UNIVERSITÉ CATHOLIQUE DE LOUVAIN
FACULTÉ DE PHILOSOPHIE ET LETTRES

UNE APPROCHE HYBRIDE DE L'EXTRACTION D'INFORMATION

Sous-langages et lexique-grammaire

Thèse présentée par Patrick Watrin en vue de l'obtention du titre de Docteur en Linguistique.

Jury : Pr. C. Fairon (directeur)
Pr. F. Guenthner
Pr. J.-R. Klein
Pr. E. Laporte
Pr. P. Mertens
Pr. J.-C. Polet (président)

LOUVAIN-LA-NEUVE
ANNÉE ACADÉMIQUE 2006-2007
CENTRE DE TRAITEMENT AUTOMATIQUE DU LANGAGE

Table des matières

I Notions Préliminaires 23

1 Lexique-grammaire : entre lexique et syntaxe 25

1.1 Fondements théoriques 25

1.1.1 Quelle unité ? 26

1.1.2 Quelle méthodologie ? 31

1.1.3 Quels critères ? 32

1.1.4 Quelle représentation ? 34

1.2 Une dialectique linguistique 36

1.2.1 Introspection *vs* déduction 38

1.2.2 Généralisation *vs* discrétisation 42

1.3 Matrices, automates et grammaires locales 46

1.3.1 Applications linguistiques 46

1.3.2 Applications didactiques 47

1.3.3 Applications informatiques 48

1.4 Conclusion 52

II Le lexique-grammaire au service de l'extraction d'information 55

2 Extraction d'information : état de la question 57

2.1 Des conférences à l'origine de l'extraction : MUC et ACE 59

2.1.1 Analyse locale 60
2.1.2 Analyse du discours 62
2.2 Une analyse locale entre recherche et compréhension 64
2.2.1 Analyse profonde *vs.* analyse de surface 65
2.3 Une nécessaire adaptabilité 67
2.3.1 Approche statistique et apprentissage 67
2.3.2 Approche linguistique et modularité 70
2.4 La phrase élémentaire comme patron informationnel 71
2.5 Conclusion 76

3 De la table à l'automate d'extraction 79
3.1 Introduction 79
3.2 Étiquetage morpho-syntaxique 84
3.2.1 Étiquetage statistique 86
3.2.2 Étiquetage linguistique 88
3.2.3 Quelle solution pour l'étiquetage ? 91
3.3 Extraction, tables et graphes paramétrés 95
3.4 Résultats en chiffres 98
3.5 Limites et perspectives 100
3.5.1 Le plan morpho-syntaxique 100
3.5.2 Le plan syntaxique 103
3.5.3 Le plan sémantique 111
3.6 Conclusion 118

III Une approche hybride et spécialisée 121

4 Vers un lexique-grammaire de spécialité 123

4.1 Introduction . . . 123
4.2 La notion de sous-langage . . . 124
4.2.1 Définition . . . 124
4.2.2 Débat terminologique . . . 128
4.3 Domaines d'application . . . 130
4.3.1 Sous-langage et recherche d'information . . . 131
4.3.2 Sous-langage et extraction d'information . . . 132
4.3.3 Sous-langage et traduction automatique . . . 134
4.4 Analyse linguistique d'un sous-langage . . . 134
4.4.1 Corpus de spécialité . . . 135
4.4.2 Lexique-grammaire, sous-langage et analyse syntaxique hybride . . . 140
4.5 Conclusion . . . 145

5 Apprentissage lexical et sous-langage 147

5.1 Introduction . . . 147
5.2 Structures lexicales complexes . . . 148
5.2.1 La collocation comme unité distributionnelle . . . 149
5.2.2 Collocation et traitement automatique des langues . . . 152
5.2.3 Collocation et sous-langage . . . 155
5.2.4 Procédure d'extraction . . . 157
5.2.5 Typologie des structures . . . 164
5.2.6 Quelques résultats . . . 174
5.2.7 Un dictionnaire de collocations . . . 180
5.3 Entités nommées . . . 183
5.3.1 Un dictionnaire d'entreprises . . . 183
5.3.2 Une formalisation générique . . . 184

5.4 Conclusion 186

6 Analyse syntaxique de surface 189

6.1 Introduction 189

6.2 Segmentation en syntagmes non récursifs 190

6.2.1 Typologie des structures 193

6.3 Attachement des constituants et sous-langage 201

6.3.1 Restrictions lexicales 202

6.3.2 Restrictions syntaxiques 204

6.3.3 Restrictions sémantiques 209

6.3.4 Déviance grammaticale 211

6.3.5 Symboles et unités 213

6.4 Conclusion 215

A Notations *Unitex* et *Outilex* 221

B La table 36DT : propriétés standard 223

C Manuel de PATEX 225

C.1 Introduction 225

C.2 Installation 225

C.2.1 Pré-requis 225

C.2.2 Installation de *PatEx* 226

C.2.3 Configuration 226

C.3 Utilisation 227

C.3.1 Le module *patex* 228

C.3.2 La commande *lgx* 230

C.4 Fichier de configuration 231

C.4.1 Éléments transversaux 231

C.4.2 Les corpus 233

C.5 Les motifs d'extraction 233

C.5.1 Correspondance inter-structurelle 236

C.5.2 Correspondance intra-structurelle 236

C.6 Le sous-langage 237

C.6.1 Les lexiques 237

C.6.2 La grammaire 240

Table des figures

1.1 Représentation arborescente de (*aimer*) 36

1.2 Représentation arborescente de (*vouloir*) 37

1.3 Le lexique-grammaire comme dialectique linguistique 46

1.4 Automate à états finis élémentaire 49

1.5 Automate à états finis non déterministe 49

1.6 Automate après déterminisation 49

1.7 GN minimal 51

1.8 GN minimal et structures de traits 52

2.1 Recherche simple sur *Google* 57

2.2 Recherche avancée sur *Google* 58

2.3 Extracteur automatique d'actualités de *Google* 58

2.4 Analyse syntaxique 62

2.5 Identification des motifs d'extraction 62

2.6 Réseau conceptuel des transactions 72

3.1 Article *céder* (extrait), Trésor de la Langue française 85

3.2 Étiquetage et conversion 92

3.3 Texte ambigu après consultation des dictionnaires 94

3.4 Processus de conversion des graphes paramétrés 96

3.5 Étiquetage XML 97

3.6 Automate d'extraction principal 99

3.7 Complexité lexicale : *le groupe nucléaire français* 102

3.8 Apposition prédicative 106

3.9 Syntagme nominal minimal 108

3.10 Syntagme nominal complexe 109

3.11 Graphe principal d'identification des anaphores 113

3.12 Anaphore introduite par *qui* 114

3.13 Anaphore introduite par *que* 114

3.14 Anaphore introduite par un pronom démonstratif 115

3.15 Anaphore introduite par un pronom personnel 116

3.16 Groupe nominal simplifié 118

4.1 Langue naturelle (L), générale Lstd et spécialisée (L′) 126

4.2 Méthode d'acquisition de ressources textuelles 139

4.3 Procédure incrémentale de segmentation/attachement 146

5.1 Combinatoire lexicale : marge d'exploitation 149

5.2 Combinatoire lexicale : chiffre d'affaires net 149

5.3 Graphes d'extraction principaux 159

5.4 Processus d'extraction 160

5.5 Éléments terminaux 164

5.6 Insertion 165

5.7 Structures de bigrammes 168

5.8 Distribution fréquentielle (bigrammes) 169

5.9 Distribution fréquentielle (trigrammes) 171

5.10 Structures de trigrammes *NAX* (*cf.* Exemples 147 à 152) 172

5.11 Structures de trigrammes *NNX* (*cf.* Exemples 153 à 158) 173

5.12 Structures de trigrammes $NprepNX$ (*cf.* Exemples 159 à 167) 174

5.13 Fréquences cumulées 177

5.14 Couverture lexicale 179

5.15 Graphe paramétré 181

5.16 Formalisation générique des noms propres 185

6.1 Segmentation en chunks 191

6.2 Modifieurs ($XADV$ et XA) 194

6.3 Chunks nominal et prépositionnel (XN et XP) 196

6.4 Structures verbales simples (XV et XVP) 198

6.5 Structures verbales infinitives (XVI et $XVIP$) 198

6.6 Structures verbales participiales (XVG et $XVGP$) 199

6.7 Graphe des symboles (extrait) 214

C.1 Graphe paramétré de pré-traitement 233

Liste des tableaux

1.1 Extrait de la table 36DT 35

1.2 Table de transitions 49

2.1 Extrait d'un texte issu de MUC-3 59

2.2 Exemple de formulaire issu de MUC-3 60

2.3 Illustration du scénario des transactions financières 60

2.4 Analyse lexicale 61

2.5 Résolution des coréférences 63

2.6 Formulaire complet 63

2.7 Exemple de motifs extraits par AutoSlog 68

2.8 Exemple de motifs extraits par Ex-Disco 70

2.10 Exemples de dépêches financières 73

3.1 Patrons d'extraction 80

3.2 Exemple d'ambiguïté lexicale 87

3.3 Sortie du TreeTagger 89

3.5 Extrait du dictionnaire de l'énergie pour le français 95

3.6 Prédicats retenus pour l'évaluation 98

3.7 Résultats d'évaluation 100

3.8 Exemples de mots inconnus 101

3.9 Grammaire jouet 102

4.1 Structures transitives locatives 125
4.2 Patron informationnel médical simplifié 131
4.3 Top 20 des motifs d'extraction extraits par [Riloff et Jones, 1999] 133
4.4 Illustration du processus de traduction (TAUM) 135
4.5 Caractéristiques du corpus de dépêches financières 137
4.6 Exemple d'organisation d'une dépêche 138
4.7 Carte d'identité de `firstinvest.com` 140

5.1 Extrait du corpus de médecine interne 156
5.2 Décompression des structures coordonnées 167
5.3 Structures nominales complexes 168
5.4 Exemples de structures $A-N$ et $prep-N$ 169
5.5 Exemples de structures NA et $NprepN$ 170
5.6 Exemples de structures NN 170
5.7 Dérivés de structures complexes 171
5.8 Corpus d'évaluation (extraction terminologique) 175
5.9 *Tokens* et *types* 175
5.10 Précision 176
5.11 Répartition des erreurs internes et externes 176
5.12 Répartition fréquentielle 180
5.13 Taille du dictionnaire *entreprises.dic* 183
5.14 Dictionnaire des entreprises (extrait) 184
5.15 Exemples de mots inconnus 185
5.16 Noms propres complexes 185

6.1 Processus d'analyse 190
6.2 Types de chunks 192

6.3 Étapes de la segmentation en chunks 193

6.4 Structures adverbiales de temps (extrait) 195

6.5 Évolution successive de l'ambiguïté 200

6.6 Ratio type/token spécialisé *vs* général 202

A.1 Principaux codes grammaticaux 221

A.2 Quelques codes sémantiques 221

A.3 Principaux codes flexionnels 222

A.4 Liste des symboles et unités 222

C.1 Description du module *PatEx* 228

C.2 Exemple partiel de classe d'extraction 234

C.3 Exemples de dépêches 235

Remerciements

Merci à Cédrick Fairon pour avoir fait du travail une récréation enrichissante scientifiquement et humainement. On a beaucoup travaillé et on a bien rigolé. Les conseils qu'il ma donnés, les discussions qu'on a pu avoir, les jours où on a refait le monde vont me manquer, c'est sûr. Mais comme dirait un grand penseur californien de notre temps : *I'll be back* et dans pas longtemps j'espère.

Merci à mes relecteurs – Franz Guenthner, Jean Klein, Éric Laporte et Piet Mertens – pour l'honneur et le plaisir qu'il me font en participant à mon jury. Merci également à Jean-Claude Polet qui a accepté de le présider.

Merci aux membres du Cental – Anne, Laurent, Mitch, Julia, Babske, Claude et Isabelle – pour le soutien qu'ils m'ont apporté dans des moments qui, grâce à eux, n'ont jamais été aussi difficiles qu'ils auraient pu l'être. Les promenades de santé autour de la fac, les verres jusque pas d'heure, les midis gastronomiques y sont pour beaucoup.

Merci aux naïades du Cecl – Sylvie, Claire, Gaëtanelle, Magali, Fanny, Christelle, Jenny, Marie-Aude et Céline – pour les midis... misogynes et caféinés qui n'ont jamais cessé de m'émerveiller. Merci particulièrement à leur cheffe, Sylviane Granger, qui a toujours trouvé le temps de me montrer la voie sans hésiter à déblayer le chemin.

Merci à mes potes de banlieue – Anastasia, Christian, Olivier, Matthieu et Sébastien – parce que c'est vraiment agréable de travailler avec eux mais aussi (et surtout) parce qu'ils m'ont initié à cette théorie chomskyenne qui, en leur compagnie, a comme un goût de fête.

Merci à mes potes tout court – Mat, Greg, $(Ju)^2$, Ced, Steph, A-So et tellement d'autres – qui m'ont forcé à déconnecter en invoquant des arguments toujours plus inattaquables, me couvrant ainsi auprès de ma conscience.

Merci à Pierre Watrin, Anne Dister et Laurent Kevers pour leurs relectures multiples et leur vision de l'orthographe qui dépasse parfois mon entendement. Ils m'ont épargné une honte certaine et éternelle en évacuant les libertés qu'il m'arrive de prendre par rapport au bon usage, succombant à je ne sais quelle licence poétique.

Merci aux deux premières femmes de ma vie qui se sont inquiétées à ma place ;), m'ont appelé, écrit et fait plus qu'il n'en faut sans rien attendre, comme d'habitude.

Merci, finalement, à Cathy, mon amoureuse, qui a réussi, contre toute attente, à supporter mes silences de plus en plus silencieux, mon humeur de plus en plus... relative sans compter mes oublis constants, mon activité sociale inexistante et bien d'autres choses encore. Ça va changer, promis !

Introduction

Telle qu'elle a été définie par les conférences MUC (*Message Understanding Conference*) et ACE (*Automatic Content Extraction*), l'extraction d'information consiste en l'identification, dans un document, des entités et des relations qui caractérisent un événement. En ce sens, elle se place à l'intersection de la recherche d'information et de la compréhension textuelle. En effet, elle cherche, d'une part, à identifier les sources pertinentes au regard d'une requête ($\simeq$ recherche d'information) et, d'autre part, à organiser l'information interne de ces sources en un certain nombre de champs prédéfinis ($\simeq$ compréhension textuelle).

Le développement d'un extracteur pose deux questions fondamentales :

1. *Quelle adaptabilité ?* Les systèmes d'extraction d'information actuels se basent principalement sur un ensemble de ressources lexicale et syntaxique spécialisées. L'adaptation de ces ressources à un nouveau domaine peut être très fastidieuse. C'est pourquoi les conférences MUC et ACE ont ajouté un paramètre supplémentaire à leur évaluation : l'adaptabilité. Il ne suffit pas de proposer une solution viable ; il faut également garantir la flexibilité des ressources nécessaires à son fonctionnement.
2. *Quelle analyse ?* L'analyse syntaxique, élément clé de tout moteur d'extraction, s'organise en deux courants principaux : analyse profonde et analyse de surface. L'empreinte de la compréhension textuelle sur l'extraction a longtemps conduit les chercheurs du domaine à préférer une analyse profonde. Cependant, les éléments à extraire sont bien souvent très localisés et ne justifient pas nécessairement une analyse complète du texte qui, par ailleurs, peut être une source importante d'erreurs.

Nous abordons la réponse à ces questions en orientant notre réflexion vers l'utilisation des bases de données syntaxiques du lexique-grammaire. Depuis 1975, la communauté qui s'est formée autour de Maurice Gross, instigateur de cette théorie, cherche à formaliser et à classifier les faits linguistiques en considérant chaque prédicat au sein d'une phrase élémentaire, unité minimale de sens. Une telle unité est particulièrement appropriée, nous le verrons, pour répondre aux besoins de l'analyse syntaxique automatique qui, nous l'avons dit, constitue le cœur d'un système d'extraction. Considérée dans un contexte particulier (*i.e.* un domaine de

spécialité), sa pertinence dépasse toutefois la seule syntaxe. Nous pouvons, en effet, préciser sa composante sémantique jusqu'à atteindre une représentation informationnelle des énoncés. Cette représentation nous permet d'identifier les différentes entités du domaine et les relations qu'elles entretiennent, répondant ainsi aux besoins principaux d'une procédure d'extraction.

L'utilisation de ce formalisme dans un cadre applicatif réel n'est pas pour autant triviale. Nous pouvons déjà présumer de deux problèmes majeurs. Le premier concerne le lexique-grammaire lui-même. Parce qu'il vise la description de la langue standard, il ne reflète pas toujours les usages lexicaux et syntaxiques des langues spécialisées, objets d'étude de l'extraction d'information. Le second ressortit à la formalisation syntaxique des structures élémentaires et pose la question de la définition du syntagme nominal.

Nous ajoutons à ces discussions un débat plus fondamental. Bien que nous nous avancions vers une solution résolument symbolique, la dichotomie *linguistique – statistique*, propre au traitement automatique des langues, reste très présente dans notre travail. Nous ancrons ce débat autour de la question de l'adaptabilité à laquelle nous apportons une réponse hybride mêlant apprentissage statistique et modularité linguistique.

Pratiquement, nous procéderons en trois temps : présentation théorique du lexique-grammaire, appréciation des enjeux et des difficultés que pose l'extraction d'information et mise au point d'une solution effective.

La première articulation de notre travail tentera d'identifier les fondements théoriques à l'origine d'un lexique-grammaire et d'en exprimer la pertinence au regard d'hypothèses linguistiques qui, de prime abord, peuvent apparaître antagonistes mais qui entretiennent en réalité une grande complémentarité : la grammaire générative transformationnelle et la linguistique de corpus. De cette discussion ressortira la mise en évidence de deux outils linguistiques remarquablement pertinents pour le traitement automatique des langues : la phrase élémentaire et le transformationnalisme harrissien.

Dans la seconde partie, après une brève présentation des enjeux de l'extraction et des solutions existantes, nous envisagerons les perspectives qu'offre la phrase élémentaire considérée en tant que patron informationnel. Nous confronterons nos hypothèses à la réalité en définissant une première méthode d'extraction. Cette méthode sera volontairement naïve et nous permettra d'apprécier les réelles difficultés que suscite le formalisme du lexique-grammaire, d'une part, et la procédure d'analyse qu'il suggère, d'autre part.

Finalement, la troisième partie de notre exposé proposera une réponse aux problèmes identifiés. Cette réponse, considérée aux niveaux lexical et syntaxique, définira une nouvelle méthode d'analyse que nous croyons plus à même de répondre aux impératifs de l'extraction.

Nous exploitons nos recherches dans une solution logicielle d'extraction d'information. Les différentes composantes de cet outil (acquisition de données textuelles, apprentissage lexical, grammaire de chunks, *etc.*) seront naturellement présentées au fur et à mesure de nos développements théoriques. Par ailleurs, un manuel d'utilisation, donnant une vue plus globale de notre application, sera annexée à ce travail.

Première partie

Notions Préliminaires

CHAPITRE 1

LEXIQUE-GRAMMAIRE : ENTRE LEXIQUE ET SYNTAXE

1.1 Fondements théoriques

> *Les cadres syntaxiques théoriques sont dans une large mesure ceux de la grammaire générative transformationnelle tels qu'ils apparaissent notamment dans les travaux de ses fondateurs, Harris et Chomsky.* [Boons *et al.*, 1976]

L'influence harrissienne avouée est, nous le verrons, omniprésente dans les travaux du lexique-grammaire. Toutefois, nous devons modérer l'ascendant générativiste évoqué dans [Boons *et al.*, 1976] et, par conséquent, la part que les modèles de Chomsky ont pu jouer dans la mise en place d'un lexique-grammaire en tant que théorie mais également en tant que méthodologie linguistique. Très tôt, on va assister à la scission intellectuelle, malheureusement définitive, qui éloignera les transformationnalismes de Gross et Chomsky. En effet, [Gross, 1979] condamne explicitement les fondements de la grammaire générative tant théoriques que méthodologiques et y oppose ceux à l'origine du lexique-grammaire. Dès lors, s'il nous faut encore constater certaines similitudes dans les deux approches, nous pouvons nous demander s'il ne s'agit pas plutôt de l'héritage du transformationnalisme de Harris qui a longtemps dirigé les travaux des deux linguistes.

LA SYNTAXE COMME DÉMARCHE CLASSIFICATOIRE

> L'étude de la syntaxe se pose exactement de la même façon que l'étude de toute famille d'objets importante de par son nombre. Les nombres en jeu, ici les nombres de phrases, sont tels qu'il est humainement impossible d'appréhender directement leur organisation, ou même l'existence éventuelle d'une organisation. L'idée naturelle est donc de grouper les objets se ressemblant en classes, et d'étudier les classes constituées. Comme les classes sont moins nombreuses que les objets, on peut es-

> pérer qu'il sera possible de percevoir l'existence d'une organisation qui les lierait, et d'en déduire une organisation pour les objets de départ. [Boons *et al.*, 1976 : p. 7]

Comment ne pas déjà ressentir, dans ces lignes, le fossé qui séparera bientôt lexique-grammaire et grammaire générative ? Il est évident que la démarche taxinomique énoncée ici ne peut s'accommoder des idées génératives. Chomsky n'aura de cesse de condamner de telles idées. Il s'y opposera notamment dans [Chomsky, 1970] et [Chomsky, 1976] où il affirme avoir noté l'émergence d'une nouvelle tendance behavioriste pour le moins saugrenue qui s'apparente à de la taxinomie pré-darwinienne. Il ira jusqu'à refuser, à cette nouvelle sorte d'ingénierie, toute puissance théorique.

Nous ne pouvons cependant pas nous étonner de l'attitude de Chomsky. Ce dernier considère que la linguistique théorique doit rendre compte de la créativité du langage, créativité qu'il annonce infinie. Comment dès lors réaliser la taxinomie de l'infini ? C'est, tout simplement impossible et, en conséquence, sans fondement. Cependant, l'infinité dont parle Chomsky repose uniquement sur un système de règles récursif qui permet d'engendrer des phrases telles que :

(1) Mathieu croit que Cathy sait que Nicolas dort.

(2) Cédric veut que Patricia sache qu'Anne-Sophie connaît la vérité.

Si le modèle représenté sommairement au travers de ces deux exemples est effectivement récursif, l'infinité qui en résulte n'est pas à proprement parler celle de la richesse sémantique mais bien plutôt une richesse de variantes organisée autour d'un nombre relativement restreint, et par conséquent nombrable, de phrases minimales. Nous constatons donc qu'en déplaçant le focus de la linguistique sur ces unités minimales, rien ne s'oppose plus à la démarche classificatoire.

1.1.1 Quelle unité ?

Ce déplacement du focus de la linguistique de la phrase complexe à la phrase simple répond à une des questions fondamentales que suppose tout classement, à savoir, *quoi classer ?*

Le concept de la phrase simple comme unité de base se retrouve également dans les théories transformationnelles de Harris et Chomsky. [Harris, 1981e] définit la phrase élémentaire comme l'unité transformationnelle minimale. Nous retrouvons également ce principe dans [Chomsky, 1969] qui définit la structure profonde comme l'ensemble des propositions élémentaires *sujet-prédicat* auxquelles s'appliquent les transformations. Toutefois, le lexique-grammaire dépasse cette vue essentiellement syntaxique. L'introduction du lexique dans la grammaire l'a amené à apprécier la phrase simple en tant qu'unité syntaxique, mais également sémantique. La pertinence de l'idée a maintes fois été soulignée au sein de la communauté du lexique-grammaire

et trouve son origine dans cette affirmation qu'«un élément ne peut être compris dans ses emplois particuliers qu'à l'aide d'informations qui relèvent exclusivement de l'agencement des éléments» ([G. Gross, 1981a : p. 35]).

La phrase élémentaire ou *phrase noyau* rejoint la conception syntaxique commune qui attribue aux phrases la forme canonique *sujet – verbe – compléments*. [Gross, 1997] note cette séquence :

$$N_0\ V\ W$$

Où,

– N_0 est le sujet,
– V est le verbe,
– W symbolise une suite indéterminée, éventuellement vide, de compléments essentiels N_i.

$\boxed{N_0\ V.}$ La définition de la phrase simple présuppose vrai le postulat qui affirme que toute phrase a un sujet situé à gauche du verbe. Que dire alors des propositions suivantes :

(3) Range ta chambre !

(4) Pas de soucis.

(5) Ainsi va la vie.

Afin de répondre à cette question, nous devons considérer l'hypothèse transformationnelle de [Harris, 1981e]. Cette hypothèse affirme que, si deux énoncés peuvent se déduire l'un de l'autre au prix d'une opération combinatoire (une transformation), ces deux énoncés participent de la même phrase noyau. Il s'agit, dans une certaine mesure, de ce que Chomsky qualifierait de différences de surface au départ d'une unique structure profonde.

Cette seule hypothèse nous permet bien de rendre compte de l'adéquation des propositions 3, 4 et 5 au modèle de la phrase élémentaire. Il nous suffit pour cela d'évoquer deux transformations : l'*effacement* et la *permutation*. 3 et 4 peuvent être déduits, par effacement, des deux énoncés suivants :

(6) (Je veux | Il faut) que tu ranges ta chambre.

(7) Il n'y a pas de soucis.

quand 5 est issu par permutation de la phrase :

(8) La vie va ainsi.

Ceci étant, il persiste certains problèmes introduits par la notion complexe de figement (nous y reviendrons par après) qui caractérise les phrases du type :

(9) Autant en emporte le vent.

La déduction est ainsi beaucoup moins aisée que dans les trois cas précédents même si l'étymologie nous enseigne l'existence de la structure :

(10) Le vent en emporte autant.

Malgré ce dernier point et quelques réelles exceptions, nous pouvons conférer au modèle $N_0\ V$ un certain caractère de généralité.

$\boxed{V.}$ La structure $N_0\ V\ W$, en ce sens qu'elle formalise l'idée commune d'une organisation phrastique *sujet-verbe-complément(s)*, confère implicitement un rôle prédicatif au verbe, chaque N_i lui étant subordonné. Toutefois, [Harris, 1981e] met en évidence, au travers de son étude de la nominalisation, une intéressante catégorie de verbes qui ne semblent pas organiser leurs compléments d'une manière ordinaire (*cf.* Exemple 11).

(11) 1. Votre théorie contredit la grammaire générative.

2. Votre théorie est en contradiction avec la grammaire générative.

[Gross, 1993] explique cette transformation par comparaison à la passivation : quand la passivation d'un verbe conduit à l'insertion de l'auxiliaire *être* et de la préposition *par*, la nominalisation, dans cet exemple, implique l'emploi du verbe *être* et de la préposition *avec*. Nous parlerons toutefois, dans le cas de la nominalisation, de *verbe support* et non d'auxiliaire. Mais, comme lui, il est sémantiquement neutre, le sens étant réalisé par le substantif (appelé *nom prédicatif*).

L'intérêt de la découverte d'Harris est primordiale pour la théorie du lexique-grammaire. Elle offre en effet la possibilité d'ajouter à l'étude des verbes celle des noms (des adjectifs et des adverbes) depuis que ceux-ci peuvent être représentés au sein d'une phrase élémentaire. Considérons la phrase :

(12) Max commet un meurtre.

Nous ne pouvons nier que cet énoncé possède les mêmes caractéristiques que (*cf.* Exemple 12). Toutefois, nous n'aurions pu, sans recourir aux verbes supports, prétendre l'étudier dans le cadre du lexique-grammaire puisque le substantif *meurtre* ne connaît pas de réalisation verbale.

$\boxed{W.}$ La principale difficulté posée par la phrase élémentaire est la définition des compléments essentiels. Une telle abstraction pose immédiatement deux questions : comment délimiter la frontière de tels compléments ?, où arrêter l'essentiel ?

Quelles limites au syntagme nominal ? Chaque N_i que compte W marque un syntagme ou groupe nominal (noté également tantôt *SN*, tantôt *GN* selon les auteurs), composant phrastique essentiel. Sa caractérisation sémantique *a priori* ne pose pas de problème et on s'accorde à dire qu'il constitue une unité désignant une réalité abstraite ou concrète. Par ailleurs, et d'un point de vue grammatical, cette unité s'articule obligatoirement autour d'un nom, tête du complément, comme le montrent les quelques exemples repris ci-dessous, exemples inspirés de l'étude menée par [Gross, 1997].

(13)

N_0	V	N_1	
La jeune fille	aime	Cathy	a) *Npr*
		les enfants	b) *DetN*
		les enfants qui restent sages	c) *DetNRel*
		l'idée qu'elle se marie	d) *DetNqueP*
		les enfants du boulanger	e) *DetNPrepN*

Si nous retrouvons bien les notions syntaxico-sémantiques que nous venons d'exposer, nous constatons également qu'une caractérisation formelle plus fine, sinon plus complète, nécessiterait une étude largement plus complexe. Les exemples c) et d) montrent qu'un syntagme nominal peut étendre sa complexité jusqu'à couvrir celle de la phrase. Par conséquent, l'appareillage formel nécessaire à son étude engloberait la grammaire complète de la langue.

Ce constat a conduit le lexique-grammaire à répondre à la question de la limite du syntagme nominal d'une manière plus intuitive (mais aussi plus immédiate) et c'est une nouvelle fois l'hypothèse transformationnelle qui nous permet d'évacuer la difficulté. Dans le très bref aperçu que nous donnons de la permutation, nous constatons que cette transformation affecte (tout comme l'effacement ou la pronominalisation d'ailleurs) un syntagme nominal entier qu'elle déplace. C'est précisément ce déplacement qui nous permet d'apprécier intuitivement les frontières des syntagmes nominaux.

Parmi l'ensemble des transformations, l'extraction semble la mieux adaptée. En effet, «elle

délimite un GN sujet ou complément (et un seul) et le déplace entre *C'est* et *Que*» (*cf.* [Gross, 1997 : p. 27]. Dès lors, au départ de la phrase,

(14) Laurence parle à Olivier dans le petit salon.

l'extraction nous permet d'isoler les deux syntagmes nominaux suivants :

(15) C'est *à Olivier* que Laurence parle dans le petit salon.

(16) C'est *dans le petit salon* que Laurence parle à Olivier.

et répond donc au problème de la délimitation des N_i[1].

Où s'arrête l'essentiel ? L'usage grammatical commun distingue les compléments essentiels, objets direct et indirect, des circonstanciels, marquant le lieu, le temps et la manière. Toutefois, un problème subsiste : les circonstants n'ont pas toujours ce caractère facultatif qui les oppose aux essentiels. Ainsi les phrases,

(17) Patricia habite **à Braine-l'Alleud**.

(18) Pierre passe **deux jours par semaine** à s'entraîner à la course.

ne peuvent se défaire de leur complément circonstanciel et demeurer grammaticales.

(19) *Patricia habite.

(20) *Pierre passe à s'entraîner à la course.

Néanmoins, au sein des deux constructions suivantes, ces mêmes compléments sont bien optionnels.

(21) Patricia fait ses courses **à Braine-l'Alleud**.

(22) Pierre travaille **deux jours par semaine**.

Il n'est donc pas de règle absolue nous permettant d'isoler les compléments essentiels. L'usage fluctue selon l'environnement verbal et doit être remis en question pour chacun des verbes de la langue.

[1]Cette réponse est théorique. Nous verrons que la formalisation du *GN*, dans le cadre d'une application de traitement automatique, pose encore problème.

1.1.2 Quelle méthodologie ?

Si c'est bien au prix d'une étude systématique du lexique verbal que nous pouvons espérer couvrir la notion de phrase élémentaire, ce nécessaire systématisme dépasse cette seule notion et constitue même la clef de voûte méthodologique du lexique-grammaire.

(23) Ce problème *concerne* Luc – Luc est *concerné* par ce problème

(24) Ce problème *regarde* Luc – *Luc est *regardé* par ce problème

En considérant les phrases 23 et 24, on constate que seule la première accepte la forme passive. Il n'est donc pas, à proprement parler, de généralisation possible. Cette opposition syntaxique, loin d'être ponctuelle ou limitée à la passivation, est généralisée à l'ensemble du lexique et de la syntaxe. De la même manière que l'étendue des compléments essentiels varie d'un verbe à l'autre, les traits syntaxiques qu'ils vérifient peuvent varier. A tel point qu'il n'a pas été possible, au terme de l'étude de 15 000 formes verbales, d'isoler deux verbes ayant exactement le même comportement syntaxique.

> If we compare, two-by-two, the rows of the matrix – i.e., if we compare the syntactic properties of any two lexical items – it is observed that no two lexical items have identical syntactic properties. If we compare the columns two-by-two, i.e. the domains of the rules, the result is the same. [Gross, 1979 : p. 860]

Par conséquent, le systématisme taxinomique du lexique-grammaire n'est pas, comme Chomsky aimerait le faire croire, une étrange vue de l'esprit sans fondement théorique, mais bien une absolue nécessité pour qui souhaite formaliser la langue d'une manière un tant soit peu générale et scientifique. L'importance d'une accumulation systématique de faits linguistiques se comprend donc doublement. Elle permet, d'une part, de démontrer efficacement et sans équivoque l'adéquation, mais, également, la portée des équivalences transformationnelles que les génératives se contentent de justifier au départ de quelques exemples. Et, d'autre part, c'est bien par l'observation d'un très grand nombre de phénomènes que nous pouvons espérer focaliser notre attention sur les récurrences et en tirer des généralisations pertinentes, évitant ainsi la généralisation de certains faits isolés.

L'inadéquation des modèles formels, tels que celui de la grammaire générative, qui refusent tout à propos théorique à la systématisation, se mesure à la lumière de l'énorme disproportion entre le langage, extrêmement redondant, engendré par application du modèle et la réalité des faits linguistiques acceptables. En somme, quand Chomsky s'attaque à l'étude du possible, Gross limite ses recherches au probable.

1.1.3 Quels critères ?

Toute démarche classificatoire suppose à terme la distribution des faits observés en classes d'équivalences. La distribution des faits s'organise nécessairement autour d'un ensemble de caractéristiques. En effet, pour intégrer une classe, deux faits doivent partager différents traits distinctifs. Il est évident que l'établissement d'un lexique-grammaire intègre bon nombre de critères syntaxiques. Toutefois, cette classification ne se limite pas à la syntaxe ; des critères numérique et sémantiques, bien que subsidiaires, entrent en ligne de compte.

Critère numérique. La commodité d'usage impose une limitation numérique aux éléments d'une classe. Nous ne pouvons, en effet, prétendre manipuler efficacement une classe si cette dernière compte plus de cent ou deux cents éléments ([Boons *et al.*, 1976 : p. 17]).

Critères sémantiques. L'association forme-sens, point théorique récurrent en linguistique, se traduit en lexicographie par la combinaison de la syntaxe et de la sémantique, la mise en contexte permettant la désambiguïsation. Le lexique-grammaire quant à lui réalise l'assujettissement inverse et utilise des critères sémantiques distinctifs afin de pallier certaines insuffisances syntaxiques. Il s'agit exclusivement de critères de sélection visant à contraindre les arguments d'un prédicat. Le lexique-grammaire distingue communément deux types de propriétés selon qu'elles sont appréciables sur des substantifs isolés (propriétés absolues – A) ou qu'elles prennent sens une fois en contexte (propriétés relatives – R)[2].

A	$N =: Nhum$	humain, animé
	$N =: N-hum$	non-humain
	$N =: Npl$	pluriel
	$N =: Nconcret$	concret
	$N =: Nloc$	locatif
R	$N =: V-n$	morphologiquement lié au verbe
	$N =: "classifieur"$	classifieur sémantique
	$N =: Npc$	partie du corps

Critères syntaxiques. Ces critères visent avant tout à caractériser la nature des compléments essentiels associés au prédicat dans la phrase élémentaire. Cette idée rejoint le distributionnalisme harrissien qui confère au contexte syntaxique une valeur distinctive.

> Assignment of an element to a particular class is always relative to the assignment of its environment. The elements are not defined except in relation to their environment. [Harris, 1974 : p. 13]

[2] Nous proposons dans l'annexe A, un résumé des différentes notations utilisées dans ce travail.

La création de classes ressortit à cette analyse, chacune étant articulée autour d'un patron syntaxique du type de la phrase élémentaire, $N_0\ V\ W$. Un test de substitution doit permettre au linguiste de déterminer l'appartenance d'un élément à une classe donnée en fonction du nombre mais aussi de la forme des compléments essentiels. Ainsi, le verbe *acheter* a pour construction de base celle de la table *36DT* :

$$N_0\ V\ N_1\ \grave{a}\ N_2 \text{ (Max achète un livre à Lisa)}$$

construction à laquelle sont associés certains critères sémantiques, N_0 et N_2 devant être humains (noté $N_0 =: Nhum$, $N_2 =: Nhum$).

Le résultat de l'application systématique de cette analyse conduit à donner une image précise de ce que [Harris, 1964] appelle la *liberté d'occurrence* des éléments. En effet, nous pouvons, sur base des classes constituées, déterminer quel énoncé appartient à quelle langue et quel autre lui est étranger. Toutefois, nous n'agissons ici que dans le cadre de la phrase élémentaire et de la langue standard qui ne peuvent bien entendu prétendre couvrir l'ensemble des productions langagières. C'est pourquoi le lexique-grammaire complète l'analyse distributionnelle par celle transformationnelle. Dès lors, le patron syntaxique de chaque classe se voit augmenté d'un certain nombre de transformations (celles exposées dans [Harris, 1981c]) comme autant de critères qu'un élément lexical doit remplir pour s'inscrire dans cette classe. Aux critères distributionnels, tant sémantiques que syntaxiques, déjà évoqués pour *acheter* s'ajoutent donc ceux transformationnels, dans ce cas-ci la pronominalisation du N_2 humain en *lui* (*lui* doit être considéré ici comme la forme canonique de *me*, *te*, *lui*, *se*, *nous*, *vous*, *leur*).

Nous l'avons vu en début de chapitre, Gross s'oppose à Chomsky en ce qu'il se refuse à limiter la créativité linguistique aux seuls procédés transformationnels récursifs agissant par combinaison de deux phrases élémentaires : les transformations binaires (coordination, subordination circonstancielle, relativisation, subordination complétive). Par conséquent, les transformations prises en compte sont celles, unaires, agissant au départ d'une unique phrase élémentaire (passivation, pronominalisation, extraposition et nominalisation). Malgré cela, et puisque les transformations binaires s'obtiennent par opération sur deux phrases élémentaires, [Gross, 1990 : p. 45] considère les classes d'équivalences comme un «ensemble qui sert de générateur pour les phrases complexes».

Il résulte de l'application systématique des trois critères – syntaxique, sémantique et, dans une moindre mesure, numérique – l'organisation du lexique en classes disjointes. Il est intéressant de noter que cette organisation oppose à nouveau lexique-grammaire et grammaire générative. [Chomsky, 1976], dans sa description du composant de base, montre un lexique essentiellement binaire. Chaque entrée lexicale est caractérisée par un certain nombre de traits syntaxico-sémantiques communs spécifiés positivement ou négativement.

(25) 1. (sincerity, [+N, +Det-, -Count, +Abstract,...])

2. (boy, [+N, +Det-, +Count, +Animate, +Human,...])

Au contraire du lexique-grammaire qui présente une réelle organisation du lexique réalisée sur la base de classes disjointes, la grammaire générative, par utilisation de traits binaires, obtient un ensemble de classes se recouvrant à partir duquel ne peut être inférée une quelconque structure. [Boons *et al.*, 1976 : p. 16] voit là «la cause principale de l'absence d'études lexicales sérieuses dans le cadre de la grammaire générative». Une telle situation s'explique très certainement par la priorité donnée aux règles, le lexique ne servant qu'à remplir les variables que ces règles renferment.

1.1.4 Quelle représentation ?

Les données linguistiques étudiées dans le cadre d'un lexique-grammaire sont représentées sous la forme de matrices binaires dans lesquelles chaque ligne représente une entrée lexicale et chaque colonne une propriété syntaxique. Les intersections ligne-colonne sont remplies d'un signe '+' si l'entrée courante vérifie la propriété syntaxique ou sémantique et d'un signe '-' dans le cas contraire[3]. Pour reprendre la terminologie établie par [Gross, 1975 : p. 152] : «Une ligne de table correspond donc à un paradigme syntaxique».

Les transformations permises pour chaque entrée lexicale sont directement observables au sein de la matrice sous la forme d'une imbrication de colonnes. Les colonnes imbricantes représentent tantôt la phrase élémentaire initiale, tantôt les conditions inhérentes à la transformation ; de la même manière, les colonnes imbriquées formalisent la structure dérivée, transformée. Associée au formalisme binaire déjà évoqué, cette représentation permet d'apprécier, pour chaque entrée, l'existence (ou non) de la forme transformée.

L'intérêt d'une représentation matricielle binaire est avant tout informatique. Les informations transformationnelles, de même que les contraintes distributionnelles, rendent possible l'utilisation du formalisme tant en génération qu'en reconnaissance.

La représentation matricielle, si elle se prête relativement bien à divers traitements automatiques, doit également être comprise comme une réponse aux insuffisances des représentations arborescentes utilisées en grammaire générative.

> Constituent analysis characterizes the sentences of a language as follows : Each sentence consists of a sequence of constituents (say, noun-phrase and verb-phrase in the case of English and many other languages), each of which in turn is a se-

[3]Nous proposons un extrait de la table 36DT dans le tableau 1.1. Les différentes propriétés sont, quant à elles, explicitées dans l'annexe B.

Table 36DT

$N_0 =: Nhum$	$N_0 =: V-n$	N_2 *bénéficiaire*		$N_1 =: Nhum$	N_0VN_0pc	$N_1 =: N-hum$	$N_1 =: V-n$	$N_1 =: DnumN$	$N_1 =: coup$	*Prép* = : *à*	*Prép* = : *de*	$N_2 =: N-hum$	$Ppv =: lui$	$Ppv =: y$	$N_2 =: V-n$	$N_2 =: Nplobl$	$N_1 estVpp$	$N_1 =: Nabst$	N_0VN_1DnumN	$N_0VN_2(deN_1)$	N_0VN_1LocN	N_0VN_1 *à* N_2Advp
+	-	-	accepter	-	-	+	-	+	-	-	+	-	-	-	-	-	-	-	-	-	-	-
+	-	+	acheter	-	-	+	+	-	-	+	-	-	-	-	-	-	-	+	+	-	-	+
+	-	-	acquérir	-	-	+	+	-	-	-	+	+	+	-	-	-	-	+	+	-	-	+
+	-	+	céder	-	+	+	-	-	-	+	-	-	-	-	-	-	-	-	+	-	-	+
+	-	+	débourser	-	-	+	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-
+	-	-	enlever	+	-	+	-	+	-	+	-	+	+	-	-	-	-	+	-	-	+	-
+	-	+	mettre	-	-	+	-	-	+	+	-	+	+	-	-	-	-	-	-	-	+	-
+	-	+	offrir	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-
+	-	+	payer	+	-	+	-	-	-	+	-	-	-	-	-	-	+	-	+	-	-	-
+	-	+	payer	+	-	+	-	+	-	+	-	-	-	-	-	-	+	-	-	+	-	-
+	-	-	prendre	+	-	+	+	+	-	+	-	+	+	-	-	-	-	+	-	-	+	+
+	-	+	préparer	-	-	+	-	-	-	+	-	+	+	-	-	-	-	+	-	-	-	-
+	-	-	racheter	+	-	-	-	-	-	+	-	-	-	-	-	-	-	-	+	-	-	+
+	-	-	racheter	-	-	+	-	-	-	+	+	-	-	-	-	-	-	-	+	-	-	+
+	-	-	rafler	+	-	+	-	+	-	+	-	-	-	-	-	-	-	+	-	-	+	-
+	-	+	rendre	+	-	+	-	+	+	+	-	+	+	-	-	-	-	+	-	-	-	-
+	-	-	reprendre	+	-	+	-	-	-	+	-	-	-	-	-	-	-	-	+	-	-	+
+	-	-	retirer	-	-	+	-	+	-	+	-	+	+	-	-	-	-	+	-	-	+	-
+	-	+	revendre	+	-	+	-	-	-	+	-	-	-	-	-	-	-	-	+	-	-	+
+	-	+	vendre	+	-	+	-	-	-	+	-	-	-	-	-	-	+	+	+	-	-	+

Tableau 1.1 : Extrait de la table 36DT

> quence of lower-level constituents (say, noun and its modifiers in the case of a noun-phrase), and so on down until the final constituents are single morphemes or syntactically unitary (but morphemically complex) words (*e.g.* conceive, boyhood). [Harris, 1962 : p. 11]

Ainsi, dès lors qu'un constituant s'est vu décomposé en une séquence de constituants d'un niveau inférieur, les différents membres ne peuvent plus entretenir de relation sinon celle qui les unit au seul constituant supérieur. Cette hiérarchisation ne peut donc qu'échouer à représenter certaines contraintes de dépendance telles que celles évoquées dans [Gross, 1997]. La contrainte que le verbe exerce sur l'article du *SN*-complément n'est aucunement perceptible dans la représentation arborescente.

(26) Luc aime l'argent – *Luc aime de l'argent

(27) *Luc veut l'argent[4] – Luc veut[5] de l'argent

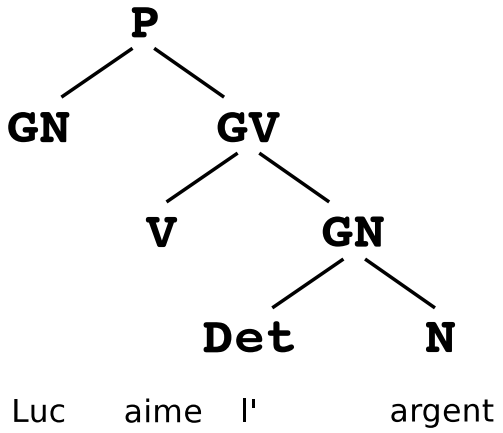


Figure 1.1 : Représentation arborescente de (aimer)

1.2 Une dialectique linguistique

> DIALECTIQUE *[...] 3.* SPECIALT. *D'après Hegel, Marche de la pensée reconnaissant le caractère inséparable des propositions contradictoires (thèse et antithèse), que l'on peut unir dans une catégorie supérieure (synthèse***).* [Rey-Debove et Rey, 1994]

Dans notre poursuite de l'étude du lexique-grammaire, nous considérons Maurice Gross, avant tout comme quelqu'un qui ne pouvait se satisfaire de solutions exclusives. Son œuvre et celle

[4]Précisons toutefois qu'en contexte cette phrase peut s'avérer être acceptable et grammaticale.

[5]Notons que, contrairement au prédicat *aimer*, la valeur du prédicat *vouloir* n'est pas générique.

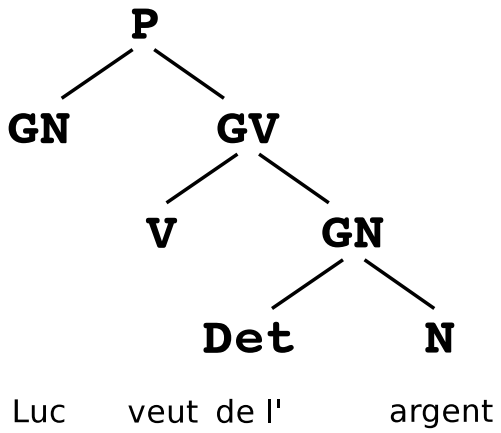


Figure 1.2 : Représentation arborescente de (vouloir)

des chercheurs qui l'ont entouré consacrent, dans une certaine mesure, la dialectique hégélienne. En effet, il tire d'éléments méthodologiques, antagonistes en apparence, une complémentarité qui transcende les réalités particulières.

Dans son union de la syntaxe et du lexique, le lexique-grammaire réconcilie dans une large mesure les positions de la linguistique formelle et des études de corpus. Cette constante dualité s'observe à tous les niveaux de la recherche linguistique telle qu'elle est menée, encore aujourd'hui, au sein de la communauté du lexique-grammaire. Nous retrouvons les deux idées unifiées du recueil des données aux conclusions que leur analyse amène.

Si, à ce stade de notre exposé, l'influence harrissienne n'est plus à justifier, nous avons voulu illustrer l'idée que nous défendons en comparant les points de vue conciliateurs du lexique-grammaire à ceux de deux autres linguistes : Chomsky, en tant que représentant de la linguistique formelle, d'une part, et Sinclair, représentant de la linguistique de corpus, d'autre part. Il ne s'agit pas de juger l'importance théorique ou historique des deux personnalités citées, mais seulement d'utiliser leurs propos très largement antagonistes afin d'étayer le nôtre.

LA SYNTAXE COMME DÉMARCHE EMPIRIQUE

> As a philosophy of science 'empiricism' consists of the use of particular methods of observation, data collection, hypothesis testing, and theory building. Empirical generalizations are arrived at by induction from observation. [Romaine, 1993 : p. 1111]

Historiquement, on oppose l'empirisme à l'introspection psychologique qui caractérise particulièrement la grammaire générative. Par ailleurs, l'usage prête souvent à la linguistique de corpus le statut de science empirique. Nous préférons cependant modérer ce propos. L'empirisme, tel que nous le concevons et tel qu'il est présenté dans l'extrait ci-dessus, participe d'une démarche hypothético-déductive, démarche qui ne peut être appréciée dans ce que nous appellerons plus tard le déductivisme de Sinclair. Traditionnellement, le monde anglo-saxon distingue deux cou-

rants en linguistique de corpus.

> It should be noted that general work which makes use of a corpus as evidence for language description is usually referred to as corpus-based. I use this term in a more restricted sense to refer specifically to work where the corpus is used mainly to expound on, or exemplify, existing theories, that is theories which were not necessarily derived with initial reference to a corpus. [...] The corpus-driven approach (Tognini Bonelli 1996b), in contrast to the corpus-based approach, constitutes a methodology that uses a corpus beyond the selection of examples to support linguistic argument or to validate a theoretical statement. [Bonelli, 2000 : pp. 73-4]

Ainsi, en utilisant le corpus comme source d'hypothèses, nous nous inscrivons dans une approche *corpus-driven* propre à Sinclair tandis qu'en recourant au corpus dans un but de vérification d'hypothèse, nous participons d'une approche *corpus-based*, plus largement répandue de nos jours.

1.2.1 Introspection *vs* déduction

> More generally, practically no GG specialist seems conscious of the fact that DEMONSTRATING THE EXISTENCE of phenomena is a prime necessity. Even if a clear difference of behaviour has been observed in some sentences, the observation may constitute an artifact. Only comparison with a large number of cases can bring plausibility as to the presence of significant data. [Gross, 1979 : p. 863]

Nous ne pouvons bien entendu limiter le travail du grammairien à la seule accumulation de données mais il est indéniable que, si elle ne constitue pas une fin en soi, cette accumulation est le préalable nécessaire à toute généralisation théorique. Il n'est donc pas étonnant que ce soit la définition même du concept de *donnée* qui soit à l'origine du dilemne linguistique exposé par [Sinclair, 1997] : doit-on focaliser son étude sur les données effectivement observables ou plutôt sur un ensemble d'énoncés artificiels, créés dans le cadre de l'étude ? En d'autres termes, doit-on préférer les données empiriques à celles introspectives ?

Chomsky répond d'une manière très claire à cette question dès lors qu'il envisage, dans [Chomsky, 1976], l'intuition comme étant l'objet même de la linguistique.

> Thus a proposed operational test for, say, segmentation into words, must meet the empirical condition of conforming in a mass of crucial and clear cases, to the linguistic intuition of the native speaker concerning such elements. Otherwise, it is without value. The same, obviously, is true in the case of any proposed operational procedure or any proposed grammatical description. [Chomsky, 1976 : p. 19]

Nous pourrions dire grossièrement qu'il ne s'agit pas de vérifier l'adéquation d'une théorie en la confrontant à la réalité, mais plutôt d'évacuer cette réalité si elle se voyait être en contradiction avec l'intuition du «linguiste-Dieu».

Nous sommes à même de mieux appréhender, au travers de cette conception de la linguistique, la nécessaire, mais, peut-être dans le même temps, artificielle, distinction chomskyenne entre *compétence* et *performance*. La compétence est la connaissance *a priori* qu'un *locuteur-auditeur idéal* («ideal speaker-listener», [Chomsky, 1976 : p. 3]) a de sa propre langue et constitue de ce fait l'unique objet de la linguistique. La performance, quant à elle, en ce qu'elle désigne l'emploi effectif d'une langue, ne peut être sujette à une quelconque généralisation. En effet, toute réalisation effective intègre, selon Chomsky, nombre de facteurs étrangers à la linguistique, parce qu'imputables aux locuteurs, facteurs qui interfèrent avec l'activité théorisante : limitations mémorielles, manque d'attention ou d'intérêt, simples erreurs,...

Un tel propos ne peut décidément s'accommoder du déductivisme de Sinclair. Ce dernier définit sa position, dans [Sinclair, 1997 : p. 1], comme reposant sur deux principes majeurs : *accept the evidence* et *reflect the evidence*. Cette *évidence* (indice, preuve) est celle de l'usage, ce que Chomsky désigne précisément par performance. Loin de contredire l'irrégularité qu'évoquait Chomsky dans sa définition de la performance, Sinclair y voit l'origine de tout travail théorique.

> Grammar is not easily applied to text, despite the claims that are sometimes made. Exceptions, gradations, continua and the like are commonplace. Likewise, the descriptions that start with text will gradually find reasons for apparent exceptions, and expose generalities. [Sinclair, 1987c : p. 159]

Cet extrait laisse transparaître son principal objectif : décrire la langue telle qu'elle est et non telle qu'elle pourrait (ou devrait) être. Cette idée de la linguistique nécessite l'appréhension des données linguistiques sans jugement *a priori*. Une telle gageure se traduit, dans les faits, par la création d'une nouvelle sorte de corpus, le *Monitor Corpus*[6] ([Sinclair, 1997 : pp. 25-6]), qui reflète quatre visions différentes de l'objectivité.

1. **Données authentiques** : nous ne pouvons prétendre étudier la langue au départ d'énoncés artificiels tels que ceux que la grammaire générative obtient par l'introspection.
2. **Données non linguistiquement motivées** : Sinclair s'oppose à l'idée d'une sélection des données dans un but illustratif.
3. **Données illimitées** : s'il est important, sinon capital, d'asseoir une théorie sur une accumulation importante de données, il convient de ne pas la limiter, mais plutôt de la rafraîchir constamment afin de ne pas mettre en avant les cas particuliers, mais également de ne pas figer la description linguistique.

[6] Cette notion dérive de celle de *corpus dynamique* développée dans [Renouf, 1993; Fairon, 1999].

4. **Données brutes** : aucune information autre que les données ne doit interférer avec le travail du linguiste ce qui revient à refuser toute annotation.

L'exploration du corpus doit, elle aussi, répondre à cette volonté d'objectivité. Dès lors, ce sont des méthodes statistiques qui mettent le linguiste au contact de l'évidence : un ordinateur dressera automatiquement la liste des mots les plus fréquents qui constituera le matériau primaire de l'analyse linguistique.

La divergence dans les objectifs poursuivis par les deux linguistes peut, pour une part, expliquer l'antagonisme des deux approches. Si nous nous plaçons dans le contexte didactique qui caractérise le travail de Sinclair, le clivage *usuel* / *non-usuel* s'avère des plus important. Il convient en effet de mettre l'apprenant au contact de la réalité de la langue afin qu'il puisse produire un langage approchant au plus celui du natif (*cf.* [Hunston et Francis, 2000 : p. 10]). De même, si nous considérons la formalisation du langage, nous ne pouvons ignorer un fait sous le seul prétexte d'une fréquence d'apparition très faible. Toutefois, nous pensons que la partialité des deux approches conduit inévitablement à un échec théorique.

Les limites de l'introspection sont évidentes et directement imputables au théoricien qui ne peut prétendre avoir une connaissance exhaustive d'une langue, fût-ce la sienne. Parallèlement, limiter son approche aux seuls corpus ne permet d'inférer la grammaire que du corpus utilisé et non de la langue dans toute sa généralité. Par conséquent, l'adoption exclusive de l'une ou l'autre position ne pouvait satisfaire la rigueur scientifique, pas plus que la prétention systématique, du lexique-grammaire. On ressent tout particulièrement cette situation dans [Gross, 1975] qui le premier exprime l'absolue nécessité de la combinaison des deux approches. Nous pourrions, encore ici, percevoir l'impact qu'a eu Harris sur les théories de Gross et du lexique-grammaire puisque [Harris, 1961] insiste sur la nécessité de maintenir le corpus, qui sous-tend une analyse syntaxique, ouvert[7] et de ne pas s'y limiter si certains faits avérés ne devaient pas s'y retrouver[8].

> Ainsi, le fait que de nombreux linguistes se soient astreints à n'étudier que des formes observées dans des textes (*i.e.* dans des corpus), a bloqué toute systématisation des recherches en syntaxe. [Gross, 1975 : p. 20]

La *servitude du corpus* qui sera également dénoncée dans [Boons *et al.*, 1976] ne peut ef-

[7] *Investigation in descriptive linguistics consists of recording utterances in a single dialect and analyzing the recorded material. The stock of recorded utterances constitutes the corpus of data, and the analysis which is made of it is a compact description of the distribution of elements within it. The corpus does not, of course, have to be closed before analysis begins.* [Harris, 1961 : p. 12]

[8] *Although our whole investigation has been in a particular corpus of utterances, we may consider this corpus to be an adequate sample of the language from which the corpus was taken. With this assumption, the methods of descriptive linguistics enable us to say that certain sequences of certain elements occur in the utterances of the language. This does not mean that other sequences of these elements, do not occur ; they may have not yet occurred without entering in our records or they may have not yet occurred in any utterance of the language, only to occur next day.* [Harris, 1961 : p. 372]

fectivement s'envisager dans la perspective transformationnelle du lexique-grammaire. Cette perspective, moins théorique qu'expérimentale, n'ambitionne pas, comme c'est le cas en grammaire générative, l'induction d'un vaste ensemble de phénomènes linguistiques au départ d'une base étroite d'énoncés. Il s'agit toutefois d'un moyen d'explorer la structure des phrases élémentaires et d'en isoler ainsi les contraintes syntaxiques. Nous ne pouvons donc prétendre nous limiter au corpus qui, aussi grand soit-il, ne pourra jamais contenir l'ensemble des variations transformationnelles de toutes les combinaisons lexicales permises par la langue standard. Par conséquent, et à l'instar de la grammaire générative chomskyenne, l'intuition servira de générateur d'exemples, contredisant ainsi l'impératif de neutralité des données cher à Sinclair[9]. Toujours à la manière générativiste, nous recourrons au jugement d'acceptabilité afin de déterminer la pertinence linguistique des énoncés créés intuitivement.

> L'analyse des textes écrits met souvent en évidence des formes considérées comme incorrectes par de nombreux locuteurs, il ne semble pas que ces formes constituent des fautes ayant échappé à l'attention de leur auteur, il semble plutôt que l'appel à l'introspection sous-évalue l'ensemble des formes grammaticales. [Gross, 1975 : p. 24]

Les corpus doivent donc être considérés, dans un cadre linguistique, comme de précieux adjuvants, non seulement parce qu'ils permettent de compléter le matériau obtenu intuitivement mais aussi parce qu'ils évitent ainsi certaines erreurs qu'un jugement trop hâtif aurait pu impliquer. Cependant, nous devons noter que les jugements d'acceptabilité, bien que terriblement subjectifs parce qu'intuitifs, respectent le principe de reproductibilité revendiqué dans la plupart des productions du lexique-grammaire et caractéristique de toute démarche scientifique. L'expérience montre qu'exercé sur un bloc de phrases minimales apparentées tant d'un point de vue lexical que syntaxique, le jugement d'acceptabilité connaît une certaine constance pour une même personne au cours du temps mais aussi au sein d'un groupe de personnes.

> La reproductibilité plus ou moins satisfaisante des jugements d'acceptabilité portant sur des séquences d'éléments lexicaux inventés la rend indépendante des corpus de mots. A cet égard, les dictionnaires et lexiques d'usage sont beaucoup plus que des adjuvants précieux. [Boons *et al.*, 1976 : p. 43]

Cette affirmation nous amène à relativiser notre idée de l'utilisation de corpus dans l'établissement d'un lexique-grammaire. Rappelons que les visées de cette théorie ne portent aucunement sur une quelconque notion d'attestation. Ainsi, il s'agit moins de recenser les séquences attestées, comme le suggère Sinclair, que de délimiter les frontières du *probable* (que nous opposons au *possible* chomskyen ; cf. *1.1.2 Quelle méthodologie ?*).

[9] *Thus examples have to be constructed with a particular testing aim in view. This can only be done by professional linguists.* [Gross, 1986a : p. 180]

Finalement, bien qu'il ne valide pas leur utilisation conjuguée, [Sinclair, 1997 : p. 37] distingue trois familles d'évidences : «existing description of the language», «native-speaker introspections», «text, i.e. language in use». Nous avons déjà pu entériner l'utilisation de deux d'entre elles : le corpus et l'introspection. L'extrait précédent, de par le recours aux dictionnaires et lexiques pré-existants qu'il suggère, complète l'appareillage exploratoire du lexique-grammaire. Tant les corpus que les jugements introspectifs participent, nous l'avons vu, de cet impératif de reproductibilité inhérent à la démarche résolument scientifique adoptée par les linguistes du lexique-grammaire. L'utilisation de descriptions antérieures doit être comprise de la même manière. Le choix du lexique, matière première des expérimentations, doit en effet faire l'objet d'un soin tout particulier. Il n'est pas envisageable de proposer des entrées dont l'emploi, tantôt trop technique, tantôt trop littéraire, serait largement méconnu.

1.2.2 Généralisation *vs* discrétisation

Au travers de la dichotomie introspection-déduction se dessinent plus exactement deux ambitions théoriques différentes. La linguistique de corpus n'est pas, il est vrai, à proprement parler une théorie en ce sens qu'elle dresse plutôt un cadre méthodologique pouvant s'appliquer à l'étude des diverses branches de la linguistique. C'est pourquoi nous parlerons plus volontiers, dans la suite de ce travail, de linguistique descriptive comme théorie résultant de l'application, au paradigme de la linguistique, des méthodes propres aux études de corpus. Nous retrouvons, par ailleurs, ce parallèle dans la description que donne [Chomsky, 1976] de l'approche descriptive qui semble tout particulièrement se prêter à définir la linguistique telle que nous pouvons la rencontrer dans les travaux de Sinclair.

> It is the descriptivist limitation-in-principle to classification and organization of data, to 'extracting patterns' from a corpus of observed speech, to describing 'speech habits' or 'habits structure', insofar as these may exist, etc., that precludes the development of a theory of actual performance. [Chomsky, 1976 : p. 15]

Le cadre théorique de Chomsky s'oppose radicalement, quant à lui, à cette étude articulée autour de la performance. En effet, il appréhende la linguistique comme l'étude des réalités mentales qui sous-tendent TOUTE performance ([Chomsky, 1976 : p. 4]) et relèvent directement de la compétence : les universaux. Cette notion des universaux ne peut être approchée en linguistique descriptive.

> For the purposes of descriptive linguistic investigations a single LANGUAGE or dialect is considered over a brief period of time. This comprises the talk which takes place in a language community, i.e. among a group of speakers, each of whom speaks the language as a native, and may be considered an informant from the

> point of view of the linguist. [Harris, 1961 : p. 13]

En focalisant son attention sur un langage particulier, le linguiste ne peut effectivement prétendre apprécier une quelconque réalité universelle. Au contraire, dans sa quête d'universalité, Chomsky a constamment cherché à éliminer de la grammaire toute information relative à l'usage linguistique, opération qui peut s'apprécier doublement dans le cadre de la grammaire générative :

– d'une part, nous pouvons dire que Chomsky inverse le paradigme distributionnaliste de Harris. Quand ce dernier partait de l'observable – du corpus – afin d'étudier la langue et d'en extraire la grammaire, Chomsky développe son formalisme au départ de la seule introspection. Nous pourrions caricaturer en disant que le premier part du réel pour tendre vers l'abstraction quand le second part de l'abstraction pour l'appliquer ensuite au réel.
– d'autre part, nous l'avions annoncé dans notre introduction, Chomsky refuse toute intervention lexicale[10] dans la conception des règles grammaticales du composant catégoriel qu'il veut le plus général possible.

> The non-terminal vocabulary of the context-free grammar is drawn from an universal and rather limited vocabulary, some aspects of which will be considered below. The context-free grammar generates phrase-markers, with a dummy symbol as one of the terminal elements. A general principle of lexical insertion permits lexical entries to replace the dummy symbol in ways determined by their feature content. [Chomsky, 1972b : p. 12]

Cet extrait correspond parfaitement à ce que [Sinclair, 1997] désigne par *open-choice principle*. Ce principe a une implication directe sur la notion de sémantique. Chomsky voit le sens comme uniquement compositionnel refusant dans [Chomsky, 1969] toute implication de l'usage ou du contexte. Dans cette optique, le sens d'une phrase se calcule par simple addition du sens des constituants. A l'opposé, l'exploration textuelle a permis à Sinclair de dégager un second principe qui relativise la compositionnalité du sens, l'*idiom principle*.

> The principle of idiom is that a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments. [Sinclair, 1997 : p. 110]

La théorie sémantique de la grammaire générative[11], ou serait-ce plutôt l'absence d'une telle théorie, de par son obsession de la généralisation, échouera toujours à rendre compte des phrases

[10] Notons que nous arrêtons notre étude des théories chomskyennes à *Aspects* (*cf.* [Chomsky, 1976]). L'introduction de l'hypothèse lexicaliste (*cf.* [Chomsky, 1972b]) relativisera cette position qui sera complètement abandonnée dans la suite de ses travaux (programme minimaliste, théorie du gouvernement et du liage, *etc.*

[11] Nous ne désignons pas ici la sémantique générative mais bien les éléments de la théorie chomskyenne qui ressortissent à la sémantique telle qu'elle est exposée dans *Aspects* (*cf.* [Chomsky, 1976])

telles que celles-ci :

(28) Il n'y a plus de foin dans les rateliers.

(29) Elle cherche une aiguille dans une botte de foin.

(30) Max est bête à manger du foin.

La systématique du lexique-grammaire, dans son étude des propriétés verbales, a permis d'isoler trois natures sémantiques :

– verbes usuels : Mat mange de la dinde,
– verbes supports[12] : Laurence déborde d'affection,
– verbes figés : Greg a diagonalisé la matrice.

Mais, si ces trois catégories diffèrent de par leur sémantisme, leurs propriétés syntaxiques sont comparables. C'est très certainement ce constat qui doit être compris comme origine de l'idée générativiste. Si nous ne pouvons nier une prise de position sémantique progressive, Chomsky ne semble pas y porter un grand intérêt dans [Chomsky, 1976], n'y consacrant qu'une note et renvoyant simplement aux travaux de Katz et Postal. L'hypothèse lexicaliste qui sera introduite dans la Théorie standard étendue (EST : Extended Standard Theory) jette les premières bases d'un réel composant sémantique. Toutefois, nous n'avons pas poussé notre analyse de la grammaire générative au-delà de la rupture que marque [Gross, 1979] et limitons donc notre propos à la théorie chomskyenne d'*Aspects*.

[Gross, 1988], dans son étude syntaxique des trois catégories verbales mentionnées ci-dessus, identifie les propriétés communes suivantes :

– identité du transformationnalisme : les figements, autant que les verbes supports et usuels, acceptent au même titre l'hypothèse transformationnelle. Ainsi, nous pourrons observer, le cas échéant, la passivation ou encore la nominalisation au sein des trois catégories.
– identité structurelle : l'étude lexicale systématique a permis de limiter la succession de compléments, si ce n'est quelques rares exceptions, à deux quel que soit le type du verbe. Les études du lexique-grammaire ont donc pu circonscrire la structure de la phrase élémentaire à la séquence suivante, où *Prep*, en tant que désignant une préposition, peut équivaloir à *zéro* :

$$(\varepsilon + N_0)V(\varepsilon + (Prep)N_1)(\varepsilon + (Prep)N_2)$$

Nous pouvons donc, en considérant encore la similitude fonctionnelle et de conjugaison, prétendre à une certaine constance grammaticale. De plus, et bien que certains éléments lexicaux

[12] Nous avons déjà pu mettre en évidence la neutralité relative du sémantisme des verbes de ce type dont la seule fonction est de supporter l'élément (substantif ou adjectif) qui joue le rôle de prédicat.

soient propres au figement[13], nous devons constater que les trois types verbaux partagent le même fonds lexical comme nous pouvons l'observer dans les trois actualisations du verbe *porter* :

– USUEL : Max porte une caisse.
– SUPPORT : Max porte de l'affection à Lisa.
– FIGÉ : Max porte le chapeau.

Dès lors, nous ne pouvons pas généraliser le principe idiomatique à l'ensemble de la langue. Malgré cela, il est entendu que la notion de figement est la preuve de la nécessité d'une discrétisation syntaxico-sémantique et une nouvelle justification de l'étude systématique du lexique-grammaire. Nous ne pouvons, en effet, abonder dans le sens de Chomsky qui, négligeant l'enjeu lexical, généralise à outrance. Toutefois, il ne convient pas non plus de sombrer dans l'extrémisme inverse. Pour reprendre [Gross, 1988 : p. 22], «les phrases figées ne sont qu'exceptionnellement entièrement figées», presque toutes peuvent être soumises à une analyse syntaxique régulière même si nous devons bien admettre que les figements semblent plus nombreux que les phrases libres.

La double dialectique que nous venons de présenter (*cf.* Sections 1.2.1 et 1.2.2) se traduit par la synthèse du formalisme transformationnel inhérent à la grammaire générative et de l'inductivisme propre aux études de corpus. Il en résulte une linguistique descriptive systématique qui, nous l'avons vu, se prête tout particulièrement au paradigme du lexique-grammaire et rejoint très largement la conception grammaticale harrissienne.

> La théorie grammaticale consiste en deux opérations. La première (chapitre II) construit en deux étapes le sous-ensemble des phrases de concaténation, qui est un langage fermé contenant toute l'information véhiculée par la langue prise dans sa totalité. On verra que ces phrases présentent des inégalités de vraisemblance d'occurrence. La seconde (chapitre III) construit les autres phrases de la langue, en opérant certaines variations de forme qui paraphrasent les phrases de stricte concaténation. [Harris, 1976b : p. 23]

Dès lors, si l'impact des théories de Harris est indéniable et par ailleurs, nous l'avons dit, avoué, nous pouvons conclure en l'influence conjuguée de la linguistique théorique, au sens chomskyen, et de la linguistique descriptive en tant que conduite au départ du réel.

[13] *Escampette*, par exemple, n'apparaît qu'au sein de la locution verbale *prendre la poudre d'escampette*.

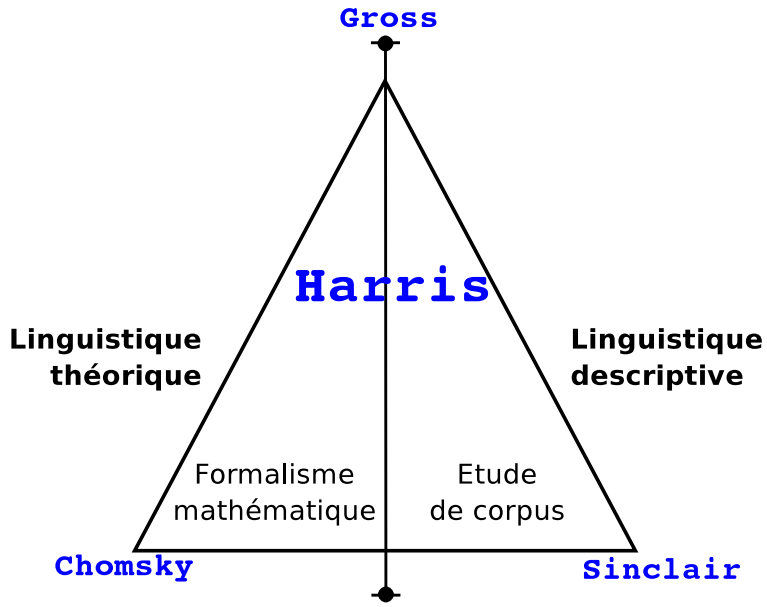


Figure 1.3 : Le lexique-grammaire comme dialectique linguistique

1.3 Matrices, automates et grammaires locales

La prétention théorique du lexique-grammaire est indéniable autant que les visées informatiques rendues directement envisageables par le caractère binaire de la représentation matricielle. En marge des applications informatiques et d'un cadre théorique adapté aux études linguistiques, l'exploration lexicale systématique, base de l'implémentation d'un lexique-grammaire, rend possible, selon nous, une approche plus didactique du formalisme.

[Leclère, 1998] rassemble un vaste ensemble de publications relativement récentes et s'inscrivant dans la dynamique créée autour du lexique-grammaire. Nous ne donnerons qu'un aperçu partiel de ce domaine applicatif et renvoyons à cet article pour plus d'informations.

1.3.1 Applications linguistiques

La diversité des études théoriques s'apprécie tant du point de vue des sujets abordés que des langues mises en œuvre. Si le français reste, à l'heure actuelle, la langue dont la couverture lexicale est la plus aboutie, le formalisme s'est propagé à l'ensemble du domaine roman, mais également, même si la progression semble plus lente, aux langues non romanes. De plus, des études ont été menées, au sein d'une langue, autour de ses différentes variantes régionales.

Le cadre théorique qu'impose le lexique-grammaire suppose la combinaison des niveaux lexical et syntaxique dont l'étude s'organise au départ de la phrase élémentaire. De même, l'outillage formel utilisé reste identique quelle que soit la langue. Dès lors, l'abstraction descriptive qui résulte de ces études autorise, et c'est là un des principaux avantages linguistiques du lexique-grammaire, les approches contrastives.

Les études lexicales, tout d'abord dédiées aux mots simples, ont très vite été étendues aux combinaisons de mots. Les structures prédicatives à verbes supports, les séquences figées et les mots composés, qui présentent un sémantisme non compositionnel, demandent en effet à être traités comme autant d'unités lexicales et font l'objet d'entrées uniques à l'intérieur du formalisme. Nous constatons donc certaines considérations sémantiques à l'origine des choix qui se posent dans la création d'un lexique-grammaire. Nous avons également mentionné dans notre travail certains critères sémantiques visant à contraindre l'environnement des prédicats quand les traits syntaxiques ne suffisent pas. La sémantique, telle qu'elle est envisagée ici, est donc essentiellement destinée à soutenir l'étude syntaxique ; il n'est pas, à proprement parler, d'*études* sémantiques.

Cette dernière constatation est très certainement à l'origine des travaux de Gaston Gross qui, en parallèle à la classification syntaxique, opère une hiérarchisation sémantique. Les *classes d'objets* (*cf.* [G. Gross, 1994]) qui procèdent de ses recherches listent les substantifs en fonction de leur propension à apparaître dans certains contextes syntaxiques. Ces contextes décrivent, pour chaque classe, un champ sémantique (*cf.* Section 2.4). Nous voyons dans ces travaux un complément précieux aux tables du lexique-grammaire.

1.3.2 Applications didactiques

Nous ne pouvons faire état ici de l'existence d'une large littérature sur le sujet contrairement aux deux autres domaines d'application. Toutefois, [Leclère, 1998] recense quelques travaux pionniers qui annoncent une certaine ambition pédagogique. En effet, le lexique-grammaire en tant qu'étude systématique de la langue (et par conséquent exhaustive) offre une source d'informations particulièrement intéressante pour l'apprentissage. Ainsi, bien que sa conception ressortisse plus exactement au traitement automatique de la langue, la correction automatique, ou seulement assistée, nous apparaît comme directement implémentable. De même, nous pourrions envisager le reformatage des tables afin de les présenter d'une manière plus ergonomique et de permettre la consultation des constructions syntaxico-sémantiques autorisées par l'usage[14].

Ces quelques pistes ne circonscrivent pas les perspectives qu'un lexique-grammaire peut offrir à l'enseignant comme à l'apprenant mais annoncent l'ébauche d'une réelle recherche didactique.

[14]Cette idée a déjà vu le jour dans un contexte théorique différent, mais proche, du lexique-grammaire. Les projets DAFA (Dictionnaire d'apprentissage du français des affaires) et DAFLES (Dictionnaire d'apprentissage du français langue étrangère ou seconde) visent à mettre à disposition de l'apprenant certaines informations syntaxiques. Ainsi, le lexicographe a recensé, pour chaque entrée lexicale, les structures de phrases élémentaires et un exemple qui les actualise (`http://www.kuleuven.ac.be/dafles/`).

1.3.3 Applications informatiques

L'orientation linguistique du lexique-grammaire le destine tout naturellement à l'analyse syntaxique automatique. La reconnaissance de motifs syntaxico-sémantiques connaît effectivement une grande tradition au sein du réseau RELEX (ensemble de laboratoires organisés autour de cette idée d'une implémentation informatique des théories linguistiques du lexique-grammaire). Notre étude l'envisage également dans le contexte particulier de l'extraction d'information. Nous serons donc amené à revenir longuement sur cette notion et nous limiterons la présentation que nous en faisons ici au formalisme général, les automates à états finis.

> D'une manière générale, les lexiques et grammaires électroniques construits au LADL sont des fonctions caractéristiques de mots ou de phrases. En d'autres termes, ce sont des automates ou grammaires qui engendrent ou acceptent les formes observées et uniquement celles-ci. [Gross, 1989 : p. 14]

Automates et transducteurs

Un **automate** est un mécanisme mathématique abstrait permettant de *reconnaître* une séquence. Dans le contexte d'une langue naturelle, un automate permet de dire si une phrase ou une séquence d'entrée appartient à la langue.

Pratiquement, on définit un automate comme un tuple de 5 éléments :

$$A = (\Sigma, Q, e_0, F, \delta)$$

Où,

- Σ est un ensemble fini de symboles constituant l'alphabet d'entrée,
- Q est un ensemble fini d'états,
- $e_0 \in Q$ est l'état initial,
- $F \subseteq Q$ est l'ensemble des états terminaux,
- δ est la fonction de transition qui associe une sortie à chaque couple (e_x, y) où $e_x \in Q$ et $x \in \Sigma$.

Les différents états et transitions sont habituellement représentés sous la forme d'un réseau de transitions simples où les états sont symbolisés par les nœuds et les transitions par les arcs qui lient deux nœuds entre eux. La procédure de reconnaissance se déroule comme suit : après chaque transition, l'automate lit un élément de la séquence à reconnaître ; si le dernier élément de la séquence mène vers un état final (nœud entouré de deux cercles), la séquence est *reconnue*. Le réseau de la figure 1.4 permet l'acceptation de suites telles que *ac*, *abac*, *ababac*, *etc.* mais rejettera les suites *bc abc*, *etc.* Notons que nous pouvons également représenter ce type de mécanisme au moyen d'une table de transitions qui indique l'état à atteindre pour chaque couple `(état, transition)` autorisé par le formalisme. Nous obtiendrons alors, pour le graphe de la figure 1.4, la table 1.2.

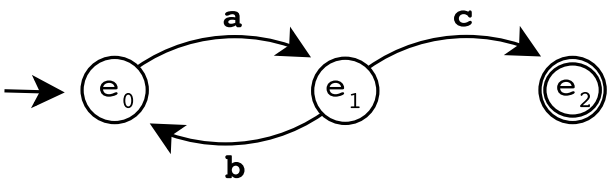


Figure 1.4 : Automate à états finis élémentaire

	a	b	c
e_0	e_1	$\emptyset$	$\emptyset$
e_1	$\emptyset$	e_0	e_2
e_2	$\emptyset$	$\emptyset$	$\emptyset$

Tableau 1.2 : Table de transitions

Les automates peuvent s'utiliser tant en génération qu'en reconnaissance. Toutefois, selon l'option choisie, certains choix peuvent s'avérer importants.

> Un réseau est dit déterministe si, pour chaque couple (e_i, a), la fonction $\delta(e_i, a)$ détermine une valeur unique. Autrement dit, un réseau de transitions est déterministe si, pour chaque état du réseau, les arcs quittant cet état portent tous une étiquette distincte. [Wehrli, 1997 : p. 97]

En génération, un automate non déterministe ne pose aucun problème ; l'énumération de l'ensemble des séquences qui constituent les phrases du langage reste possible. La reconnaissance, par contre, nécessitera l'adjonction d'une fonction supplémentaire de **marche-arrière** (*backtracking*) afin d'éviter que le choix d'une transition bloque l'automate dans un état sans issue (*cf.* Figure 1.5).

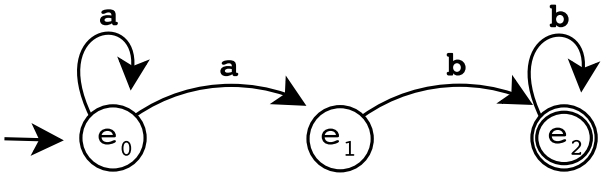


Figure 1.5 : Automate à états finis non déterministe

Il est toutefois possible de déterminiser ce type d'automate. Ainsi l'automate de la figure 1.5, non déterministe, donnera, après déterminisation, l'automate de la figure 1.6.

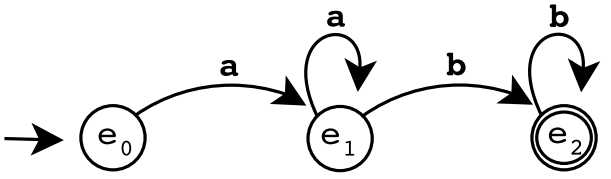


Figure 1.6 : Automate après déterminisation

Un **transducteur** est un cas particulier de reconnaisseur qui associe une information (ou **sortie**) à une unité préalablement reconnue. Il s'agit donc d'un 6-tuple :

$$A = (\Sigma_1, \Sigma_2, Q, e_0, F, \delta)$$

Où,

- Σ_1 est un ensemble fini de symboles constituant l'alphabet d'entrée,
- Σ_2 est un ensemble fini de symboles constituant l'alphabet de sortie,
- Q est un ensemble fini d'états,
- $e_0 \in Q$ est l'état initial,
- $F \subseteq Q$ est l'ensemble des états terminaux,
- δ est la fonction de transition qui associe une sortie à chaque couple (e_x, y) où $e_x \in Q$ et $x \in \Sigma$.

Ce mécanisme est directement pertinent pour le traitement automatique des langues. Il permet, en effet, d'associer à une séquence reconnue, une sortie qui peut, par exemple, préciser la nature grammaticale d'un mot ou la fonction d'un syntagme.

Unitex et les grammaires locales

La procédure traditionnelle de l'analyse syntaxique consiste en la reconnaissance pas à pas (menée de droite à gauche ou de gauche à droite selon l'implantation) d'un contexte grammatical. Une phrase est parcourue par une tête de lecture qui concatène, au fur et à mesure, les séquences identifiées afin de reproduire l'organisation en syntagmes de la phrase initiale.

Les analyseurs construits au départ des tables du lexique-grammaire procèdent, quant à eux, d'une idée entièrement nouvelle. Rappelons que l'organisation en tables obéit à deux principes :

- d'une part, chaque entrée lexicale est contextualisable sous la forme d'une phrase élémentaire *sujet-verbe-compléments essentiels*,
- d'autre part, les transformations qui leur sont associées permettent la construction de l'ensemble des variations possibles.

Nous pouvons dès lors représenter, sous forme d'automate, la grammaire associée à chaque lexie et ainsi couvrir exhaustivement la créativité linguistique d'une langue. L'analyse syntaxique proprement dite consiste alors en la simple reconnaissance de la phrase d'entrée[15]. *Unitex*[16]

[15] Nous ne pouvons bien entendu envisager de réaliser ce travail manuellement, les tables verbales comptant déjà à elles seules plus de 15 000 entrées. [Roche, 1993], [Senellart, 1999] et [Paumier, 2003] évoquent à ce propos la méthodologie des *graphes patrons* ou *graphes paramétrés*, sur laquelle nous reviendrons plus longuement au cours de ce travail (*cf.* Section 3.3). Elle consiste sommairement à associer à chaque table un automate de référence qui formalise l'ensemble des propriétés distributionnelles et transformationnelles de la table. Une procédure automatique permet ensuite de générer un automate particulier pour chaque entrée de la table.

[16] *Unitex* est une réimplémentation *open-source* du système *Intex* (*cf.* [Silberztein, 1993]). En marge des possibilités de ce dernier, il intègre de nouvelles fonctionnalités dont l'*Unicode* qui permet, sans traitement préalable, l'analyse de langues, telles que l'arabe ou le grec, qui ne présentent pas un alphabet latin. De plus, la mise à disposition du *code source* rend possible une collaboration à très grande échelle qui lui confère une vitalité nécessairement plus importante. Ce sont ces deux considérations qui nous ont amené à privilégier l'utilisa-

offre en ce sens un cadre de travail très intéressant et accessible au linguiste non-informaticien (*cf.* [Paumier, 2003]). Il permet la formalisation graphique des automates sous la forme de *grammaires locales* (*cf.* Figure 1.7[17]) et offre un nombre considérable d'outils permettant leur application au texte. Nous renvoyons, pour une description complète de ces outils, au manuel d'*Unitex* (*cf.* [Paumier, 2006]).

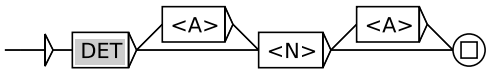


Figure 1.7 : GN minimal

Outilex et l'unification

Outilex (*cf.* [Blanc *et al.*, 2006][18]) s'inscrit dans le même formalisme qu'*Unitex* (*e.g.* lexique-grammaire et automates). Il ajoute toutefois certaines facilités qui viennent enrichir ce formalisme : les *structures de traits* et l'*unification*.

> On définit des traits, c'est-à-dire des couples attribut-valeur, les valeurs pouvant être des symboles atomiques ou des traits. Des ensembles de traits (conjoints) peuvent être associés à un mot ou à un syntagme. Ils sont appelés structures fonctionnelles en LFG, catégories en GPSG ou signes en HPSG. Ils peuvent se présenter comme des matrices ou comme des graphes orientés sans cycles (l'arc portant l'attribut et pointant vers sa valeur). [Abeillé, 2002 : pp. 33-34]

L'unification permet, quant à elle, de *réunir*, en une nouvelle structure, deux structures de traits dont les composantes sont compatibles. La nouvelle structure contiendra l'ensemble des traits communs aux deux structures sources. Pratiquement, ce processus nous permet de contraindre l'analyse. Appliqué à l'automate de la figure 1.8[19], par exemple, elle nous permettra de refuser

---

tion d'*Unitex*. Ce logiciel est téléchargeable depuis le site de l'*Institut Gaspard-Monge*, promoteur du logiciel : `http://www-igm.univ-mlv.fr/~unitex/`

[17] Au sein de la figure 1.7, `<N>` désigne un nom commun et `<A>`, un adjectif. La boite grisée `DET` renvoit à un sous-graphe. Un sous-graphe peut en contenir un autre.

[18] *Outilex* n'est pas encore disponible pour le public. Les contacts privilégiés que nous entretenons avec le laboratoire d'informatique-linguistique de l'*Institut Gaspard-Monge*, en général, et Olivier Blanc (développeur principal de l'analyseur syntaxique associé à la plate-forme logicielle *Outilex*), en particulier, nous ont permis de profiter de cet outil. La plupart des développements proposés dans ce travail n'aurait pas été possible hors de cette collaboration. Le lecteur, curieux, trouvera certaines informations sur la page associée au projet : `http://www.at-lci.com/Outilex/Outilex.html`.

[19] Les différences entre les formats *Unitex* et *Outilex* sont minimes. Nous pouvons d'ailleurs utiliser l'éditeur de graphes du premier afin de développer les automates du deuxième. Il existe toutefois quelques différences de notation. Dans la figure 1.8, la balise `<det>` désigne un déterminant quelconque (`<DET>`, dans *Unitex*) ; `<noun>`, un nom commun (`<N>`, dans *Unitex*) ; et `<adj>`, un adjectif (`<A>`, dans *Unitex*). Associés à ces balises, les couples *attributs–valeurs* `gender = $$.gender` et `number = $$.number` forment la structure de traits de chaque élément. Le processus d'unification vérifie que la valeur des attributs `gender` et `number` est identique pour tous les éléments de l'automate.

les analyses qui ne partagent pas les mêmes valeurs de genre (*gender*) et de nombre (*number*). Nous ajoutons ainsi une validation morpho-syntaxique à la reconnaissance. Cette combinaison se traduit par une réduction des analyses possibles pour une séquence donnée. Notons, également, qu'*Outilex* permet la pondération de certaines séquences au sein d'un automate, favorisant ces parcours dans le processus d'analyse.

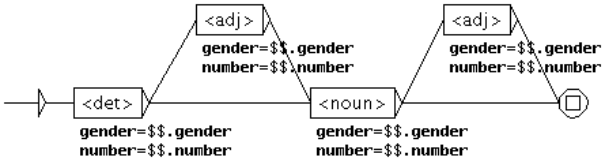


Figure 1.8 : GN minimal et structures de traits

Nous avons décrit, très brièvement, le formalisme de même que deux des outils que nous exploitons dans ce travail (*Unitex* et *Outilex*). Nous aurions pu largement nous étendre sur leur fonctionnement mais, il nous semble plus intéressant d'envisager leur approfondissement dans un contexte applicatif particulier. Nous poursuivrons donc leur description tout au long de ce travail, en fonction de l'utilisation que nous en ferons.

1.4 Conclusion

L'énorme puissance descriptive du modèle linguistique initié par Maurice Gross répond, nous en avons l'intime conviction, à la plupart des attentes tant linguistiques qu'informatiques que nous pourrions formuler au cours de l'étude d'une langue naturelle. Toutefois, un modèle formel comme le lexique-grammaire nécessite le recours à certaines approximations qui impliquent inévitablement certaines limites :

1. Les tables du lexique-grammaire présentent l'emploi des prédicats dans leur forme canonique. Cependant, en situation réelle d'énonciation telle que nous pouvons l'observer en corpus, nous rencontrons des incises, des inversions stylistiques, *etc.*, autant de situations qui nécessitent un traitement particulier. Certaines d'entre elles ont d'ailleurs fait l'objet d'études lexico-grammaticales approfondies.
2. Les tables du lexique-grammaire ne traitent que les compléments dits essentiels. Nous ne retrouvons pas les compléments circonstanciels non essentiels[20].
3. Certains compléments essentiels peuvent, dans certaines sous-structures d'une structure syntaxique donnée, disparaître.
4. Finalement, le lexique-grammaire vise, avant tout, la description de la langue standard

[20][Gross, 1986a], dans sa définition du complément adverbial, rassemble les adverbes au même titre que les compléments circonstanciels et les propositions subordonnées circonstancielles. Nous opposons à ce complément adverbial étendu (*cf.* Exemples 17 et 18) les compléments circonstanciels *non essentiels* (*cf.* Exemples 21 et 22).

sans considération pour d'éventuelles spécialisations propres à un domaine particulier. C'est un point que nous développerons de manière plus conséquente dans le cadre de l'extraction d'information.

Ceci étant, au vu de ces quelques problèmes, qui ne sont pas à proprement parler insurmontables, nous ne croyons pas pouvoir parler de failles. Nous pensons que le modèle initial peut être complété facilement afin de répondre à ces limites. Les classes d'objets sont par ailleurs une tentative en ce sens.

Le reste de notre travail visera l'implémentation des concepts du lexique-grammaire au sein d'une application d'extraction d'information. Nous présentons cette discipline au chapitre suivant au terme duquel nous jetterons les bases d'un système nouveau qui profite de la puissance descriptive du lexique-grammaire.

Deuxième partie

Le lexique-grammaire au service de l’extraction d’information

CHAPITRE 2

EXTRACTION D'INFORMATION : ÉTAT DE LA QUESTION

La complexité des systèmes d'information a augmenté au même rythme que les données se sont elles-mêmes multipliées. Aujourd'hui, l'essor du support électronique, et plus particulièrement d'Internet, a conduit à un chaos informationnel indéniable. La croissance d'Internet est à ce point vertigineuse que nous ne pouvons plus, au départ d'une simple recherche, espérer toucher à l'essentiel. En avril 2003, *Netcraft*[1] recensait 40 000 000 de sites contre 92 615 362 en août 2006. Le nombre a donc plus que doublé en l'espace de seulement trois ans[2].

Figure 2.1 : Recherche simple sur Google

Cette situation a conduit la recherche d'information à s'enrichir de diverses routines visant à garantir une certaine pertinence au résultat. Dans cette optique, la plupart des moteurs de recherche opère un archivage préalable des données qui reflète d'une part la thématique de la source et, d'autre part, son degré d'autorité (une source massivement référencée sera considérée comme plus pertinente que les autres et présentée en priorité à l'utilisateur). De même, nous disposons d'interfaces permettant de contraindre la recherche en en précisant l'étendue mais

[1] `http://news.netcraft.com`

[2] Les figures 2.1 et 2.2 donnent une illustration «grand public» du problème mentionné. Notons que les saisies d'écran que nous proposons datent de 2003. À cette époque, les mots-clés *extraction d'information* se retrouvaient dans 51 000 documents. Actuellement (août 2006), cette même requête conduit à 1 280 000 résultats.

surtout les limites (*cf.* Figure 2.2).

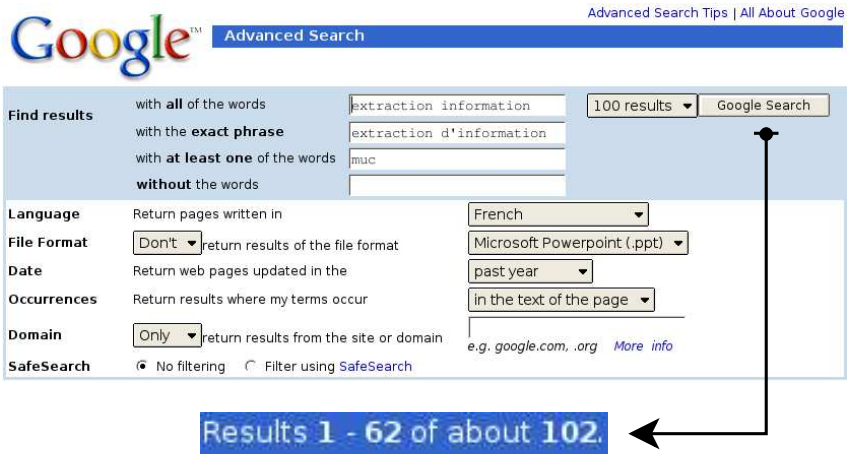


Figure 2.2 : Recherche avancée sur Google

Au-delà de la seule recherche, le boom électronique qui a marqué ces deux dernières décennies a introduit de nouvelles perspectives dans le traitement de l'information : le filtrage d'information organise les documents en une arborescence de domaines et sous-domaines ; le résumé automatique permet, quant à lui, de donner un aperçu plus ou moins global des documents archivés ; *etc.* C'est dans ce contexte que se sont inscrites les premières idées d'une extraction d'information.

Organisés jusqu'alors autour du nécessaire classement des documents, les systèmes d'information se sont peu à peu intéressés à l'organisation de leur contenu. Ainsi, ont émergé des projets qui tendent à organiser l'information en champs distincts, soit en fonction du sujet qu'elle traite (*cf.* Figure 2.3, A.), soit autour des acteurs qu'elle met en scène – lieux, personnes, organisations, etc. – (*cf.* Figure 2.3, B.).

Figure 2.3 : Extracteur automatique d'actualités de Google

2.1 Des conférences à l'origine de l'extraction : MUC et ACE

Comme c'est le cas pour beaucoup de disciplines, la procédure d'évaluation qui vise à juger les systèmes d'extraction a dans une large mesure participé à la définition de la discipline elle-même.

De 1987 à 1998, c'est-à-dire de la première conférence organisée par la *Message Understanding Conference* (MUC) à la septième et dernière en date, la procédure d'évaluation est restée fondamentalement la même et s'articule en trois temps. La première étape spécifie la tâche à effectuer. Une fois la tâche définie, les participants reçoivent un programme d'évaluation automatique et un ensemble de documents pour lesquels les différentes informations à extraire sont spécifiées. Ces éléments doivent leur permettre tantôt d'apprendre un nouveau système, si la méthode est statistique, tantôt de développer une grammaire d'extraction. Après une période pouvant couvrir un à six mois, les participants reçoivent un nouvel ensemble de documents (*cf.* Tableau 2.1 [Appelt *et al.*, 1993]). Ceux-ci sont soumis aux différents systèmes et les résultats obtenus sont ensuite renvoyés afin d'être évalués par comparaison aux données extraites manuellement.

San Salvador, 19 Apr 89 (ACAN-EFE) Salvadoran President-elect Alfredo Cristani condemned the terrorist killing of Attorney General Roberto Garcia Alvarado and accused the Farabundo Marti National Liberation Front (FMLN) of the crime. [...] Garcia Alvarado, 56, was killed when a bomb placed by urban guerillas on his vehicle exploded as it came to a halt at an intersection in downtown San Salvador. [...] Vice President-elect Francisco Merino said that when the attorney general's car stopped at a light on a street in downtown San Salvador, an individual placed a bomb on the roof of the armored vehicle. [...] Guerrillas attacked Merino's home in San Salvador 5 days ago with explosives. There were seven children, including four of the vice president's children, in the home at the time. A 15-year-old niece of Merino's was injured. [...] According to the police and Garcia Alvarado's driver, who escaped unscathed, the attorney general was traveling with two bodyguards. One of them was injured.

Tableau 2.1 : Extrait d'un texte issu de MUC-3

La nature des tâches a, de son côté, beaucoup évolué depuis la première conférence, précisant dans le même temps les objectifs de la discipline. Tout d'abord articulée autour de la compréhension de texte (MUC-1 et MUC-2), la *Message Understanding Conference* a, depuis sa troisième manifestation, focalisé son attention sur le remplissage de formulaire (*cf.* Tableau 2.2 [Appelt *et al.*, 1993]).

Nous retrouvons, depuis 1999, les tâches associées à l'extraction d'information dans les campagnes d'évaluation ACE qui ont succédé à la défunte MUC et existent encore actuellement. Ces tâches se sont complexifiées et couvrent un nombre plus important de situations : la détection et la reconnaissance des **entités**, des **valeurs**, des **expressions temporelles**, des **relations** et des **évènements** (*cf.* [ACE, 2006]). Notons qu'elles correspondent précisément aux différents éléments intervenant dans le remplissage d'un formulaire. Nous les envisagerons donc au sein

de cette procédure que nous illustrerons au départ de la dépêche financière et du formulaire présentés dans le tableau 2.3.

Date :	19 Apr 89
Location :	El Salvador : San Salvador
Incident Type :	Bombing
Perpetrator ID :	"urban guerillas"
Perp. Org. :	"FMLN"
Confidence :	Suspected or Accused by authorities : "FMLN"
Physical Target :	"vehicle"
Effect :	Some Damage : "vehicle"
Human Target :	"Roberto Garcia Alvarado"
Description :	"attorney general" : "Roberto Garcia Alvarado" "driver" "bodyguards"
Effect :	Death : "Roberto Garcia Alvarado" No Injury : "driver" Injury : "bodyguards"

Tableau 2.2 : Exemple de formulaire issu de MUC-3

[Grishman, 1997 : p. 14] associe à la procédure d'extraction deux niveaux d'analyse consécutifs : l'*analyse locale* qui procède à l'extraction des séquences pertinentes en regard des champs du formulaire et l'*analyse du discours* qui réalise l'intégration des résultats.

AEGON VEUT SE RECENTRER SUR L'ASSURANCE VIE

L'assureur allemand, Aegon, a annoncé qu'il va vendre les activités de crédit de sa filiale américaine, Transamerica Finance, au conglomérat industriel General Electric, pour 5,4 milliards de dollars. Ainsi, le groupe compte à terme se désengager de sa filiale, ce qui est conforme à sa stratégie de recentrage sur son cœur de métier qu'est l'assurance vie.

05/08/03 10 :07

FORMULAIRE

VENDEUR	NOM	
	DESCRIPTION	
ACQUÉREUR	NOM	
	DESCRIPTION	
TRANSACTION	OBJET	
	DESCRIPTION	
	DÉTAILS	
MONTANT		

Tableau 2.3 : Illustration du scénario des transactions financières

2.1.1 Analyse locale

L'analyse lexicale

Après avoir effectué la segmentation en phrases, l'extracteur délimite les frontières des mots et assigne à chacun d'eux une catégorie grammaticale utilisée ensuite lors de l'analyse syntaxique. L'analyse lexicale inclut également la reconnaissance des entités, des valeurs et des expressions temporelles.

Par *entité*, nous désignons principalement les entités nommées, c'est-à-dire les noms propres (noms de personne, d'organisation, de lieux, *etc*), mais également (depuis ACE-1) un ensemble d'autres séquences spécifiques à l'un ou l'autre scénario et renvoyant, par exemple, à des infrastructures, des véhicules ou des armes.

De même que les entités, les valeurs et les expressions temporelles peuvent également intervenir au niveau des arguments d'un évènement et doivent donc être identifiées (*cf.* Tableau 2.4). Une valeur peut être propre à un scénario d'extraction particulier (délit, peine, titre de fonction) ou indépendante du contexte (informations de contact, valeur numérique). Les expressions temporelles appellent un traitement particulier en ce qu'elles peuvent être de type (absolu ou relatif) et de nature (durée, ancrage temporel) différents. Il faut donc envisager leur normalisation afin de pouvoir *utiliser* l'information qu'elles véhiculent.

Pour l'identification de ces structures, on utilise habituellement une succession d'expressions régulières dont l'application est déclenchée par certains mots-clés[3] :

– un titre (*M.*, *Dr*, *Mme*, *etc.*), un prénom, un suffixe (*Jr*, *Sr*, *etc.*) ou encore une initiale centrale pour les entités nommées,
– un chiffre ou une unité pour les valeurs, et
– une année, un mois, un jour ou une structure adverbiale de temps, pour les expressions temporelles.

De nombreuses études ont été menées en ce sens et l'extraction des entités nommées, pour ne citer qu'elle, obtient d'excellentes performances (plus de 92% en rappel et 95% en précision, *cf.* [Maurel et Friburger, 2002; Poibeau, 2003; Fairon et Watrin, 2003; Kevers, 2006]).

ENTITÉS	Aegon Transamerica Finance General Electric
VALEURS	5,4 milliards de dollars
EXPRESSIONS TEMPORELLES	Jour : 05 Mois : 08 Année : 2003 Heure : 10 :07

Tableau 2.4 : Analyse lexicale

L'analyse syntaxique

L'analyse syntaxique (*cf.* Figure 2.4) est responsable de l'identification des relations que nous pouvons observer entre deux entités au moins. Le plan d'évaluation ACE distingue relations

[3]Dans le cadre des entités nommées, [McDonald, 1996] distingue les *indices externes* des *indices internes*, c'est-à-dire les mots-clés internes (*e.g. Ltd.* ou *G.m.b.H.* pour les noms d'entreprises) et externes (*e.g. the company X*) à la séquence à reconnaître.

symétriques (pour lesquelles l'ordre des entités n'est d'aucune importance) et asymétriques. Ces dernières doivent également assigner un rôle à chaque entité.

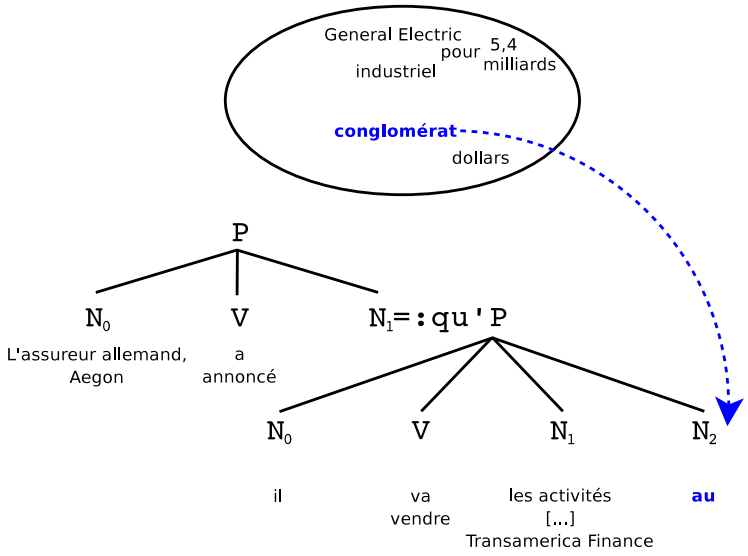


Figure 2.4 : Analyse syntaxique

La profondeur de l'analyse reste, encore actuellement, une question fondamentale. Doit-on se limiter à la reconnaissance des têtes de syntagmes ou, au contraire, analyser chaque élément de la séquence ? Bien qu'une analyse de surface puisse occasionner la perte de certains fragments d'information, l'analyse profonde est souvent jugée par trop coûteuse, la plupart des informations traitées en pareil cas n'étant aucunement pertinentes pour le remplissage des champs du formulaire (*cf.* Section 2.2).

2.1.2 Analyse du discours

Identification des évènements

Un évènement définit un scénario d'extraction (transaction financière, conflit armé, arrestation, *etc.*). Il est donc composé d'une ou plusieurs relations et d'au moins deux arguments pour lesquels un rôle précis doit être assigné. Les attributs d'un évènement sont le type (*i.e.* financier), le sous-type (*i.e.* transaction), la modalité, la polarité et le temps. Il est habituellement représenté sous la forme d'un motif ou patron d'extraction (*cf.* Figure 2.5)

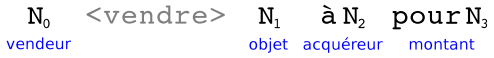


Figure 2.5 : Identification des motifs d'extraction

Résolution des coréférences

Les différents acteurs d'un évènement de même que les relations mises en œuvre peuvent être désignés de multiples façons au sein d'un texte (*cf.* Tableau 2.5). Il convient, toutefois, de considérer cette diversité comme les actualisations d'une unique réalité afin de ne pas présenter une information fragmentée ou redondante.

Aegon	assureur allemand il le groupe
Transamerica Finance	sa fililale américaine sa filiale
General Electric	conglomérat industriel

Tableau 2.5 : Résolution des coréférences

Inférence et rassemblement

Un unique fait peut être disséminé sur plusieurs phrases. Il est donc nécessaire de combiner l'ensemble des éléments afin d'exposer le fait complet (*cf.* Tableau 2.6). De même, l'information peut être implicite et doit alors être inférée d'autres séquences textuelles.

VENDEUR	NOM	Aegon
	DESCRIPTION	assureur allemand
ACQUÉREUR	NOM	General Electric
	DESCRIPTION	conglomérat industriel
TRANSACTION	OBJET	Transamerica Finance
	DESCRIPTION	filiale américaine
	DÉTAILS	activités de crédit
MONTANT		5,4 milliards de dollars

Tableau 2.6 : Formulaire complet

Chacune des tâches présentées ci-dessus donne lieu à une évaluation séparée. Pour juger leur qualité, les conférences MUC utilisent la précision et le rappel[4]. Le rappel estime la couverture de l'extracteur (*i.e.* le rapport entre le nombre de séquences extraites et le nombre de séquences à extraire) et la précision, la pertinence des séquences extraites (*i.e.* le rapport entre le nombre d'extractions correctes et le nombre de séquences extraites).

$$\mathbf{R} = \frac{x}{x+y}$$

[4]En complexifiant les tâches, les conférences ACE ont également complexifié les mesures qui ne distinguent plus à présent le rappel de la précision. La valeur de chaque élément (entité, relation, *etc.*) n'est plus binaire (extrait ou non) mais intègre la reconnaissance des coréférences, la caractérisation en types, sous-types et classes, *etc.* Nous n'envisagerons pas ces ajouts dans ce travail. Une description complète de ces mesures peut toutefois être téléchargée depuis le site du NIST (*National Institute of Standards and Technology*), promoteur de la conférence (`http://www.nist.gov/speech/tests/ace/`).

$$\begin{cases} \mathbf{x} = Nbre\ d'éléments\ extraits, \\ \mathbf{y} = Nbre\ d'éléments\ manquants, \\ \mathbf{x} + \mathbf{y} = Nbre\ total\ d'éléments. \end{cases}$$

$$\mathbf{P} = \frac{x}{x+z}$$

$$\begin{cases} \mathbf{x} = Nbre\ d'éléments\ extraits\ corrects, \\ \mathbf{z} = Nbre\ d'éléments\ extraits\ erronés \\ \mathbf{x} + \mathbf{z} = Nbre\ total\ d'éléments\ extraits. \end{cases}$$

Le rappel et la précision sont inversément proportionnels ; la précision diminue au rythme où le rappel augmente (et inversément). Selon le choix d'implémentation et l'objectif poursuivi, nous privilégierons l'un ou l'autre. Ainsi, par exemple, si l'extraction est un préalable à une indexation intelligente des documents, la précision doit être favorisée afin de minimiser les erreurs potentielles qui viendraient altérer, par la suite, la qualité de la recherche.

Notons finalement que, pour estimer l'efficacité globale d'un système, sans considérer indépendamment le rappel et la précision, la recherche et l'extraction d'information utilisent la mesure F :

$$\mathbf{F} = \frac{2 \times (P \times R)}{(P+R)}$$

2.2 Une analyse locale entre recherche et compréhension

Bien que l'usage commun mélange volontiers les deux notions, il convient, au vu de ce que nous venons d'exposer, de distinguer la recherche de l'extraction d'information. Les progrès tant théoriques que techniques de ces dernières années ne permettent plus d'établir la distinction en termes de complexité des fonctions langagières mises en œuvre ou encore de profondeur d'analyse. Toutefois, les deux disciplines relèvent de deux tâches bien définies et plus complémentaires que comparables.

La recherche d'information répond en termes de documents à une requête donnée. Les procédures qu'elle met en œuvre consistent à identifier, au sein d'un ensemble textuel hétérogène de taille conséquente, les documents pertinents en regard de la thématique que recouvre la demande. De son côté, l'extraction organise l'information interne au document en un certain nombre de champs prédéfinis.

> In this paper we shall use a narrower definition : the identification of instances of a particular class of events or relationships in a natural language text, and the extrac-

> tion of the relevant arguments of the event or relationship. Information extraction therefore involves the creation of a structured representation (such as a data base) of selected information drawn from the text. [Grishman, 1997 : p. 10]

Par conséquent, l'extraction s'inscrit plus précisément dans une démarche de compréhension textuelle. Nous ne pouvons pas organiser l'information à partir des seuls mots que compte un texte, les relations qu'ils entretiennent doivent être prises en compte. Il n'est donc pas étonnant que les premiers moteurs d'extraction se soient inscrits dans une démarche de compréhension globale. Ainsi, des systèmes comme TACITUS (*cf.* [Hobbs *et al.*, 1993]) et GE NLTOOLSET (*cf.* [Jacobs et Rau, 1990]) associent à chaque phrase du texte d'entrée une représentation logique ou conceptuelle. Ces structures *sémantiques* servent ensuite à isoler les informations pertinentes.

Du point de vue de la précision et du rappel mais également de la rapidité d'exécution, les performances d'un système de compréhension globale sont assez médicores. TACITUS, par exemple, atteint seulement 24% de rappel et 40% de précision. La principale raison de ce manque de performance est l'inadéquation des méthodes utilisées au but poursuivi. Quand un système de compréhension doit, pour un texte donné, rendre compte de toute sa complexité (incluant en cela les nuances sémantiques et les objectifs de l'auteur), un moteur d'extraction vise le remplissage d'un formulaire prédéfini et simple. Par conséquent, il n'est pas nécessaire, dans ce dernier cas, d'envisager l'analyse complète du texte, l'information pertinente se limitant la plupart du temps à 10% du texte (*cf.* [Appelt *et al.*, 1993]). L'extraction d'information relève donc plus d'une démarche de *compréhension locale*.

Afin de limiter l'analyse aux seuls éléments pertinents, les extracteurs plus récents ont envisagé différentes solutions qui s'observent au niveau du *pré-traitement* par l'identification de mots-clés déclencheurs (*trigger words*, *cf.* [Appelt *et al.*, 1993; Yangarber et Grishman, 1998]), mais aussi, et surtout, à un niveau plus global, dans le choix de la méthode d'analyse.

2.2.1 Analyse profonde *vs.* analyse de surface

L'analyse syntaxique est indéniablement le composant central d'un système d'extraction. Les éléments à extraire sont, en majorité, des syntagmes nominaux et les relations qui les lient les uns aux autres sont bien souvent formalisables sous la forme d'une structure prédicat – arguments (*cf.* Exemples 31 et 32).

(31) Airborne, le spécialiste du frêt aérien, va **être racheté par** Deutsche Post **pour** 1,05 milliard de dollars, soit 21,25 dollars par action.

(32) Le groupe allemand d'énergie et de services aux collectivités E.ON prévoit de **vendre** Viterra Energy Services **à** CVC Capital Partners **pour** un montant d'envi-

ron 1 milliard d'euros, selon des agences de presse.

C'est cette importance de la syntaxe qui a poussé, dans un premier temps (MUC-1 et MUC-2), nombre de systèmes à privilégier une analyse syntaxique profonde afin d'assurer une reconnaissance maximale (*e.g.* TACITUS, [Hobbs *et al.*, 1993] ; GE NLTOOLSET, [Jacobs et Rau, 1990] ; PROTEUS, [Yangarber et Grishman, 1998]). Cependant, nous l'avons vu, cette procédure n'est pas optimale. Les phrases complexes induisent une combinatoire lexico-syntaxique qu'il n'est pas toujours nécessaire de gérer, toutes ces phrases ne relevant pas forcément du domaine d'extraction. De plus, l'expérience montre que les arguments pertinents en regard de l'information à extraire s'organisent fréquemment en de simples fragments qu'une analyse de surface suffit à extraire. Par conséquent, dès la troisième édition de la *Message Understanding Conference*, ces réflexions ont conduit les chercheurs à préférer une analyse syntaxique de surface (*cf.* FASTUS, [Appelt *et al.*, 1993] ; NYU, [Grishman, 1995]). Dans ce cas, les fragments à extraire sont formalisés sous la formes d'automates à états finis. Il ne s'agit plus alors d'analyser la phrase complète mais d'isoler, au sein de celle-ci, les éléments pertinents pour l'extraction (*pattern matching*).

Pratiquement, une analyse de surface se limite à l'identification de structures syntaxiques nominales et verbales non récursives. L'attachement des groupes prépositionnels ou adjectivaux et des relatives est considéré au niveau de la reconnaissance des patrons d'extraction (et donc limité à ces derniers) par le biais de règles génériques comme :

– N_0 (<PREP> N)* V
– N_0 <PRO+rel> N* V (N + <PREP> N)* V
– N_0 V (N + <PREP> N)* <CONJC> V
– *etc.*

Ainsi, pour l'exemple 33, une analyse de surface aboutira à la structure 33. L'application de patrons d'extraction tels que `killing of <victime>`, `accused <suspect> of the crime` identifiera ensuite les séquences *Attorney General Roberto Garcia Alvarado* comme la victime d'un incident de type *killing* dont le suspect présumé est *Farabundo Marti National Liberation Front*.

(33) Salvadoran President-elect Alfredo Cristani condemned the terrorist killing of Attorney General Roberto Garcia Alvarado and accused the Farabundo Marti National Liberation Front (FMLN) of the crime.

(34) [$_N$ Salvadoran President-elect Alfredo Cristani] [$_V$ condemned] [$_N$ the terrorist killing] of [$_N$ Attorney General Roberto Garcia Alvarado] and [$_V$ accused] [$_N$ the Farabundo Marti National Liberation Front (FMLN)] of the crime.

Une telle analyse permet de limiter la complexité syntaxique aux structures relevant du domaine d'extraction, accélérant de cette manière le temps de traitement et limitant l'impact des erreurs potentielles. Il s'agit néanmoins d'un compromis avec la part d'imperfection qu'une telle notion implique.

> Traditional syntactic analysis has the role not only of identifying syntactic structure but also of regularizing syntactic structure. In particular, different clausal forms, such as active and passive forms, relative clauses, reduced relatives, etc. are mapped into essentially the same structure. This regularization simplifies the scenario pattern matching - it needs deal with fewer forms of each scenario pattern. [Grishman, 1997 : p. 22]

En conséquence de ces différentes remarques, nous constatons que l'analyse syntaxique quelle qu'elle soit pose de nombreux problèmes. Bien que les systèmes actuels privilégient une analyse de surface, plus rapide, les résultats peuvent être, dans le même temps, plus fragmentaires, et doivent être rassemblés ultérieurement (les liens syntaxiques n'étant pas tous exprimés). Il n'y a donc pas une réponse idéale mais un faisceau de compromis plus ou moins pertinents.

2.3 Une nécessaire adaptabilité

Au-delà de l'objectif poursuivi (*i.e.* le remplissage d'un formulaire) l'extraction d'information, telle qu'elle a été définie par les conférences MUC et ACE, suppose une certaine adaptabilité des ressources. Le temps imparti aux concurents pour adapter un extracteur à un scénario est très court (de un à six mois). Cette situation a conduit beaucoup de chercheurs à privilégier une solution *statistique* reposant sur l'apprentissage automatique des ressources, par opposition aux solutions *linguistiques* (ou *à base de connaissances*), considérées plus coûteuses.

2.3.1 Approche statistique et apprentissage

L'approche statistique s'inscrit dans une démarche automatique d'apprentissage[5] des ressources. Cet apprentissage nécessite l'utilisation d'un corpus. Les méthodes se distinguent selon que le corpus utilisé est annoté ou non. L'annotation dont il est question concerne les éléments pertinents pour l'extraction, c'est-à-dire les séquences propres à un scénario donné (les différents acteurs, l'objet et le montant d'une transaction, par exemple).

[5]Nous considérons dans cette section l'apprentissage en contexte libre (*e.g.* non structuré). Il existe également des algorithmes s'appuyant sur la structure de certains documents comme les horaires, les annonces immobilières, les offres d'emploi, *etc.* Ce type de documents définit un cadre d'extraction particulier et plus immédiat que nous n'abordons pas dans ce travail.

Corpus annotés

[Riloff, 1993] part d'un ensemble de séquences, résultats d'une extraction antérieure. Elle isole, ensuite, les phrases qui les contiennent et, au moyen de l'analyseur CIRCUS (*cf.* [Lenhert, 1991]), génère pour chacune d'entre-elles une structure conceptuelle. L'articulation fondamentale de cette procédure d'apprentissage consiste en l'identification, au sein de ces structures, d'un *point d'ancrage conceptuel*, c'est-à-dire un mot qui peut suffire à activer le concept. Considérons, par exemple, la phrase *the diplomat was kidnapped* et la séquence de déclenchement *the diplomat*. L'analyseur CIRCUS attribuera à la phrase la représentation `<subject> passive-verb` et *AutoSlog* retournera le terme *kidnapped* comme point d'ancrage. Cette méthode permet d'acquérir très rapidement une masse importante de motifs d'extraction. Cependant, ces motifs délimitent une seule information et ne permettent pas d'extraire les relations qui lient cette information aux autres (*cf.* Tableau 2.7). L'apprentissage substitue, dans une certaine mesure, la quantité à la qualité.

Patron linguistique	Patron d'extraction
<subject> passive-verb	<victim> was murdered
<subject> active-verb	<perpetrator> bombed
<subject> verb infinitive	<perpetrator> attempted to kill
<subject> auxiliary noun	<victim> was victim
passive-verb <dobj>	killed <victim>
active-verb <dobj>	bombed <target>
infinitive <dobj>	to kill <victim>
verb infinitive <dobj>	threatened to attack <target>
gerund <dobj>	killing <victim>
noun auxiliary <dobj>	fatality was <victim>
noun prep <np>	bomb against <target>
active-verb prep <np>	killed with <instrument>
passive-verb prep <np>	was aimed at <target>

Tableau 2.7 : Exemple de motifs extraits par AutoSlog

[Soderland *et al.*, 1995] propose une solution d'apprentissage différente. Il part d'un ensemble de *nœuds conceptuels*. Chaque nœud représente l'information pertinente d'une phrase du corpus d'entraînement (*e.g.* les séquences extraites) de même que les contraintes lexicales et syntaxiques qui ont conduit à leur extraction. Au départ de ces nœuds conceptuels, le travail principal de CRYSTAL est de réduire progressivement les jeux de contraintes afin d'augmenter leur couverture. Pour ce faire, la procédure sélectionne un premier nœud et isole ensuite les nœuds similaires (selon une mesure de similarité non précisée). Une nouvelle définition conceptuelle est créée en relâchant les contraintes des deux nœuds de manière à les unifier. Le nouveau concept est ensuite proposé au corpus d'entraînement pour validation. Si l'extraction qui en résulte est pertinente, le candidat est retenu et les deux nœuds d'origine sont supprimés de l'ensemble de départ.

Corpus bruts

Bien qu'efficaces, les méthodes que nous venons d'exposer reposent, nous l'avons dit, sur l'existence d'un corpus annoté. Il est évident que si ce corpus n'existe pas, la procédure d'annotation (nécessairement manuelle) alourdit considérablement le processus global qui, dès lors, perd de son intérêt. Par conséquent, diverses expériences ont été menées pour évaluer la faisabilité d'un apprentissage sur corpus brut, c'est-à-dire non annoté.

[Riloff, 1996] est une réimplémentation d'*AutoSlog* (*cf.* [Riloff, 1993]). Ce nouvel algorithme ne requiert plus l'annotation interne du corpus et se base uniquement sur un ensemble préclassé de documents pertinents et non pertinents (*e.g.* relevant ou non du domaine d'extraction). Le processus opère en deux temps. Dans un premier temps, les phrases du corpus sont analysées par CIRCUS et une structure conceptuelle est générée pour chaque syntagme nominal. Ce processus est semblable à celui d'*AutoSlog*. Dans un second temps, les nouvelles structures sont réappliquées sur le corpus et leur pertinence est estimée en mesurant leur fréquence d'apparition au sein des textes du domaine par rapport à leur fréquence d'apparition totale. Cette seconde passe permet de classer les candidats motifs en fonction de leur degré d'appartenance au domaine.

[Riloff et Jones, 1999] apportera une troisième et dernière amélioration à *AutoSlog* en introduisant l'idée d'un *déclenchement mutuel*. L'algorithme commence par appliquer une série de motifs sur un corpus. Il ressort de cette analyse l'extraction d'un lexique d'arguments. La soumission des arguments au corpus permet ensuite d'obtenir de nouveaux candidats-motifs qui viendront compléter la bibliothèque de patrons d'extraction. La procédure est appliquée plusieurs fois jusqu'à convergence.

La solution proposée par [Yangarber *et al.*, 2000] rejoint cette idée d'un déclenchement mutuel. Au départ d'une base de patrons d'extraction existants, le corpus d'entraînement est divisé en deux sous-corpus. Le premier rassemble les documents contenant au moins une occurrence d'un patron de la base et le second, tous les autres. Pour chaque document du premier corpus, la procédure infère de nouveaux patrons qui sont classés selon une mesure comparable à celle utilisée dans [Riloff, 1996] (*e.g.* par comparaison de leur vitalité au sein des deux sous-corpus). Les patrons les plus pertinents sont ajoutés à la base initiale et la procédure recommence au début. Notons, toutefois, qu'au contraire de [Riloff et Jones, 1999], [Yangarber *et al.*, 2000] apprend des patrons qui ne se limitent pas toujours à une information (*i.e.* l'objet ou le sujet d'un prédicat verbal ou le complément prépositionnel d'un prédicat nominal) et peuvent exprimer une relation (*cf.* Tableau 2.8[6]).

[6] *C-Company* et *C-Person* représentent respectivement les entités *compagnie* et *personne*. *C-Appoint*, *C-Resign* et *C-Buy* désignent des classes de verbes exprimant la relation qui lie la ou les entités au prédicat. Ces patrons montrent une structure qui s'apparente à une phrase élémentaire. Nous rapprochons donc cette solution de celle

Subject	Verb	Direct Object
C-Company	C-Appoint	C-Person
C-Person	C-Resign	–
	C-Buy	C-Company
C-Company	merge	*

Tableau 2.8 : Exemple de motifs extraits par Ex-Disco

Nous n'avons pas dressé une liste exhaustive des solutions d'apprentissage. Nous pourrions encore mentionner des systèmes tels que [Huffman, 1996], [Soderland, 1999], *etc.* Toutefois, les procédures restent globalement les mêmes. L'apprentissage opère soit au départ d'un ensemble de motifs pré-existants, soit au départ de syntagmes nominaux, résultats d'une extraction antérieure. Par ailleurs, nous avons orienté notre description des méthodes statistiques autour de l'acquisition de patrons d'extraction. Or, la spécialisation à un domaine peut également nécessiter un apprentissage lexical afin de couvrir au mieux les termes de ce domaine. Nous reviendrons plus longuement sur cette notion au chapitre 5.

2.3.2 Approche linguistique et modularité

L'approche linguistique, appelée également extraction à base de connaissances, repose sur le développement de grammaires et suppose la combinaison d'une démarche introspective et d'une méthodologie basée sur corpus, plus précisément un corpus de spécialité. Comme nous l'exposions dans le premier chapitre, une telle réunion permet d'une part d'anticiper les situations non observées au sein du corpus et, d'autre part, de compléter certaines omissions du linguiste. Toutefois, un travail de ce type demande beaucoup de temps et d'efforts, la procédure étant itérative. Chaque règle doit, en effet, être testée et déboguée de nombreuses fois. De plus, la performance du système est étroitement liée à l'attention du linguiste responsable de son développement.

S'l est évident qu'aujourd'hui, nous ne pouvons pas substituer la machine au linguiste sans compromettre la qualité de la description linguistique, la machine peut toutefois intervenir à différents niveaux du travail de formalisation pour automatiser ou simplement faciliter la procédure. Ainsi, [Yangarber et Grishman, 1998] décrit le fonctionnement d'une interface graphique, PET, permettant l'acquisition semi-automatique de connaissances. Sur base d'exemples d'évènements et d'une description des structures informationnelles attendues, PET crée le patron approprié pour réaliser l'extraction. Il suggère également certaines généralisations afin d'en augmenter la couverture. Ce principe d'acquisition semi-automatique a été envisagé de manière comparable dans [Cunningham *et al.*, 2002] et [Poibeau, 2003].

Dans le cadre d'une analyse monolithique, cette procédure ne pourrait être envisagée. Les connaissances linguistiques à mettre en œuvre au sein des patrons d'extraction seraient trop nombreuses et le travail d'inférence et de généralisation résultant trop complexe pour être pré-

que nous présentons en fin de ce chapitre (*cf.* Section 2.4).

cis. PROTEUS, l'extracteur pour lequel a été développé PET, montre une architecture modulaire, séparant connaissances générales (immuables) et spécialisées (dépendantes du domaine d'extraction). De cette manière, la formalisation des patrons d'extraction (relevant des connaissances spécialisées) est facilitée puisque débarrassée des considérations lexicales et syntaxiques générales. De cette manière, PET travaille au départ d'un nombre très réduit d'éléments qui sont par conséquent plus facilement manipulables.

Si nous n'observons pas d'autres approches visant l'acquisition semi-automatique de patrons d'extraction, la modularité est au centre de la majorité des solutions dites linguistiques (*cf.* [Appelt *et al.*, 1993], [Jacobs *et al.*, 2001] et [Jacobs et Rau, 1990], par exemple).

Pendant longtemps, nous avons dû constater qu'une nette priorité a été accordée aux méthodes statistiques, priorité certainement imputable à leur apparent automatisme. Selon la situation d'apprentissage, cet automatisme peut, néanmoins, s'avérer des plus fastidieux. [Appelt et Israel, 1999] expose à ce propos deux situations qui peuvent se présenter lors de la mise au point d'un extracteur d'entités nommées. Si ce système voit sa spécification première, la sensibilité à la casse, s'inverser, une méthode statistique n'aura qu'à réappliquer son algorithme d'entraînement à un corpus préalablement débarrassé de ses majuscules. Par contre, l'approche à base de connaissances se verra contrainte de réécrire toutes ses règles. Cette première situation favorise donc une approche statistique mais, s'il s'agit plutôt d'étendre le champ d'action du système, une approche linguistique n'aura qu'à étendre la couverture de sa grammaire contrairement aux méthodes statistiques qui auront à établir un nouveau corpus d'entraînement.

Ces dernières années, la linguistique computationnelle a commencé à considérer une potentielle intégration des solutions statistiques et linguistiques afin de profiter des avantages des deux. En extraction d'information, cette idée se traduit par l'association de la puissance descriptive des solutions linguistiques et des facilités statistiques d'apprentissage[7]. C'est un point que nous croyons essentiel et que nous envisagerons au cours de ce travail. Notons que certains développements récents (*cf.* [Feldman *et al.*, 2006]) intègrent déjà cette notion.

2.4 La phrase élémentaire comme patron informationnel

Dans ce travail, nous nous concentrerons plus particulièrement sur le domaine financier. Ce domaine a été introduit lors de la cinquième édition de *la Message Understanding Conference* (*cf.* [Onyshkevych *et al.*, 1993]). Il s'agissait alors d'identifier les informations relatives aux

[7]Le recours aux statistiques ne se limite pas à l'apprentissage mais intervient également dans la procédure d'extraction proprement dite par l'utilisation de formalismes stochastiques à différents niveaux de l'analyse (étiquetage, analyse syntaxique, *etc.*). Nous ne considérons pas ces solutions dans ce chapitre. Elle ne sont pas propres à l'extraction mais au traitement automatique de la langue dans son ensemble. Nous les introduirons au cas par cas si nous envisageons leur utilisation.

créations de co-entreprises (*joint-ventures*). De notre côté, nous avons privilégié les informations relatives aux transactions financières et plus particulièrement celles relatives à l'achat ou la vente d'entreprises. Dans ce cas précis, les informations pertinentes concernent les entités qui entrent en jeu dans la transaction de même que la nature de l'échange. Certaines informations, comme le montant de la transaction ou la description des entités, peuvent également venir compléter le patron d'extraction. Schématiquement, le concept d'*échange* peut être représenté dans un réseau conceptuel organisant ces informations (*cf.* Figure 2.6).

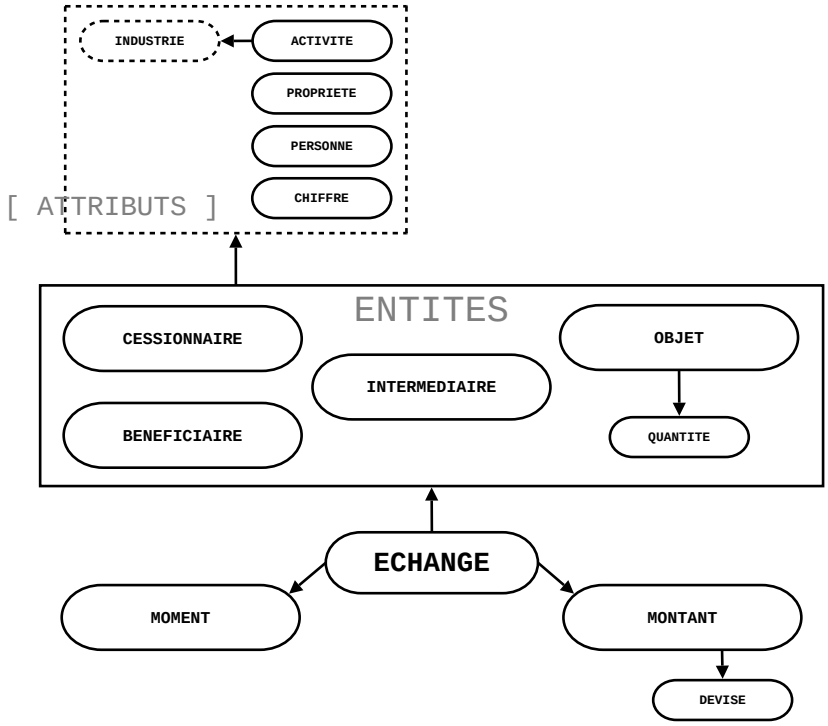


Figure 2.6 : Réseau conceptuel des transactions

Nous illustrons, dans le tableau 2.10, différents contextes énonciatifs relatifs au domaine des transactions financières. Les séquences soulignées représentent les informations pertinentes au regard des champs que nous avons définis ci-dessus.

D'un point de vue linguistique, une transaction suggère un prédicat datif (bénéfactif ou privatif selon la direction de l'échange). Le lexique-grammaire (plus particulièrement [Leclère, 1978] et [Guillet et Leclère, 1992]) définit une classe, la table 36DT, qui rassemble ces verbes autour de la structure N_0 *V* N_1 *à* N_2 (*Advp*)[8]. Cette classe nous permet d'isoler les séquences suivantes :

(35) La CPR *cède* son pôle obligataire

(36) la *cession* de son pôle obligataire à Natexis BP

[8]Ce dernier complément n'est pas essentiel mais caractéristique des prédicats financiers. Il désigne un montant éventuel noté N_3.

(37) Dexia *a acquis* 329 621 actions Dexia France au prix unitaire de 150 euros

(38) Générale Location *a cédé* à Unibail 389 620 actions Paris

(39) *Cette cession* concerne seulement une vingtaine d'opérateurs de marché

(40) *l'offre* lancée par son concurrent Unibail

La CPR cède son pôle obligataire Comme nous l'annoncions dans nos colonnes le 26 octobre dernier, la CPR a déclaré hier soir avoir signé un accord avec Natexis Banques Populaires (issu de la fusion de la Caisse Centrale des Banques Populaires et de Natexis) portant sur la cession de son pôle obligataire à Natexis BP. Cette cession concerne seulement une vingtaine d'opérateurs de marché dont le transfert effectif interviendra dans les prochaines semaines. Pour la CPR, l'opération permet de se recentrer sur ses activités de base, les métiers actions.
Dexia détient 99,27% du capital de Dexia France Le Conseil des Marchés Financiers annonce ce soir que Dexia a acquis 329 621 actions Dexia France au prix unitaire de 150 euros pendant la durée de l'offre publique de retrait suivie d'un retrait obligatoire, soit du 23 février au 21 mars 2000 inclus. Dexia détient désormais 99,27% du capital et des droits de vote de la société. Le communiqué du CMF indique que les actions de Dexia France ont été radiées aujourd'hui du premier marché au comptant, et spécifie que les actions non présentées à l'offre publique de retrait par les actionnaires minoritaires sont transférées à l'initiateur de l'offre.
Paris Expo : fin du suspense Le Conseil des Marchés Financiers, a annoncé hier que Générale Location avait apporté ses titres à l'offre lancée par son concurrent Unibail. Générale Location a cédé à Unibail 389 620 actions Paris Expo et ne conserve plus que 480 actions de cette société, le prix de transaction étant de 117 euros par action. Ainsi, la bataille que s'étaient livrée Générale Location et Unibail est finie. Le titre Unibail, à 15h30, s'apprécie de 3,5%, à 139 euros.

Tableau 2.10 : Exemples de dépêches financières

Dans un contexte générique, nous ne pouvons pas caractériser finement la sémantique des constituants. Ainsi, les traits associés à la table se limitent à :

– N_0 = : N_{hum}
– N_1 = : $N_{concret}$
– N_2 bénéficiaire
– N_2 = : N_{hum}

Toutefois, la restriction du focus sémantique au domaine des transactions financières nous permet de préciser davantage ces traits et, selon que le datif est privatif ou non, d'imaginer les deux représentations suivantes :

	BÉNÉFACTIF	PRIVATIF
N_0	ACQUÉREUR	VENDEUR
N_1	OBJET	
N_2	VENDEUR	ACQUÉREUR
N_3	MONTANT	

La phrase élémentaire forme donc une unité informationnelle qui, considérée dans un contexte sémantique bien délimité, offre trois avantages pertinents pour l'extraction d'information :

– **sémantique :** il nous est possible de caractériser finement la sémantique des classes d'équivalence que forment les constituants essentiels de la phrase élémentaire ;
– **syntaxique :** les transformations associées à chaque table nous permettent de décliner la forme élémentaire canonique en un ensemble de patrons d'extraction ;
– **informationnel :** en plus de délimiter les informations à extraire, la phrase élémentaire nous permet de conserver les relations que ces informations entretiennent. Les liens syntaxiques qui réunissent *acheteur*, *vendeur*, *objet* et *montant* correspondent, au niveau informationnel, aux différents liens de notre réseau conceptuel.

Bien entendu, nous ne pouvons pas nous limiter à une unique structure élémentaire. Certains constituants essentiels ne sont pas toujours actualisés (*cf.* Exemples 35 à 40) et peuvent s'inscrire dans une autre relation *prédicat – arguments* :

(41) la CPR *a déclaré hier soir avoir signé un accord* avec Natexis Banques Populaires

(42) Dexia *détient* 99,27% du capital de Dexia France

(43) Dexia *détient* désormais 99,27% du capital et des droits de vote de la société

(44) Générale Location *avait apporté* ses titres à l'offre lancée par son concurrent Unibail

Ces prédicats ne définissent pas à proprement parler une transaction. Mais, parce que leurs constituants expriment des faits nécessaires à l'extraction, ils entrent, avec la structure dative, dans la construction du *sous-domaine* financier des transactions. Cette notion de *(sous-)domaine* (*cf.* [Guenthner, 1998]) introduit deux éléments particulièrement intéressants : la spécialisation lexico-syntaxique et les classes sémantiques. Ces deux éléments trouvent un cadre théorique commun au travers de la notion de *classes d'objets* (*cf.* [G. Gross, 1994; Pesant et Mathieu-Colas, 1998; G. Gross et Geunthner, 2002], par exemple)[9].

> On est alors mené à postuler, pour rendre compte de l'emploi précis d'un opérateur donné, des informations supplémentaires concernant les arguments. Il s'agit de

[9]Il ne s'agit certainement pas du seul cadre théorique réunissant les deux notions. Nous pourrions entre autres évoquer les travaux de Mel'cuk (*cf.* [Mel'cuk *et al.*, 1984]). Toutefois, les liens évidents qui associent *classes d'objets* et *lexique-grammaire*, nous ont amené à privilégier l'approche de [G. Gross, 1994].

> sous-catégoriser les traits en sous-classes sémantiques, que nous appelons **classes d'objets**, et qui seules permettent de discriminer les sens de l'opérateur avec la précision nécessaire à la reconnaissance ou à la génération de phrases correctes. Que les propriétés exclusivement syntaxiques soient insuffisantes pour le traitement automatique est un fait reconnu par tous ceux qui ont travaillé dans le domaine. [G. Gross, 1994 : p. 18]

Les classes d'objets sont des classes d'arguments (*e.g.* des classes d'équivalence qui caractérisent les constituants essentiels d'une phrase élémentaire). Leur objectif est de préciser davantage la nature sémantique des arguments d'un prédicat. C'est donc par relation à ce prédicat qu'elles sont définies. Toutefois, [Pesant et Mathieu-Colas, 1998] souligne qu'un prédicat est bien souvent insuffisant pour circonscrire une classe complète, le recours à une conjonction de prédicats (ou de classes de prédicats) est nécessaire. Cette observation rejoint celle que nous avons faite quant à l'extraction des informations relatives aux transactions.

Du point de vue de l'extraction, nous pouvons considérer chaque information comme formant une classe d'objets. Les prédicats qui actualisent ces classes définissent alors les patrons d'extraction propres au domaine. Ce formalisme offre l'avantage de mettre les informations (*e.g.* les constituants essentiels) en relation d'un prédicat à l'autre et d'ainsi faciliter leur rassemblement au sein du formulaire.

Bien que la procédure suggérée ici offre des perspectives très intéressantes, elle ne suffit pas d'emblée à résoudre tous les problèmes posés par l'extraction. Intuitivement, nous voyons deux éléments supplémentaires à prendre en compte.

1. Certaines informations, parce que circonstancielles, peuvent ne pas s'inscrire dans un schéma de phrase élémentaire (les adverbes de temps tels que *hier* ou *du 23 février au 21 mars 2000 inclus*, par exemple). Nous pouvons toutefois traiter ces structures adverbiales assez aisément au moyen d'automates spécifiques (*cf.* [Senellart, 1999]).
2. La phrase élémentaire ne suffit pas à séparer informations *effectives* et *potentielles*. Le monde de la finance est, en effet, un monde de spéculations qui projette à l'avance certaines transactions non effectives. Heureusement, les verbes et les structures de ce type sont limités (au sein du domaine financier) et peuvent être recensés exhaustivement sans difficulté (*cf.* Exemples 45 et 46).

(45) **Selon** la presse allemande (le quotidien économique Handelsblatt), Deutsche Telekom **aurait** proposé à Lagardère de lui racheter 50% de Club-Internet pour 450 millions d'euros.

(46) Rappelons en effet que le nom de France Telecom a souvent été évoqué au mo-

ment de la cession d'Orange (voire de One2One) et les observateurs **pensent** que **si** Vodafone réussit son offre hostile sur Mannesmann, il revendra l'opérateur britannique au groupe dirigé par Michel Bon.

Au terme de cette première réflexion, nous constatons que le lexique-grammaire est *a priori* un formalisme pertinent pour l'extraction. Associé à la notion de classe d'objets, il constitue une base de connaissances couvrant la majorité des besoins de la discipline.

2.5 Conclusion

Nous avons défini ici les principaux enjeux de l'extraction d'information tant du point de vue des résultats (précision et rappel) que de la conception d'une application (adaptabilité). Nous avons également défini notre hypothèse de travail qui s'inscrit dans le formalisme du lexique-grammaire et vise à exploiter la phrase élémentaire comme patron informationnel.

Nous ne nous étendrons pas sur la création des classes sémantiques qui, pour le domaine des transactions financières, ont déjà été envisagées par [Poibeau, 2003]. Il définit un ensemble de prédicats formant une classe sémantique `Opération`. Ces prédicats peuvent servir de base à une solution de type lexique-grammaire. Sa démarche d'analyse, à base d'automates, rappelle d'ailleurs ce formalisme. Il s'est toutefois centré sur l'aspect sémantique de l'extraction et a proposé une solution d'apprentissage de classes que nous pouvons considérer comme formant un ensemble de tables de lexique-grammaire (bien qu'il ne leur associe pas de propriétés distributionnelles). De notre côté, nous nous sommes plus particulièrement intéressé à l'analyse syntaxique, c'est-à-dire à l'identification des informations plus qu'à leur conceptualisation.

Pratiquement, nous poursuivons deux buts. Nous cherchons, d'une part, à étudier l'utilisation d'un lexique-grammaire en extraction d'information (du point de vue de la phrase élémentaire mais également du formalisme lui-même) et, d'autre part, à identifier une solution d'analyse rencontrant les enjeux de l'extraction.

D'un point de vue plus théorique, nous tenterons de répondre aux deux questions qui sous-tendent la création d'un moteur d'extraction :

– *Solution linguistique ou statistique ? A priori*, nous nous dirigeons vers une solution linguistique exclusivement. Toutefois, nous envisagerons certaines alternatives statistiques, soit pour estimer leur intérêt, soit pour répondre à certains impératifs de la discipline.
– *Analyse profonde ou partielle ?* L'objectif du lexique-grammaire est indéniablement l'analyse syntaxique en profondeur. Ce sera donc naturellement notre point d'entrée. Selon les difficultés que nous rencontrerons, nous envisagerons néanmoins différents compromis à l'exactitude syntaxique.

Le prochain chapitre proposera une solution à l'extraction qui repose sur la phrase élémentaire du lexique-grammaire. Cette solution doit être considérée comme une première approche dont la naïveté nous permettra de mettre en évidence les problèmes induits par le formalisme utilisé. Ces problèmes seront ensuite étudiés afin d'imaginer une nouvelle solution plus efficace, compte tenu des impératifs de la discipline.

CHAPITRE 3

DE LA TABLE À L'AUTOMATE D'EXTRACTION

Nous envisageons, dans ce chapitre, la mise au point et l'évaluation d'une méthode d'extraction d'information tirant parti des ressources linguistiques du lexique-grammaire. Nous voulons, de cette manière, estimer la pertinence de la *phrase élémentaire* en tant que *patron informationnel*. Nous devons donc nécessairement nous positionner dans un contexte applicatif réel. Un tel objectif demande la prise en compte d'un grand nombre de phénomènes : l'ambiguïté lexicale, la complexité syntaxique, l'anaphore, *etc.* Dans la mesure où ces phénomènes peuvent poser problème, nous envisagerons des solutions afin, au final, d'obtenir un extracteur fonctionnel. Toutefois, parce que certains problèmes sortent du cadre précis de cette étude, les solutions que nous apporterons peuvent être partielles et doivent être considérées comme les prémices d'une réflexion plus importante.

Pratiquement, nous procèderons en trois temps. Nous introduirons tout d'abord l'analyse syntaxique propre au lexique-grammaire en regard d'un problème fondamental pour l'extraction d'information : l'ambiguïté lexicale des prédicats. Nous envisagerons ensuite les deux articulations principales d'un analyseur : l'étiquetage morpho-syntaxique et la reconnaissance des structures de phrase. Finalement, nous proposerons une évaluation chiffrée qui nous amènera à discuter les limites et les perspectives d'une extraction à base de lexique-grammaire. Ce dernier point introduira les développements ultérieurs de ce travail.

3.1 Introduction

Rappelons que l'extraction d'information consiste, au sein d'un texte donné, à isoler les différents segments phrastiques pertinents au regard d'un patron informationnel ainsi que leurs relations de dépendance (*cf.* Chapitre 2).

Dans la perspective d'un lexique-grammaire, et plus généralement d'un système à base de connaissances, un moteur d'extraction vise l'identification des syntagmes nominaux au départ d'une base prédicative propre à un domaine particulier.

Considérons brièvement les quelques exemples suivants :

(47) Lafarge ***avait cédé*** en janvier 2001 60% de Materis à trois fonds d'investissements (Carlyle, Advent International et CVC Capital Partners).

(48) En 2003, Scottish & Newcastle ***avait*** également ***cédé*** ses pubs pour un montant de 2,51 milliards de livres, ce qui va lui permettre de mener à bien une nouvelle vague d'acquisitions.

(49) En juillet dernier, HVB ***avait cédé*** sa filiale de crédit à la consommation Norisbank à la DZ Bank pour un montant de 447 millions d'euros.

Si nous nous plaçons dans le contexte des transactions financières que nous privilégions dans ce travail, l'identification du prédicat *céder* nous permet d'extraire un ensemble de groupes nominaux ciblant des informations telles que l'*acheteur*, le *vendeur*, l'*objet de la vente* ou encore le *montant de la transaction*. De cette manière, nous sommes à même de proposer, en sortie du processus d'extraction, différents patrons ou formulaires plus ou moins complets selon le cas (*cf.* Tableau 3.1).

ACHETEUR	trois fonds d'investissements (Carlyle, Advent International et CVC Capital Partners)		la DZ Bank
VENDEUR	Lafarge	Scottish & Newcastle	HVB
OBJET	60% de Materis	ses pubs	sa filiale de crédit à la consommation Norisbank
MONTANT		un montant de 2,51 milliards de livres	un montant de 447 millions d'euros

Tableau 3.1 : Patrons d'extraction

Présentée de cette manière, l'extraction d'information repose sur le repérage, au sein d'un corpus de textes, d'un ensemble de prédicats préalablement identifiés comme relevant du domaine étudié. En effet, la spécialisation sémantique de ces prédicats d'une part, et leur position directrice au sein de la phrase d'autre part, nous permettent non seulement de conférer une valeur informationnelle aux différents constituants, mais également de préserver les relations de dépendance nécessaires à l'exploitation des résultats (*cf.* Chapitre 2).

Toutefois, la simple identification de la forme verbale *céder* ne peut suffire comme le suggèrent les exemples 50 à 54.

(50) Après avoir initialement jugé cette offre insuffisante, **Henkel *a* finalement *cédé* aux sirènes de l'américain**, ce dernier ayant désormais la voie libre pour détenir plus de 95% du capital du groupe, ce qui lui permettra de lancer une procédure de

cession forcée auprès des actionnaires minoritaires, en vue de récupérer 100% du capital de Wella.

(51) En l'absence de statistiques économiques, les acheteurs du marché sont restés en retrait, ***cédant* le devant de la scène aux investisseurs qui pariaient sur la baisse du marché**.

(52) Acier : **George W. Bush *cède***.

(53) Après une ouverture à la hausse, **le CAC 40 *a* progressivement *cédé* du terrain** et perd à la mi-séance 0,93% à 2876 points.

(54) Pour rappel, l'action Legris a clôturé hier à 18,86 euros et il semble bien que **le management *ait cédé* aux pressions de Guy Wiser-Pratte (dont le fonds d'investissement Wiser-Pratte & Co détient 20,5% du capital)**.

Dans ces quelques exemples, il n'y a pas à proprement parler un unique prédicat *céder*. En nous référant à [Gross, 1975], [Guillet et Leclère, 1992] et [Boons *et al.*, 1976], nous dénombrons trois structures différentes pour cette seule forme[1] :

– N_0 *V à ce Qu P* : Max cède à ce qu'Ida parte[2]
– N_0 *V* N_1 *à* N_2 : Max cède son lit à Luc
– N_0 *V à* N_1 : Max cède au désespoir

Ces trois constructions – respectivement complétive (table 7), transitive (table 36DT) et intransitive (table 33) – désignent trois emplois sémantiques différents de la forme verbale, le second étant celui que nous exploitons pour le sous-langage de la finance. Ils ne suffisent pas cependant à rendre compte de l'ensemble des exemples. Il nous faut encore envisager les figements *céder le devant de la scène* et *céder du terrain* pour expliquer toutes les occurrences.

Cette situation est révélatrice d'un phénomène linguistique particulièrement étudié en traitement automatique des langues, l'*ambiguïté lexicale*. Que nous envisagions l'extraction d'un point de vue linguistique ou statistique, nous devons nécessairement la considérer afin d'éviter la multiplication des erreurs de précision.

Beaucoup de traitements, sinon tous, reposent sur une désambiguïsation préalable du sens des mots. C'est effectivement à ce prix que nous pourrons, par exemple, traduire automatiquement un mot d'une langue à l'autre ou encore indexer correctement un document dans un système de recherche d'information. La résolution de l'ambiguïté est dite *IA-complète* (*cf.* [Appelt *et al.*, 1993]), c'est-à-dire qu'elle nécessite la résolution de l'ensemble des problèmes posés par l'intelligence artificielle (problèmes qui reposent eux-mêmes sur une désambiguïsation en amont).

[1] Les exemples sont empruntés aux auteurs des ouvrages précités.
[2] Notons que cette phrase, peu commune, a une valeur plus théorique qu'usuelle.

Toutefois, depuis le début des recherches en traitement automatique de la langue, et plus particulièrement en traduction automatique, certaines perspectives ont pu être dégagées. Ainsi, on a rapidement constaté l'importance du rôle joué par le contexte dans la désambiguïsation du sens des mots. Un grand nombre de solutions partielles ont émergé appréhendant le contexte d'un point de vue tantôt sémantique, tantôt syntaxique.

La majorité des solutions dites sémantiques reposent sur une ou plusieurs bases de connaissances : thésaurus, réseau sémantique ou dictionnaire (*cf.* [Lesk, 1997; Yarowsky, 1992]). *WordNet* (*cf.* [Miller, 1995]) illustre particulièrement bien ces connaissances parce qu'il combine la plupart de leurs propriétés. D'une part, il associe à chaque terme une ou plusieurs définitions à la manière d'un dictionnaire et permet le calcul du sens d'un mot sur base de ses signatures terminologiques (composées de l'ensemble des mots pleins contenus dans ses définitions). Dès lors, pour un mot ambigu donné, un sens sera retenu par étude de l'intersection de ses signatures terminologiques (*e.g.* ses définitions) et de celles des mots pleins de son contexte. D'autre part, *WordNet* définit, pour chaque entrée, une liste de synonymes qui, à la manière d'un thésaurus, forment une hiérarchie conceptuelle dont nous pouvons inférer des classes de mots servant ensuite de base à la désambiguïsation ; un sens particulier sera attribué au mot selon que son contexte, typiquement une centaine de mots, renvoie majoritairement à telle ou telle classe. Finalement, *WordNet* inclut un large éventail de relations sémantiques (hyponymie, hyperonymie, antonymie, méronymie, *etc.*). Si chaque relation est appréciée en termes de distance, nous pouvons pondérer un mot ambigu en fonction de la *distance* qui le sépare des mots de son contexte. Sera retenu le sens minimisant cette distance.

Le principal problème posé par toute solution à base de connaissances est dû à la complexité des traits sémantiques et des définitions associées aux différentes entrées. De cette finesse naît inévitablement une explosion combinatoire qui conduit à un large faisceau de résultats au sein duquel il n'est parfois pas évident ni même possible d'isoler une solution en particulier. Par ailleurs, la richesse des ressources n'est pas toujours nécessaire. [Fairon et Ho, 2004] développe une procédure de désambiguïsation menée au départ du *Webster 1913*, simple dictionnaire explicatif disponible en ligne (*cf.* également, pour ce type d'approche, [Kozima *et al.*, 1993]). Cette procédure repose sur la notion de *quantité d'information échangée* (QIE) et appréhende la désambiguïsation par l'étude des relations de synonymie qui lient un mot, potentiellement ambigu, aux mots de sa définition. On retient comme synonymes les mots échangeant le plus d'informations, c'est-à-dire les mots les plus proches. L'évaluation de la méthode montre que les résultats obtenus sont équivalents à ceux que nous pourrions obtenir à partir de *WordNet*.

Les solutions syntaxiques s'opposent aux solutions sémantiques dans leur définition du contexte. Un approche sémantique organise le contexte sous la forme d'un *sac de mots*. Il s'apparente alors à une fenêtre de lecture (dans son acception la plus simple), sans considération pour les

relations que les mots entretiennent au sein de cette fenêtre sinon celles que nous pouvons déduire par consultation de la base de connaissances. Une approche syntaxique privilégiera quant à elle un contexte local, défini syntaxiquement, allant d'une suite de mots – la collocation[3] – à la phrase.

> The problem of deciding whether a given phoneme sequence is one word with all its meanings, or two homonymous words, is made easier by the fact that words do not for the most part appear alone. They appear in particular environments – at least when they are at selectional frequency – and may have different meanings under particular operators or over particular arguments. [Harris, 1988 : p. 62]

Dans ce paragraphe, Harris confirme que nous pouvons résoudre l'ambiguïté en prenant en compte le contexte lexico-syntaxique du mot ambigu. C'est précisément cette idée que [Paumier, 2003] explore en proposant, à la suite de [Roche, 1993] et [Senellart, 1999], une solution d'analyse à base de graphes paramétrés (*cf.* Section 3.3), reprise également par [Fairon *et al.*, 2005] (*cf.* Section 3.5.1). Il s'agit de formaliser chaque prédicat au sein d'une structure de phrase et d'ainsi étendre le contexte au maximum. Associée à une description systématique des prédicats de la langue telle qu'elle est suggérée par le lexique-grammaire, cette procédure devrait nous permettre de résoudre la plupart des problèmes liés aux phénomènes d'ambiguïtés (mais naturellement pas tous les cas).

Nous examinerons, dans ce chapitre, les possibilités qu'offre cette dernière solution. Nous proposons une procédure volontairement naïve, à considérer comme une première étape dans notre réflexion. Nous désirons avant tout étudier la pertinence de la phrase élémentaire et de l'outillage transformationnel tels qu'ils sont décrits au sein des tables. Nous isolerons finalement les questions que pose l'utilisation du lexique-grammaire dans un processus de traitement automatique du langage : formalisation du groupe nominal, attachement, complexité syntaxique, *etc.*

Dans le second chapitre, nous avons abordé les différentes articulations d'un système d'extraction : tokenisation, analyse lexicale, *etc.* Bien que toutes puissent, selon la langue étudiée, soulever des questions intéressantes, nous ne prétendons pas traiter ici l'ensemble des spécificités des applications de traitement automatique de la langue. Nous privilégions essentiellement les éléments transversaux à notre travail de même que les points faisant l'originalité de notre méthode au regard des solutions communément acceptées.

Ainsi, nous nous étendrons sur les éléments qui touchent plus précisément la problématique des ressources linguistiques (l'étiquetage morpho-syntaxique et l'analyse syntagmatique) afin de discuter leur pertinence au sein d'un système d'extraction. De la même manière, nous ne

[3] Pour une définition de la collocation, nous renvoyons à l'introduction du chapitre 5.

détaillerons pas les notions communes, propres à toute application linguistique automatique, et auxquelles nous n'apportons rien.

3.2 Étiquetage morpho-syntaxique

L'annotation morpho-syntaxique[4] associe à chaque mot d'un texte ou d'une phrase un jeu d'étiquettes grammaticales, morphologiques, syntaxiques et/ou sémantiques, selon le niveau de granularité souhaité. Ces étiquettes confèrent une information lexicale aux mots, information sans laquelle aucune analyse linguistique automatique ne serait possible. Ce sont également ces étiquettes qui introduisent l'ambiguïté au sein du processus d'analyse.

La principale source d'ambiguïté lexicale que nous pouvons observer en français, et dans la plupart des langues, tient au fait, nous l'avons vu, qu'une forme simple ou complexe peut connaître différentes significations ou, en d'autres mots (plus proches du lexique-grammaire), que différents termes peuvent avoir la même forme graphique. Poursuivons, en guise d'illustration, notre analyse du verbe *céder*.

La figure 3.1 rend compte de la nécessaire contextualisation que nous abordions dans l'introduction de ce chapitre. Comme nous le voyons, afin de distinguer les sens du verbe, le TLF[5] associe, à la forme, une structure phrastique caractérisée aux niveaux lexical, syntaxique et sémantique.

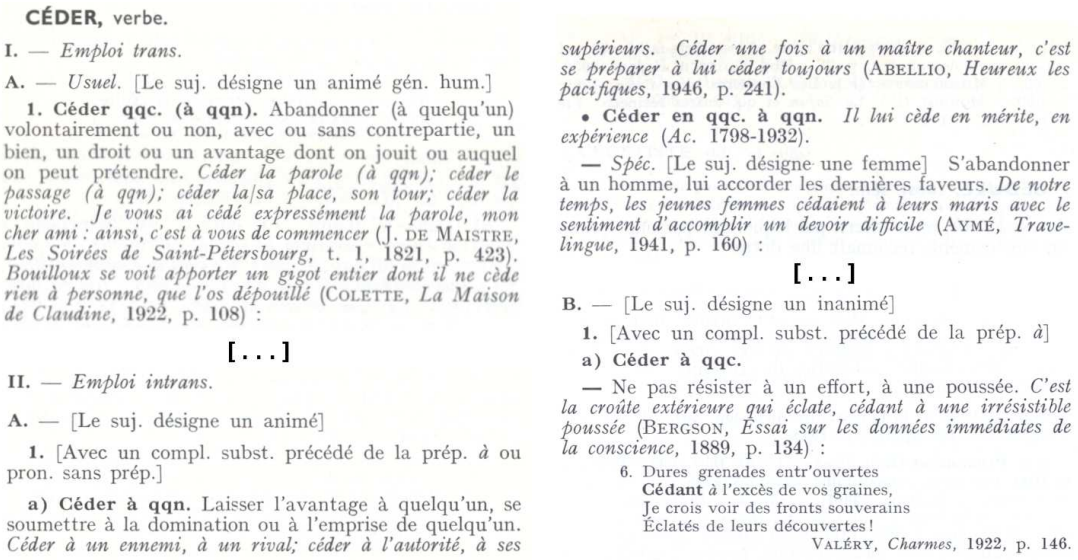
– **Traits lexicaux :** lexicalisation des prépositions (*à qqn/qqc*, *sous qqc*, *etc.*).
– **Traits syntaxiques :** différenciation des emplois transitifs directs ou indirects et intransitifs.
– **Traits sémantiques :** humain *vs* non-humain (*qqn*/*qqc*) ; animé *vs* inanimé.

Cette structure rappelle dans une certaine mesure l'idée de la phrase élémentaire du lexique-grammaire :

> La meilleure façon de décrire les différences de sens d'un mot ou d'une expression consiste à les plonger dans un contexte qui lèvera les ambiguïtés. Mais la cohérence des descriptions impose que l'on normalise les contextes, il ne peut s'agir de contextes inventés au coup par coup pour chaque expression, comme c'est le cas dans les dictionnaires papier. Les contextes naturels sont ceux qui sont constitués de phrases élémentaires. En effet, on peut voir les phrases élémentaires comme des unités minimales de sens. [Gross, 2001 : p. 8]

[4]Par *annotation* ou *étiquetage*, nous désignons uniquement la procédure d'association d'une information morpho-syntaxique à un mot. Nous ne sous-entendons aucunement une potentielle désambiguïsation. S'il y a désambiguïsation, cette dernière doit être considérée comme parallèle à l'annotation.

[5]Le *Trésor de la Langue française* offre un apparat méta-linguistique très riche que tous les dictionnaires ne possèdent pas. Notons toutefois que ces données, contrairement à celles d'un lexique-grammaire, ne sont pas exploitables automatiquement.



CÉDER, verbe.

I. — *Emploi trans.*

A. — *Usuel.* [Le suj. désigne un animé gén. hum.]

1. Céder qqc. (à qqn). Abandonner (à quelqu'un) volontairement ou non, avec ou sans contrepartie, un bien, un droit ou un avantage dont on jouit ou auquel on peut prétendre. *Céder la parole (à qqn); céder le passage (à qqn); céder la/sa place, son tour; céder la victoire. Je vous ai cédé expressément la parole, mon cher ami : ainsi, c'est à vous de commencer* (J. DE MAISTRE, *Les Soirées de Saint-Pétersbourg*, t. 1, 1821, p. 423). *Bouilloux se voit apporter un gigot entier dont il ne cède rien à personne, que l'os dépouillé* (COLETTE, *La Maison de Claudine*, 1922, p. 108) :

[...]

II. — *Emploi intrans.*

A. — [Le suj. désigne un animé]

1. [Avec un compl. subst. précédé de la prép. *à* ou pron. sans prép.]

a) Céder à qqn. Laisser l'avantage à quelqu'un, se soumettre à la domination ou à l'emprise de quelqu'un. *Céder à un ennemi, à un rival; céder à l'autorité, à ses supérieurs. Céder une fois à un maître chanteur, c'est se préparer à lui céder toujours* (ABELLIO, *Heureux les pacifiques*, 1946, p. 241).

• **Céder en qqc. à qqn.** *Il lui cède en mérite, en expérience* (*Ac.* 1798-1932).

— *Spéc.* [Le suj. désigne une femme] S'abandonner à un homme, lui accorder les dernières faveurs. *De notre temps, les jeunes femmes cédaient à leurs maris avec le sentiment d'accomplir un devoir difficile* (AYMÉ, *Travelingue*, 1941, p. 160) :

[...]

B. — [Le suj. désigne un inanimé]

1. [Avec un compl. subst. précédé de la prép. *à*]

a) Céder à qqc.

— Ne pas résister à un effort, à une poussée. *C'est la croûte extérieure qui éclate, cédant à une irrésistible poussée* (BERGSON, *Essai sur les données immédiates de la conscience*, 1889, p. 134) :

6. Dures grenades entr'ouvertes
Cédant à l'excès de vos graines,
Je crois voir des fronts souverains
Éclatés de leurs découvertes!
VALÉRY, *Charmes*, 1922, p. 146.

Figure 3.1 : Article céder (extrait), Trésor de la Langue française

Bien sûr, nous ne pouvons limiter la notion d'ambiguïté à un phénomène essentiellement sémantique. En plus d'une différence de sens entre deux formes identiques, nous pouvons également rencontrer une différence de nature grammaticale[6]. Le terme *cessions* peut, par exemple, tout à la fois être un nom et un verbe :

– **<cessions,cession.N+z1 :fp> :** Air France a confirmé son objectif de résultat d'exploitation hors *cessions* aéronautiques légèrement positif sur l'ensemble de l'exercice.[7]
– **<cessions,cesser.V+t+z1 :I1p :S1p> :** Et si nous *cessions* de nous résigner ? Et si nous *cessions* de raisonner "hexagonalement" ?

Selon que nous privilégions une procédure d'analyse approximative ou exacte[8], ces différentes ambiguïtés seront résolues ou non au niveau morpho-syntaxique, précédant l'analyse syntagmatique (c'est-à-dire le découpage en constituants). À ces deux procédures correspondent deux méthodes d'étiquetage, l'une statistique et l'autre à base de dictionnaires.

[6]Il est entendu que l'ambiguïté ne pourrait se limiter à ce que nous en disons. En marge des niveaux lexical et syntaxique qui nous concernent plus particulièrement, elle se décline également aux plans morphologique (la délimitation des frontières de mots, par exemple) et pragmatique (notamment pour le calcul du contexte référentiel d'un énoncé). [Fuchs, 1996] offre, à cet égard, une classification exhaustive et pertinente du phénomène.

[7]Notons que, dans un contexte de sous-langage, les mots propres au domaine ne rencontrent pas la même ambiguïté que dans un contexte général (*cf.* Chapitre 4). Sur un corpus de 929 occurrences du verbe *céder*, nous n'avons rencontré que 5 occurrences ne ressortissant pas à l'emploi financier. Les 1029 occurrences de *cessions* sont quant à elles unanimement nominales. Toutefois, le phénomène demeure très important pour l'ensemble du lexique restant et ne peut bien entendu être ignoré.

[8]Par *exact*, nous entendons qu'aucun choix ne sera pris en cas d'ambiguïté. De cette manière, nous n'introduisons pas d'erreurs potentiellement pénalisantes pour le reste de l'analyse. À l'inverse, la procédure approximative consiste à faire un choix, qui dans certains cas pourra s'avérer inexact.

3.2.1 Étiquetage statistique

L'annotation statistique repose généralement sur un ensemble de règles contextuelles, basées sur l'observation statistique de données textuelles et vise la résolution exhaustive de l'ambiguïté lexicale (*cf.* [Brill, 1992; Schmid, 1994]) de même que la prise en compte des mots inconnus (les noms propres, par exemple, ou encore les termes très spécialisés n'appartenant pas au langage courant).

Considérons la phrase suivante : *La banque écossaise a acquis pour 285 millions de dollars l'établissement américain CB*. En considérant la complexité lexicale dans sa totalité, nous recensons 608 256 combinaisons syntaxiques possibles[9] (*cf.* Tableau 3.2). Cette ambiguïté est pour une bonne part artificielle, toutes ces combinaisons ne sont pas syntaxiquement acceptables. Des règles très simples peuvent élaguer considérablement le champ des possibles. Prenons pour exemple le groupe nominal sujet de la phrase, *la banque écossaise*. Nous dénombrons, dans le tableau 3.2, 88 analyses différentes (4 * 11 * 2). Le mot *la* peut se comprendre comme un nom singulier ou pluriel, un pronom ou un déterminant. Le mot *banque*, s'il n'est pas substantif, peut être une forme verbale transitive ou intransitive. Le mot *écossaise*, bien qu'il puisse être nom ou adjectif, restera féminin dans les deux cas.

1. ^ [**<PRO+PpvIL>**] **<V :W>**[10] **:** un pronom personnel, premier élément d'une phrase, est nécessairement suivi d'un infinitif.
2. [**(<N :x>+<PRO :x>)**] **<MOT>*** [**<V :x>**] **:** un verbe à la nième personne (notée *x*) demande un sujet de la même personne.
3. [**<DET>**] **<MOT>* <V> :** un déterminant ne peut être directement suivi d'un verbe.

Par application de ces seules règles, l'ambiguïté se voit déjà fortement réduite. Il est entendu que cette grammaire est partielle ; nous constatons toutefois que beaucoup d'ambiguïtés peuvent être très facilement évacuées à un stade pré-syntagmatique[11]. Ainsi, le mot *la*, selon la règle 1, ne peut être pronom en début de phrase s'il n'est pas suivi d'un infinitif. De la même manière, la deuxième règle nous permet de limiter les formes verbales potentielles de *banque* aux seules troisièmes personnes. Bien entendu, il nous reste encore 8 analyses différentes pour une

[9] *Ce nombre s'obtient en effectuant le produit des degrés d'ambiguïté de chaque mot. Pour des phrases longues, la croissance du degré est exponentielle, on parle d'explosion combinatoire. Toutefois, cette manière de compter représente mal les phénomènes linguistiques en jeu. En effet, les ambiguïtés sont souvent locales, c'est-à-dire limitées à une partie de la phrase, voire à des mots simples dans le cas des homographies, elles n'interfèrent alors aucunement entre elles.* [Gross, 2001 : p. 30]

[10] Nous indiquons entre crochets l'(es) élément(s) sur le(s)quel(s) pèse la contrainte.

[11] Il ne s'agit pas là d'une solution propre aux méthodes statistiques. [Laporte et Monceaux, 1999] et [Dister, 2001], par exemple, lèvent l'ambiguïté des mots grammaticaux et lexicaux en décrivant, au sein d'automates, le contexte d'apparition des séquences ambiguës. Ces automates sont appliqués au texte avant toute analyse syntaxique.

La	4	.N+z1 :ms :mp
		le.DET+Ddef+z1 :fs
		le.PRO+PpvLE+z1 :3fs
banque	11	.N+Conc+z1 :fs
		banquer.V+i+U+z1 :P1s :P3s :S1s :S3s :Y2s
		banquer.V+t+z1 :P1s :P3s :S1s :S3s :Y2s
écossaise	2	,.N+z3 :fs
		écossais.A+d+z1 :fs
a	9	.N+z1 :ms :mp
		.XI+z1
		avoir.V+aux+z1 :P3s
		avoir.V+en+i+U+z1 :P3s
		avoir.V+i+U+z1 :P3s
		avoir.V+t+U+z1 :P3s
		avoir.V+t+z1 :P3s
		avoir.V+y+i+U+z1 :P3s
acquis	8	.A+d+z1 :ms :mp
		.N+z1 :ms :mp
		acquérir.V+t+z1 :Kms :Kmp :J1s :J2s
pour	2	.N+z1 :ms
		.PREP+z1
285	1	NBR
millions	1	million.N+Dnom+z1 :mp
de	3	.DET+Dind+z1
		.PREP+z1
		.XI+z1
dollars	1	dollar.N+Unit+z1 :mp
l	4	.N+z1 :ms :mp
		le.DET+Ddef+z1 :ms :fs
établissement	2	.N+Abst+z1 :ms
		.N+Abst+z1 :ms
américain	2	.A+d+z1 :ms
		.N+z1 :ms
CB	1	NP

Tableau 3.2 : Exemple d'ambiguïté lexicale

unique analyse correcte. Cependant, et au vu de l'aspect plus que fragmentaire de notre jeu de contraintes, il nous paraît tout à fait envisageable de désambiguïser efficacement le français de cette manière. Il restera souvent plus d'une analyse, mais d'autres solutions peuvent se combiner à celle-ci et apporter une amélioration substantielle.

En marge de ces règles syntaxiques, la morphologie permet également de résoudre certaines situations complexes. [Laporte, 2001] cite, à titre d'illustration, le suffixe *–ives* que l'on retrouve principalement dans les adjectifs féminins pluriels (*cf.* Exemples 55 à 57). Bien entendu cette règle, comme l'ensemble des règles de ce type, est une approximation d'un phénomène plus complexe. Il existe donc des contre-exemples (*cf.* Exemples 58 et 59).

(55) Le taux d'utilisation des capacités **productives** est resté inchangé à 75,6% par rapport au mois de janvier.

(56) Galeries Lafayette (+3,1%) : le groupe, via sa filiale Laser, s'est offert 60% du

capital du n°1 des cartes **privatives** aux Pays-Bas.

(57) Les deux groupes mènent des discussions **exclusives** et devraient annoncer un accord sous peu.

(58) Cependant, selon les firmes du secteur, ce projet de loi, qui n'a pas encore été ratifié au parlement, pourrait être annulé du fait de l'approche des **législatives** allemandes.

(59) Time Warner étudie différentes **alternatives** à celle d'une vente de la Warner Music à Emi Group.

Enfin, les probabilités peuvent également résoudre certains conflits. Un étiqueteur stochastique désambiguïse chaque mot par étude du contexte grammatical proche. Généralement, on privilégie, pour un mot ambigu donné, l'étiquette la plus fréquente dans un contexte grammatical précis. Une telle solution demande donc un entraînement préalable sur corpus étiqueté (manuellement) et peut mener à certains choix erronés. Notons que la taille du contexte, ou de la fenêtre de lecture, peut varier considérablement selon les choix d'implémentation. Si une solution à base de bigrammes est le plus souvent privilégiée, rien n'empêche d'élargir la fenêtre aux trigrammes, *etc.* De la même manière, si les probabilités sont habituellement calculées au regard des catégories grammaticales uniquement, certaines applications complètent ces estimations par l'étude fréquentielle des occurrences de mots graphiques sans prendre en compte cette fois l'abstraction grammaticale.

Les résultats de ce type d'étiqueteur sont actuellement de l'ordre de 95% de réussite. Le *TreeTagger*, développé à l'université de Stuttgart, affiche, par exemple, une précision de 96,36% (*cf.* [Schmid, 1994]). Par ailleurs, pour la phrase que nous explorons à titre d'exemple, la précision du *TreeTagger* est quasi maximale (*cf.* Tableau 3.3). Ce résultat peut sembler satisfaisant si nous considérons l'étiquetage en soi mais doit toutefois être circonstancié. 96,36% correspond en effet à une erreur tous les 27 mots.

3.2.2 Étiquetage linguistique

Cette méthode procède par application de dictionnaire et refuse, quant à elle, toute levée d'ambiguïté lors de l'étiquetage afin d'assurer l'exactitude des procédures ultérieures (si nous ne prenons pas de décision, nous ne sommes pas amenés à prendre une mauvaise décision). Une solution de ce type entraîne inévitablement une explosion combinatoire, plus ou moins importante selon la taille de la phrase mais également selon le niveau de granularité de l'information associée à chaque mot.

Unitex, que nous utilisons dans cette première expérience, étiquette un texte par application des dictionnaires de mots simples (DELA) et composés (DELAC) compilés et distribués par l'ins-

La	DET+Ddef	le
banque	N	banque
écossaise	A	écossais
a	V+aux :P	avoir
acquis	V :K	acquérir
pour	PREP	pour
285	NUM	@card@
millions	N	million
de	PREP	de
dollars	N	dollar
l'	DET+Ddef	le
établissement	N	établissement
américain	A	américain
CB	PRO+Poss	<unknown>
.	PUNC	.

Tableau 3.3 : Sortie du TreeTagger

titut Gaspard-Monge[12] (*cf.* [Courtois, 1990; Silberztein, 1993]) ainsi qu'un certain nombre de dictionnaires plus spécialisés (*i.e.* prénoms, lieux, *etc.*). Ces dictionnaires, créés à l'origine par le LADL, s'organisent, pour chaque langue, en quatre formats distincts (plus ou moins aboutis selon la langue, le français étant la langue la plus développée à ce jour) : le DELAS, dictionnaire morphologique, le DELAP, dictionnaire phonémique, le DELAF, dictionnaire de formes fléchies et le DELAR, dictionnaire des radicaux de la langue. Ces quatre formats renvoient à deux niveaux de description : morphologique pour les DELAS, DELAP et DELAR et orthographique pour le DELAF. Le niveau syntaxique, considérant la structuration des unités orthographiques dans un contexte de phrase, est, pour sa part, décrit au niveau des tables de lexique-grammaire. Notons que notre extracteur limite ses ressources linguistiques aux dictionnaires DELAF et DELAC.

D'un point de vue informatique, une entrée simple consiste en une suite de lettres d'un alphabet donné ne contenant pas de séparateurs (espace blanc, apostrophe, trait d'union, *etc.*). Cette conception conduit à certaines inconséquences d'un point de vue sémantique. Ainsi, nous observons des entrées comme *aujourd*, *parce*, *etc*. qui, si elles ne forment pas une unité sémantique cohérente, sont graphiquement homogènes. Un code renseigne toutefois la non-autonomie de ces constituants (*XI* dans le DELAF). Les mots composés sont quant à eux formés d'une suite de mots simples, au sens que nous venons de discuter.

[12] `http://infolingu.univ-mlv.fr/english/DonneesLinguistiques/Dictionnaires/download.html`

Les dictionnaires DELA(C)S.
Au regard de son correspondant fléchi, un dictionnaire morphologique contient relativement peu d'informations morpho-syntaxiques. Ces dernières sont injectées, *a posteriori*, lors du processus de flexion automatique. Ce processus est conditionné par les différents codes associés à l'entrée. Chaque entrée du dictionnaire est augmentée de deux informations : la catégorie grammaticale et le mode de flexion. Ces deux informations sont concaténées en un *code morphologique* et renvoient à une classe flexionnelle particulière.

– `déjeuner,.N1.V3`
– `secrétaire,.N1.N31`
– `avant,.N1.A80.PREP.ADV`

Le dictionnaire DELAP.
Ce dictionnaire augmente chaque entrée du DELAS d'une information caractérisant la signature phonémique du mot, c'est-à-dire sa ou ses prononciations (ainsi que celle des formes fléchies qu'il peut engendrer) et les possibilités d'élision et de liaison.

– `abasourdir,abazurdir.V18,abasurdir.V18`
– `discothèque,diskotek.N21`
– `mère,mer.N21`

Le dictionnaire DELAR.
Le DELA des radicaux de la langue se présente sous la forme d'une bibliothèque d'automates formalisant la structure des mots et leurs relations morphologiques. Ainsi, chaque automate décrit une entrée du DELAS et l'ensemble de ses dérivés morphologiques.

Les dictionnaires DELA(C)F.
Ce dictionnaire renseigne, pour chaque lemme du DELAS, les formes fléchies et les variantes orthographiques possibles. Le dictionnaire DELAF du français compte quelque 800 000 entrées auxquelles est associé un ensemble d'étiquettes morphologiques, syntaxiques et sémantiques.

– `financent,financer.V+t+z1 :P3p :S3p`
– `abactérien,.A+z3 :ms`
– `zygolophodons,zygolophodon.N+Anl+z3 :mp`

La première zone, précédant la virgule, désigne la forme fléchie, c'est-à-dire l'entrée lexicale proprement dite. Elle est suivie du lemme (zone précédant le point). Si ce dernier n'offre pas de différences orthographiques par rapport à la forme fléchie, la zone est vidée. Suit ensuite la partie du discours augmentée, le cas échéant, d'une information de sous-catégorisation ainsi que d'indications distributionnelles et sémantiques (séparées par un signe '+'). Les codes flexionnels sont recensés en fin d'entrée, chaque jeu d'étiquettes étant introduit par le signe ' :'.

3.2.3 Quelle solution pour l'étiquetage ?

Une solution statistique à l'étiquetage nous permettrait, *a priori*, d'évacuer un grand nombre de problèmes. En effet, si un étiquetage à base de dictionnaires ne peut affecter une catégorie grammaticale aux mots inconnus, un étiqueteur stochastique, par contre, sur base du contexte et d'une information fréquentielle, prend une décision qui, bonne ou mauvaise, permet de poursuivre l'analyse de la phrase et mener, le cas échéant, à une solution correcte. De plus, en évacuant l'ambiguïté catégorielle dès le niveau morpho-syntaxique de l'analyse, nous devrions éviter l'explosion combinatoire des résultats de même que les regroupements syntagmatiques erronés. Nous pouvons donc nous demander si cette solution n'est pas idéale. Nous avons abordé cette question dans un article écrit conjointement avec Cédrick Fairon et Sébastien Paumier : *Can We Parse Without Tagging* ? (*cf.* [Fairon *et al.*, 2005])[13].

Notre expérience repose sur un corpus de spécialité de 1 500 000 mots, consistant en un ensemble de textes trouvés sur Internet et spécialisés dans le domaine de l'énergie. Pour évaluer les deux méthodes d'étiquetage, nous avons sélectionné cinq verbes illustrant les principales structures syntaxiques du domaine :

– *alimenter* : N_0 *V* N_1 *en* N_2 *à partir de* N_3
– *chauffer* : N_0 *V* N_1 *à partir de* N_3
– *consommer* : N_0 *V* N_1
– *contenir* : N_0 *V* N_1
– *produire* : N_0 *V* N_1 *à partir de* N_3

Bien sûr, notre corpus contenait d'autres structures mais nous étions uniquement intéressés par celles présentant des constituants sémantiquement pertinents pour l'extraction.

Afin d'étiqueter notre corpus, nous avons utilisé le *TreeTagger*[14]. Le texte obtenu en sortie est ensuite transformé en un format compatible avec *Unitex*.

L'étiquetage du *TreeTagger* produit un fichier dans lequel chaque ligne contient un mot du texte associé à une information grammaticale et un lemme (*cf.* élément 2 de la figure 3.2). Notons

[13]Cette recherche s'est inscrite dans un projet plus large visant à créer, pour le bureau Van Dijck, un moteur d'extraction d'information dans le domaine de l'énergie. Il s'agissait alors d'estimer, en précision et rappel, l'apport d'une désambiguïsation en amont de l'analyse syntaxique. Par conséquent, et bien que le domaine étudié ne soit pas celui de la finance, nous croyons les résultats particulèrement pertinents pour notre propos. Par ailleurs, ces résultats tendent à montrer que la solution que nous apportons à l'extraction, de même que les problèmes qui se posent, ne sont pas propres à un domaine donné et sont généralisables à la langue dans sa globalité.

[14]La plupart des étiqueteurs affichent des performances équivalentes et ce dernier est non seulement disponible gratuitement (`http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html`) mais également très simple d'emploi. Nous avons, par ailleurs, développé une interface d'utilisation du *TreeTagger* en ligne : `http://cental.fltr.ucl.ac.be/~pat/tagger/`. Celle-ci permet à l'utilisateur d'envoyer au serveur un texte à étiqueter. Une fois le travail réalisé, le fichier résultant lui est envoyé par courrier électronique.

Figure 3.2 : Étiquetage et conversion

que nous ne disposons d'aucune information flexionnelle.

Avant de proposer le résultat à *Unitex*, nous devons le linéariser et adapter le jeu d'étiquettes grammaticales au format DELAF (*cf.* élément 3 de la figure 3.2).

Le texte non étiqueté conserve, quant à lui, toute la complexité lexicale. Le DELAF nous permet d'associer à chaque mot l'ensemble des étiquettes possibles. Nous ne proposons donc aucune information *a priori* contrairement aux méthodes statistiques. L'ambiguïté de la phrase d'exemple est illustrée par l'automate de la figure 3.3.

	PRÉCISION	**RAPPEL**
Texte désambiguïsé	92,28% (598/648)	47,84% (598/1250)
Texte non désambiguïsé	88,67% (493/556)	39,44% (493/1250)

Nous n'aborderons pas la méthode d'extraction qui sera détaillée au point 3.3. Les résultats obtenus au départ des deux méthodes montrent une faible différence de précision – 3,6% – en faveur d'une désambiguïsation dès l'étiquetage. Ce résultat confirme l'idée harrissienne, qui sous-tend notre méthode, d'une désambiguïsation naturelle par le contexte, nos graphes d'extraction formalisant chaque prédicat au sein d'une phrase. De plus, elle semble indiquer, du point de vue de la précision, que l'étiquetage, au sens du *TreeTagger*, n'est pas absolument nécessaire.

Parallèlement, l'analyse des résultats a pu dégager deux raisons principales à la différence de rappel (8,4%) :

Ressources lexicales incomplètes

Ce problème est particulièrement présent dans un domaine tel que celui de l'énergie dont le lexique est extrêmement spécialisé et sujet à l'innovation linguistique. Il est évident, dès lors, qu'un dictionnaire de la langue standard ne peut couvrir tous les besoins d'une spécialité comme l'énergie.

Description syntaxique partielle

Au moment de cette expérience, notre extracteur ne disposait pas de grammaires permettant la reconnaissance des nombres en chiffres. Ce manque ne pose pas de réels problèmes si nous utilisons un corpus étiqueté puisque l'étiqueteur va précisément les identifier comme nombre. Cependant, si nous basons notre extraction sur un corpus brut, c'est-à-dire non désambiguïsé, leur seule présence empêche toute reconnaissance.

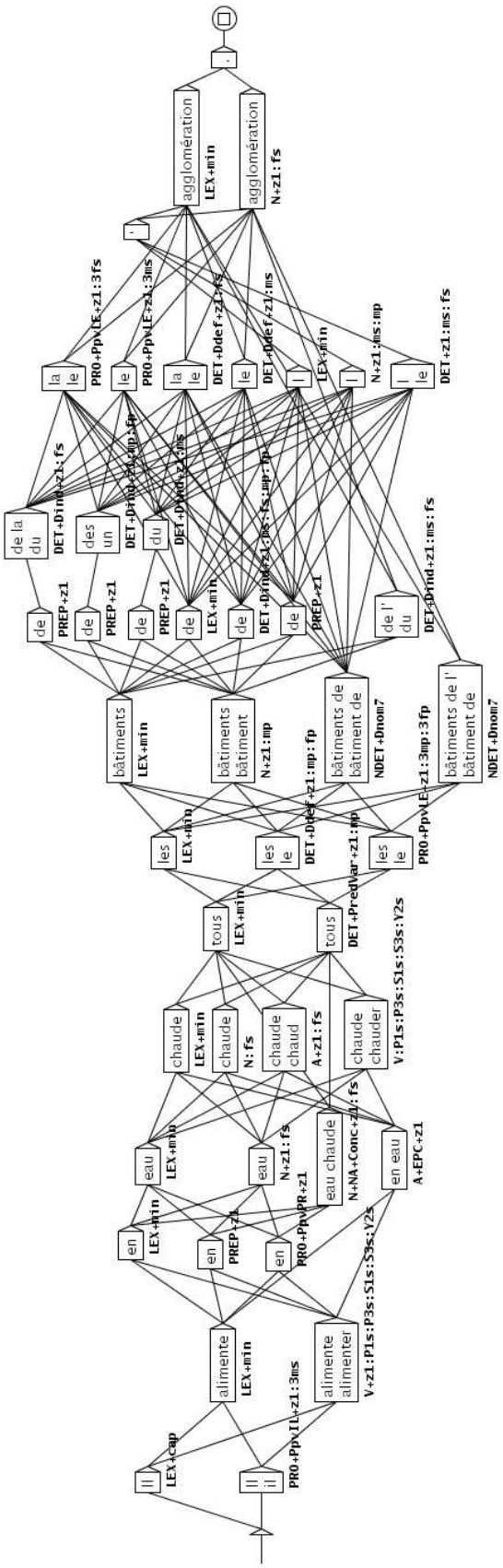


Figure 3.3 : Texte ambigu après consultation des dictionnaires

Bien que nous ayons pu expliquer la différence de rappel entre les deux solutions d'étiquetage que nous proposons, il n'en reste pas moins qu'il est insuffisant. Ces scores sont en grande partie imputables à deux choix d'implémentation. D'une part, et c'est la principale raison, nous privilégiions, dans cette expérience, la précision de l'extraction ; or, nous l'avons vu au chapitre 2, une augmentation du rappel entraîne inévitablement une diminution de la précision. D'autre part, nous avons limité notre description linguistique à la phrase simple et aux transformations unaires qui ne reflètent pas toujours la complexité syntaxique des énoncés en situation réelle d'énonciation.

biocombustible,.N :ms
biocombustibles,biocombustible.N :mp
biodiesel,.N :ms
biodiesels,biodiesel.N :mp
biométhanisation,.N :fs
biométhanisations,biométhanisation.N :fp
cocombustion,.N :fs
cocombustions,cocombustion.N :fp

Tableau 3.5 : Extrait du dictionnaire de l'énergie pour le français

Pour conclure cette expérience, les résultats montrent que, moyennant certaines améliorations au sein de nos automates et une spécialisation au domaine étudié de nos ressources lexicales (*cf.* Tableau 3.5), une solution à base de dictionnaires peut aisément se passer d'une désambiguïsation au niveau morpho-syntaxique sans diminution de performance. Nous rejoignons en cela le constat fait dans [Appelt *et al.*, 1993].

3.3 Extraction, tables et graphes paramétrés

Le contenu d'une table de lexique-grammaire peut être exploité en recourant au principe des graphes paramétrés décrit par [Roche, 1993], [Senellart, 1999] et [Paumier, 2003]. Un tel graphe rend compte des constructions linguistiques formalisées dans la table à laquelle il renvoie. Chaque construction est ainsi représentée par un chemin au sein du graphe (*cf.* éléments 2 et 3 de la figure 3.4).

Le graphe paramétré contient un ensemble de variables correspondant aux colonnes de la table et dont la valeur dépend directement de l'entrée lexicale concernée. Le nom des paramètres – *@A, @B, etc.* – renvoie aux colonnes de la table. Ainsi, *@A* désigne la première colonne, *@B* désigne la seconde colonne, *etc.* Dans la figure 3.4, par exemple, le graphe paramétré numéroté 2 utilise la table 1 pour produire le graphe 4.

Pour chacune des entrées de la table (*cf.* élément 1 de la figure 3.4), c'est-à-dire pour chacun des prédicats, un graphe (ou grammaire) est généré par substitution des variables au sein du graphe

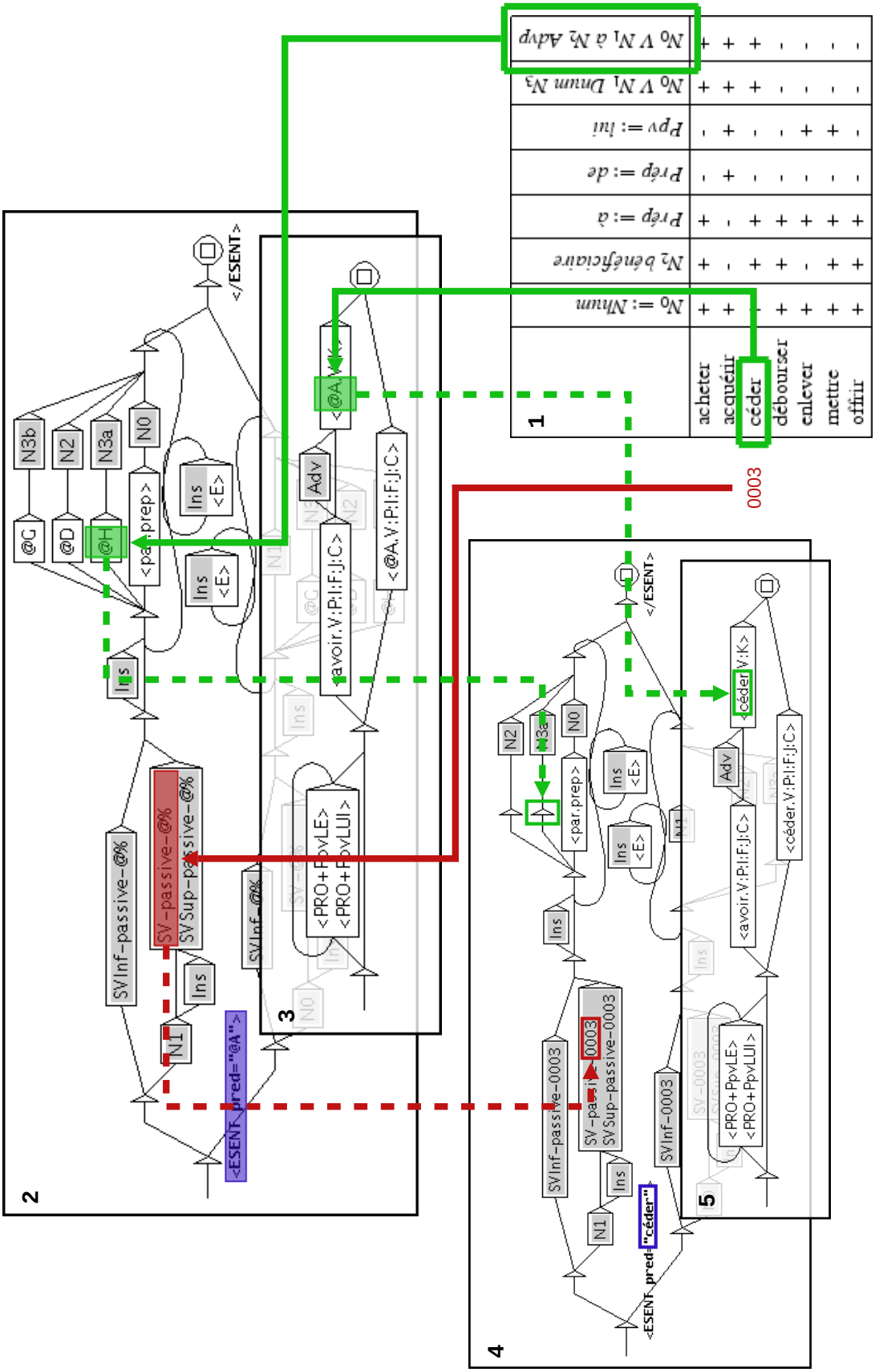


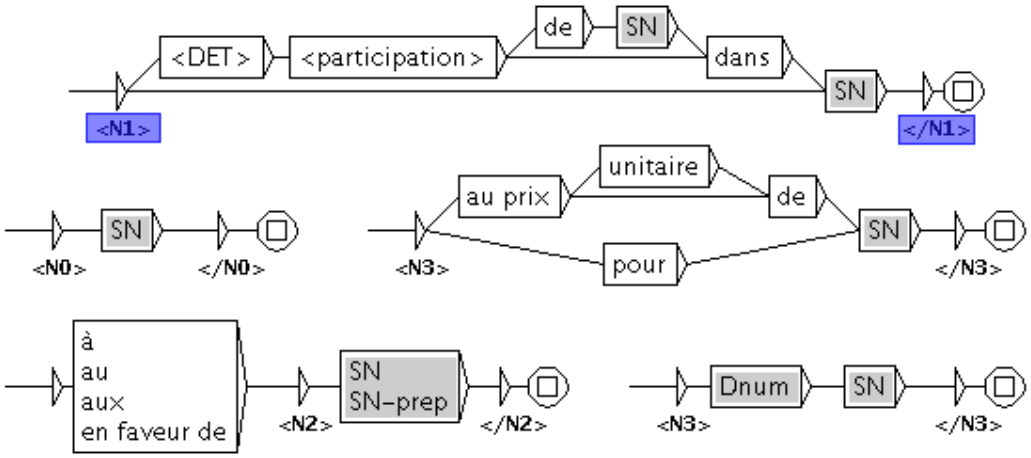
Figure 3.4 : Processus de conversion des graphes paramétrés

paramétré (*cf.* éléments 4 et 5 de la figure 3.4). Nous avons vu qu'une table peut contenir tantôt un signe positif ou négatif, tantôt des éléments lexicaux. Dès lors, la substitution opère comme suit :

– si la variable *@X* renvoie à un signe −, le chemin est supprimé ;
– si elle renvoie à un signe +, le chemin est conservé ;
– si elle renvoie à un élément lexical, elle est remplacée par ce dernier.

Une procédure automatique génère, sur base de ces trois règles, un reconnaisseur formalisant les seules constructions grammaticales et acceptables des différentes entrées.

L'analyse syntaxique réalisée au départ des graphes conditionnés par les tables ne fait aucune différence entre analyse syntaxique et identification de motifs (*pattern matching*, *cf.* [Paumier, 2003]). Nous utilisons la fonction de reconnaissance d'*Unitex* (*Locate*) pour identifier, au sein d'un texte, l'ensemble des séquences pertinentes. Si les séquences sont reconnues par le graphe, nous pouvons dire que nous les avons *parsées* (*e.g.* analysées syntaxiquement). Il est, en effet, possible d'ajouter au graphe des étiquettes qui baliseront, en sortie, les différents syntagmes reconnus (*e.g.* `<N1>` et `</N1>`).



<ESENT pred="céder"> <N0> Aventis </N0> a cédé <N1> un traitement de l'asthme </N1> au <N2> laboratoire Kos Pharmaceuticals </N2> <N3> un montant de 200 millions de dollars </N3> </ESENT>

Figure 3.5 : Étiquetage XML

Notre procédure d'extraction proprement dite tire parti de cette dernière fonction. En cours de reconnaissance, des balises XML sont insérées dans le corps du texte, et plus exactement dans les phrases reconnues (*cf.* Figure 3.5). Dès lors, l'extraction en tant que telle ne diffère pas réellement du processus d'analyse syntaxique.

Il nous suffit ensuite d'isoler, pour chaque phrase reconnue, les syntagmes étiquetés afin de récupérer l'information recherchée ainsi que leurs dépendances. Nous utilisons, pour cela, une feuille de style XSL[15] qui nous permet de transformer le format de sortie XML en un format plus propice à la lecture. Si nous avons opté pour le HTML, facilement intégrable à une application Web, le XML peut, par le biais de la technologie XSLT, se transformer en n'importe quel format.

3.4 Résultats en chiffres

Notre stratégie d'évaluation rejoint celle que nous avons mis en œuvre afin d'évaluer notre extracteur pour le domaine de l'énergie (*cf.* Section 3.2.3). Nous avons sélectionné huit verbes (*cf.* Tableau 3.6) ressortissant au domaine des transactions financières et représentatifs de la structure élémentaire dative de la table *36dt*. Il est évident que ces prédicats ne suffisent pas à couvrir totalement le domaine mais notre objectif n'est pas pas là. Des études telles que [G. Gross et Geunthner, 2002] et [Poibeau, 2003] ont déjà montré qu'il était possible de formaliser un domaine de spécialité au départ d'une base prédicative, sous forme de classes d'objets, pour le premier, ou de simple liste, pour le second. De notre côté, nous cherchons à estimer la pertinence de la phrase élémentaire comme patron d'information. En ce sens, la structure dative de la table 36DT est particulièrement intéressante tant du point de vue informationnel que syntaxique. En effet, elle couvre la plupart des informations du concept de transaction (*cf.* Figure 2.6) et offre une structure syntaxique suffisamment complexe pour espérer isoler l'ensemble des problèmes.

PRÉDICATS
acheter
acquérir
céder
offrir
racheter
reprendre
revendre
vendre

Tableau 3.6 : Prédicats retenus pour l'évaluation

Le corpus que nous avons utilisé compte 54 618 mots et est composé de 1 761 phrases contenant chacune l'un des huit prédicats évalués. Notre automate d'extraction (*cf.* Figure 3.6) formalise les propriétés distributionnelles et transformationnelles recensées dans [Guillet et Le-

[15] *XSL is a family of recommendations for defining XML document transformation and presentation. It consists of three parts : XSL Transformations (XSLT), a language for transforming XML ; the XML Path Language (XPath), an expression language used by XSLT to access or refer to parts of an XML document ; (XPath is also used by the XML Linking specification) ; XSL Formatting Objects (XSL-FO), an XML vocabulary for specifying formatting semantics.* (`http://www.w3.org/Style/XSL/`)

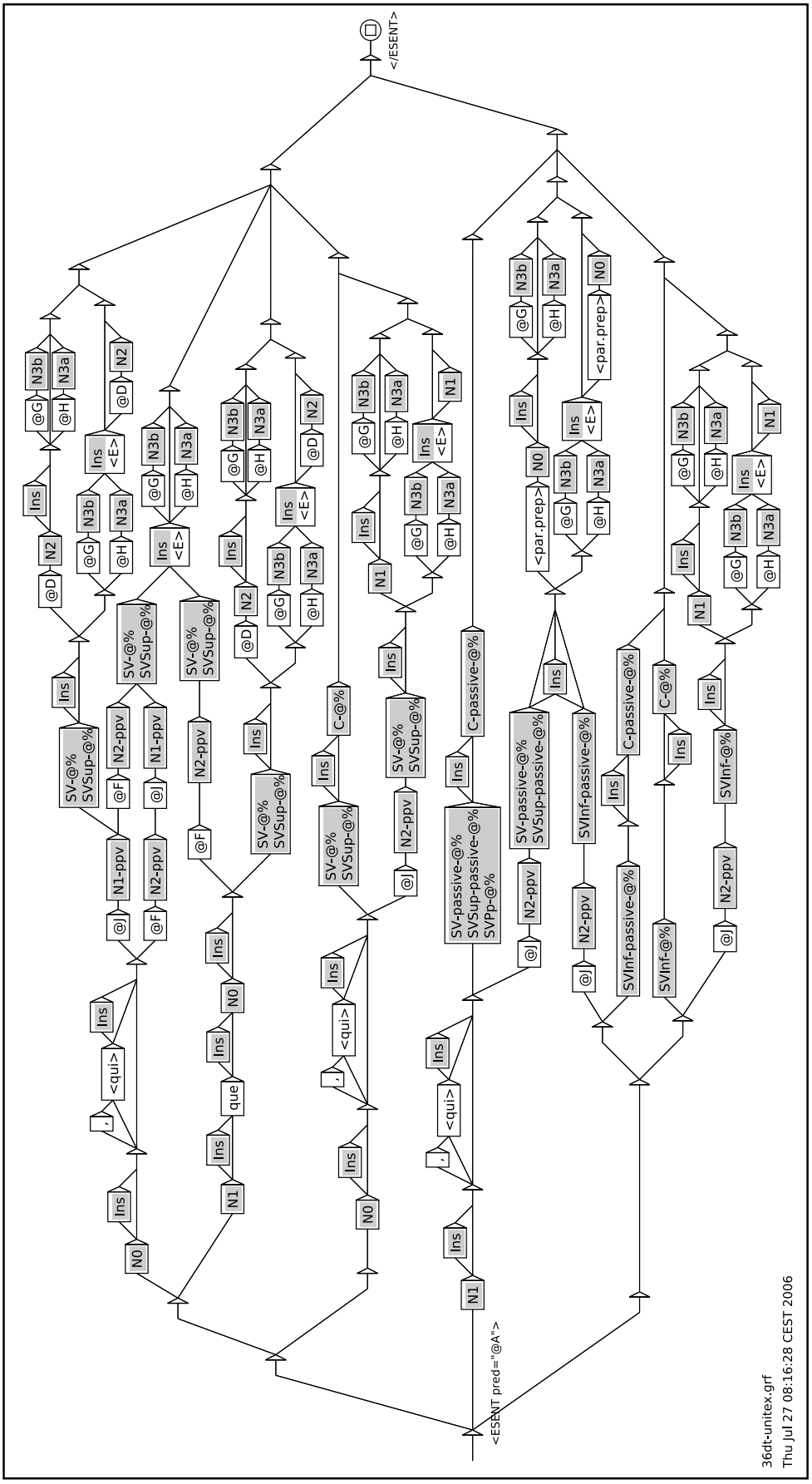


Figure 3.6 : Automate d'extraction principal

clère, 1992]. Dans cette première expérience, nous considérons uniquement les transformations unaires[16]. Nous partons du principe que cette structure reste cohérente (*e.g.* visible) au sein des phrases complexes qui, théoriquement, résultent de la simple combinaison de plusieurs structures élémentaires.

Nous proposons, dans le tableau 3.7, les résultats de l'évaluation en termes de rappel et de précision. Ces résultats pourraient, à tort, être perçus comme insatisfaisants. Rappelons que nous cherchons à isoler la difficulté de l'extraction d'information et que, par conséquent, nous avons envisagé la table 36DT telle qu'elle est décrite dans [Guillet et Leclère, 1992] sans considérer une quelconque spécialisation lexicale ou syntaxique au domaine étudié. La majorité des erreurs, nous le verrons, sont imputables à la naïveté de cette approche et seront donc aisément corrigées.

PRÉCISION	RAPPEL
87,32%	60,7%

Tableau 3.7 : Résultats d'évaluation

Les erreurs de rappel sont dues principalement à un manque de couverture aux niveaux lexical (unités de mesure, noms d'entreprises, *etc.*) et syntaxique (coordination et apposition). Les problèmes de précision relèvent quant à eux majoritairement de l'ambiguïté d'attachement. Nous étudions ces erreurs au point suivant.

3.5 Limites et perspectives

Dans cette section, nous identifions et expliquons les problèmes à l'origine des erreurs de rappel et de précision isolés par notre évaluation. Nous organisons les erreurs en trois catégories selon qu'elles relèvent des plans morpho-syntaxique, syntaxique ou sémantique. Dans la mesure où une erreur concerne un point extérieur à notre hypothèse générale, nous suggérerons une solution partielle. Les erreurs restantes seront traitées dans la troisième partie de cette étude.

3.5.1 Le plan morpho-syntaxique

La couverture lexicale

Bien que larges et détaillées, les ressources dictionnairiques développées au sein du LADL, ne peuvent prétendre couvrir l'ensemble de la langue qui évolue constamment. Selon le domaine étudié, nous rencontrons de nombreux néologismes et termes spécialisés de même que des noms

[16]Notons que nous intégrons également les relatives. Pour les prédicats en position subordonnée nous ne considérons toutefois pas la proposition complète. Nous limitons en effet notre formalisation aux séquences N_1 *que* N_0 V *à* N_2 et N_0 *qui* V N_1 *à* N_2, c'est-à-dire aux constituants ressortissants aux prédicats de transaction.

propres qui ne se retrouvent pas dans les dictionnaires fournis avec *Unitex* (*cf.* Tableau 3.8).

TERMINOLOGIE FINANCIÈRE			
actionnarial	dédupliquer	multivendeur	surcapitaliser
backlog	défaisance	nasdaq	surallocation
bancassurance	gatekeeper	onshore	suramortissement
bancassureur	gearing	origination	surcapacitaire
bilanciel	génériqueur	outsourcer	surperformance
billetique	géodépendant	outsourceur	surperformer
booster	géolocalisation	outsourcing	surpondérer
boursièrement	hub	polyservice	sursouscrire
câblopérateur	infogérance	préouverture	taxotère
cashflow	infogéreur	refacturation	technoventure
closing	infomercial	refinanceur	télébillétique
coactionnaire	multicroissance	retail	télédéclaratif
coating	multidevise	retake	télédéclaration
contracteur	multigestion	roaming	télédiffuseur
corporate	multigestionnaire	scoring	télémarketing
débouclage	multijoueur	streamer	télépéage
débouclement	multimarque	streaming	téléperformance
décertification	multiservice	slot	télépilotage
décollecte	multithématique	surallocation	téléservice

ERREURS ORTHOGRAPHIQUES			
administation	cedait	dégats	recettespublicitaires
adminsitration	cèdait	égalemant	dollarsen
admnistration	céde	egalement	dollarspour
affairs	comamndes	percéesignificative	dudossier
affaures	commmerciale	prochainesaugmentation	engénéral

Tableau 3.8 : Exemples de mots inconnus

En marge de la terminologie du domaine, certaines erreurs orthographiques peuvent compromettre l'analyse. C'est un problème récurrent en extraction d'information. [Appelt *et al.*, 1993], par exemple, intègre à la procédure de pré-traitement des documents une étape de correction des erreurs orthographiques. De notre côté, nous avons pu constater que ces erreurs correspondent pour la plupart à une distance de *Levenstein* de 1. Greffé à un analyseur destiné à la production, un algorithme de ce type nous permettrait très certainement de prendre une décision dans le cas de ces mots inconnus. Nous avons d'ailleurs exploité cet algorithme aux noms propres du type Netan*yahou*, Netan*yahu*, Netan*yehou* dans [Fairon et Watrin, 2003]. Cette idée n'alourdit pas la chaîne de traitement.

D'autres erreurs peuvent également être imputables à l'agglutination de mots distincts, certainement due au format HTML original de nos corpus. La conversion en texte des documents web n'est pas un problème trivial. Il existe de nombreuses solutions performantes, dont celle que nous avons mise au point, mais il restera toujours des erreurs résiduelles.

L'ambiguïté catégorielle.

Ce phénomène entraîne des erreurs spécifiques, engendrées pour la plupart par l'heuristique utilisée lors de la reconnaissance. Nous privilégions les séquences les plus longues (option *longest match* d'*Unitex*). Les regroupements sont donc calculés de manière à couvrir le plus long segment textuel possible.

Le mot *est*, par exemple, peut à la fois être un adjectif (point cardinal) ou un verbe (troisième personne du singulier du verbe *être*). Dès lors, si nous considérons la première solution, nous obtiendrons, pour la phrase reprise ci-dessous, et par application de l'heuristique du plus long motif, un parenthésage syntagmatique erroné.

> [...] <N0>le groupe nucléaire français est le seul à avoir déposé une</N0> offre<N1> de rachat </N1>.

Dans cet exemple, *est* est une forme verbale, le syntagme nominal correct devrait être *le groupe nucléaire français* et *offre* devrait être considéré comme un substantif.

S	→	SN SV (SN|SP)*
SP	→	(PREP|PREDET) SN
SN	→	(DET ? SA ? N SA ?) | (PRO)
SV	→	ADV* V
SA	→	ADV* A

Tableau 3.9 : Grammaire jouet

De telles erreurs se présentent quand le texte n'est pas ou mal désambiguïsé. Toutefois, l'ambiguïté n'est pas seule responsable. En limitant, au sein d'une grammaire (*cf.* Tableau 3.5.1), l'information grammaticale aux catégories syntaxiques sans prendre en compte la flexion, nous augmentons inévitablement le degré d'ambiguïté structurale. Considérons, pour nous en persuader, le syntagme nominal *le groupe nucléaire français*.

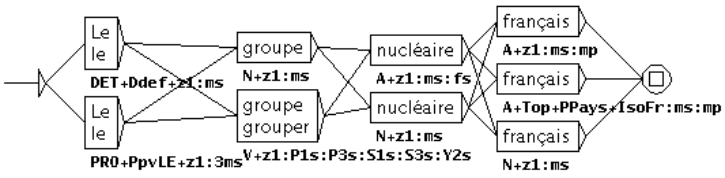


Figure 3.7 : Complexité lexicale : le groupe nucléaire français

L'automate de la figure 3.7 représente la combinatoire morpho-syntaxique de cet énoncé. Il est entendu que ces énoncés possibles ne sont pas tous probables. La grammaire du français n'accepte pas n'importe quelle suite grammaticale. Les règles de notre grammaire jouet (*cf.* Tableau 3.5.1) limitent la combinatoire aux énoncés suivants :

— <le.Ddef :ms> <groupe.N :ms> <nucléaire.A :ms> <français.A :ms>
— <le.Ddef :ms> <groupe.N :ms> <nucléaire.A :fs> <français.A :ms>
— <le.Ddef :ms> <groupe.N :ms> <nucléaire.A :ms> <français.A :mp>
— <le.Ddef :ms> <groupe.N :ms> <nucléaire.A :fs> <français.A :mp>
— <le.PRO :3ms> <groupe.V :P1s :P3s :S1s :S3s :Y2s> <nucléaire.A :fs> <français.N :ms>
— <le.PRO :3ms> <groupe.V :P1s :P3s :S1s :S3s :Y2s> <nucléaire.A :fs> <français.N :mp>
— <le.PRO :3ms> <groupe.V :P1s :P3s :S1s :S3s :Y2s> <nucléaire.N :fs> <français.A :ms>
— <le.PRO :3ms> <groupe.V :P1s :P3s :S1s :S3s :Y2s> <nucléaire.N :fs> <français.A :mp>

Il reste, néanmoins, certaines solutions non acceptables. En effet, si elles se conforment à l'agencement catégoriel décrit par notre grammaire, la syntaxe n'est pas respectée au niveau flexionnel. Pour notre exemple, seule la première solution respecte cette dernière contrainte.

Selon [Paumier, 2003 : p. 88], il n'est pas nécessaire de vérifier la cohérence de l'accord en genre et en nombre des différents éléments d'un syntagme. La formalisation d'une grammaire au niveau de la phrase doit nous permettre de résoudre cette ambiguïté. Il s'agit d'un point théorique couramment accepté au sein de la communauté du lexique-grammaire. Cependant, la complexité syntaxique des énoncés linguistiques en contexte réel est telle que, afin d'éviter une explosion combinatoire, nous devons lexicaliser[17] de manière conséquente les automates, ce qui ne nous semble pas toujours réaliste ni adéquat vu le problème traité.

Pour finir, l'ambiguïté catégorielle entraîne également, pour un motif d'extraction, un ensemble trop important de solutions concurrentes devant lesquelles il n'est pas toujours envisageable de faire un choix automatique.

3.5.2 Le plan syntaxique

Complexité structurale

Notre analyse repose sur la notion de phrase élémentaire et sur l'ensemble des transformations unaires que celle-ci accepte. Cependant, la réalité syntaxique des énoncés en corpus ne reflète pas toujours ce modèle. Au-delà des transformations simples ou unaires, la langue connaît un ensemble de transformations complexes ou binaires combinant plusieurs schémas phrastiques élémentaires. L'entrelacement de phrases élémentaires peut entraîner un masquage de leur signature syntaxique qui peut, si elle n'est pas prise en compte, pénaliser la reconnaissance et, par conséquent, l'extraction proprement dite.

[17] Un automate lexicalisé est un automate qui contient des transitions lexicales en marge des catégories grammaticales abstraites. Ce formalisme permet de garder un plus grand contrôle sur l'analyse mais demande également un développement plus important qui ne rejoint pas forcément les impératifs de l'extraction d'information.

(60) Dans cette optique, AXA a créé une joint-venture à 50–50 avec le groupe britannique Hammerson ***qui*** va racheter (pour 700 MF) **et** rénover (pour 800 MF) l'ex-siège social de la compagnie d'assurance.

(61) Surtout, ***la direction*** a également indiqué avoir racheté **et** supprimé 5,290 millions d'obligations convertibles à échéance 2006, émises en décembre 2000, pour un montant global de 109,6 millions d'euros.

(62) El Paso : ***l'opérateur gazier*** a obtenu un prêt de 1,2 milliard de dollars **et** vendu des actifs pour 500 millions de dollars.

(63) Walt Disney : ***le groupe de divertissement*** a annoncé étudier l'idée de fermer certains de ses magasins **ou** de vendre une partie de ses magasins de distribution qui ne sont pas rentables.

(64) La SSII française, A Novo, a annoncé que ***sa filiale Carte&Services***, avait signé trois nouveaux contrats **et** repris la majeure partie du parc monétique de la société C2S, filiale Sodifrance.

La **coordination** de prédicats verbaux peut conduire à une mise en facteur du sujet, pour la coordination de séquences verbe-compléments, d'un complément dans le cas d'une coordination sujet-verbe ou d'une combinaison des deux. Par conséquent, certains constituants essentiels de la phrase élémentaire se voient réduits à zéro pour l'un des deux prédicats coordonnés (*cf.* Exemples 60 à 64). En nous limitant aux phrases élémentaires, nous ne pouvons donc prétendre à mieux qu'une analyse partielle, si encore il y a analyse.

[Salkoff, 1979], de même que [Dominguès, 1999] pour la coordination interne au groupe nominal, proposent la décompression des segments coordonnés et la reconstruction des éléments conjonctionnels réduits à zéro. À cette fin, ils ont établi une typologie des différentes conjonctions possibles et un ensemble de règles de réécriture permettant ladite reconstruction.

Nous ne cherchons pas, de notre côté, l'analyse profonde de la phrase mais l'extraction de certains constituants. Dès lors, la reconstruction des deux segments conjoints nous semble superflue. Chaque extraction est conditionnée par la reconnaissance d'un prédicat préalablement identifié comme pertinent au regard du patron informationnel. Si ce prédicat est coordonné à un autre (relevant ou non du domaine), il nous paraît préférable d'identifier, au sein de la structure complexe, les syntagmes *zéros* sans aller jusqu'à la reconstruction du tout. Encore faut-il pouvoir, bien entendu, identifier les syntagmes communs aux deux prédicats.

> Quand deux verbes sont conjoints dans une chaîne centrale, on aboutit à une formule du type :
>
> (7) Σ {tV *et* tV′} Ω' ; exemple : *Pierre {achète et mange} des pommes*[18]

[18]Dans [Salkoff, 1979], C abstrait la chaîne complexe centrale, Σ le sujet, tV un verbe fléchi et Ω une suite

> dans l'analyse où *des pommes* est l'objet commun de *achète* et *mange*.
>
> Dans ce cas, la liste d'objets Ω' doit contenir l'intersection des compléments d'objets possibles pour tV et tV$'$, car seuls les objets communs aux deux verbes sont permis. Après avoir placé le verbe tV et éventuellement la séquence conjointe *et* tV$'$, une restriction sur la chaîne d'objets prend l'intersection des objets tV et tV$'$ (chaque verbe portant une liste d'objets dans son entrée lexicale), et seuls ces objets communs sont essayés dans Ω'. [Salkoff, 1979 : p. 34]

Nous pourrions dès lors envisager la gestion de la coordination de manière très simple en formalisant une structure coordonnée canonique au sein de laquelle nous attribuerions, sans ambiguïté, les différentes articulations syntagmatiques aux deux prédicats. Il s'agirait donc en définitive d'augmenter simplement nos graphes patrons d'un chemin décrivant cette structure.

Toutefois, cette idée, bien qu'intéressante, nous semble très théorique. Le phénomène de la coordination dépasse largement cet universel. Au sein des exemples 60 à 64, nous observons la structure générique suivante :

$$\mathrm{C} = N_0\ V_1\ (W_1)\ \textit{et}\ V_2\ (W_2)\ W$$

où, W_1, W_2 et W désignent, respectivement, les objets optionnels de V_1, V_2 et ceux potentiellement partagés par les deux prédicats.

Cette structure complexifie nettement la procédure de reconnaissance parce qu'elle montre une certaine *ambiguïté d'attachement*. Un premier ensemble de constituants, noté W_1, peut s'insérer entre les deux prédicats coordonnés (*cf.* Exemples 60, 62, 63 et 64). Cette suite d'objets est la plus évidente à traiter. Elle n'est pas, pour sa part, ambiguë et se raccroche nécessairement au verbe V_1. La véritable complexité posée par la coordination s'observe au niveau des compléments W_2 et W dont la frontière centrale n'est pas nécessairement triviale à établir.

Les exemples 60 et 61 semblent, de prime abord, confirmer l'hypothèse de [Salkoff, 1979]. Les seuls constituants présents sont ceux dont la structure est effectivement partagée par les prédicats *racheter* et *rénover*, d'une part, et *racheter* et *supprimer*, d'autre part. Ainsi, par exemple, nous ne retrouvons pas le complément *à* N_2, caractéristique de la phrase élémentaire définitoire de la classe 36DT dont relève le verbe *racheter*. Cependant, il s'agit là d'une confirmation partielle de l'hypothèse émise par Salkoff. La structure interne des compléments *pour 700 MF* et *pour 800 MF* est syntaxiquement équivalente mais, parce qu'ils renvoient à deux réalités sémantiques distinctes, ils ne peuvent être rassemblés. Dès lors, dans l'énoncé 60, seul le complément *l'ex-siège social de la compagnie d'assurance* peut être attribué aux deux prédicats.

---

d'objets.

Les phrases 62, 63 et 64 présentent, quant à elles, une structure du type N_0 V_1 W_1 *et* V_2 W_2 au sein de laquelle les séquences W_1 et W_2 ne correspondent aucunement à une quelconque intersection syntaxique des prédicats coordonnés qui partagent uniquement le sujet N_0.

Ainsi, si nous pouvons conclure à une proximité syntaxique des prédicats mis en coordination, nous ne sommes pas à même de vérifier l'équivalence suggérée par Salkoff. Le problème et la solution que nous devons envisager sont par conséquent plus complexes que ce que nous annoncions plus haut. Nous ne préconisons pas pour autant une méthode différente. Il nous paraît toujours intéressant de considérer la coordination comme un chemin supplémentaire au sein de notre automate d'extraction. La plupart des compléments ne sont pas réellement ambigus ; leur structure syntaxique, principalement l'ancrage prépositionnel des objets indirects, et leur sous-catégorisation sont suffisamment contraignantes pour éviter toute confusion. Nous reviendrons sur ce point au chapitre 6.

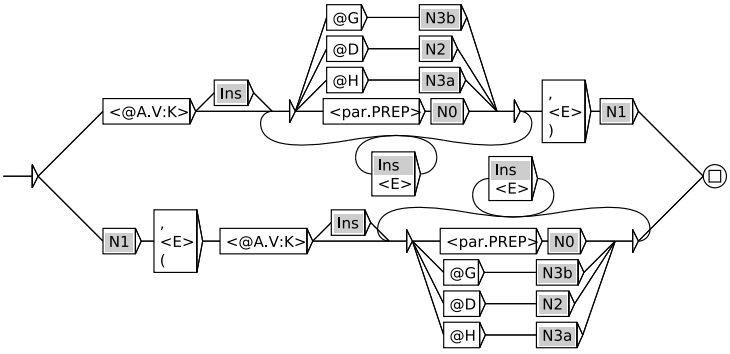


Figure 3.8 : Apposition prédicative

De même que la coordination, l'**apposition** entraîne un masquage partiel de la phrase élémentaire (*cf.* Exemples 65 à 67).

(65) Microsoft : l'éditeur de logiciels va vendre sous licence les droits d'Unix, le système d'exploitation concurrent à Windows, **acquis** auprès du groupe SCO, ce qui pourrait pousser d'autres entreprises du secteur à faire de même.

(66) Nous allons donc céder notre ligne de titres Pechiney, **acquis** il y a deux mois au prix unitaire de 43,20 euros, sans attendre le lancement de l'offre publique.

(67) Selon le magazine britannique Business, **repris** par l'agence Reuters, Nokia, fabricant finlandais de téléphones mobiles, envisagerait une OPA sur Psion, fabricant britannique de PDA (assistants personnels).

Il s'agit, dans ces exemples, de structures prédicatives participiales. L'apposition, considérée seule, ne nous permet pas d'identifier la totalité des constituants pertinents pour l'extraction. Il est nécessaire d'envisager le contexte gauche de l'apposition pour les extraire. La difficulté

n'est pas insurmontable et les structures nettement moins ambiguës que celles issues de la coordination. Nous pouvons, par conséquent, aisément les formaliser au sein de notre automate de reconnaissance (*cf.* Figure 3.8[19]).

Attachement

> Reconstruire la structure syntaxique de la phrase, c'est reconstituer les emboîtements successifs des syntagmes, c'est-à-dire la hiérarchie des constituants : pour savoir où rattacher le syntagme, il faut pouvoir déterminer si ce syntagme est contenu ou non dans le syntagme précédent, c'est-à-dire délimiter les frontières des syntagmes à mesure que l'on tente de les hiérarchiser. Or, ce travail, nécessaire à la compréhension de la phrase, se heurte généralement à une multiplicité de solutions possibles. [Fuchs, 1996 : p. 111]

La complexité du phénomène d'attachement est très importante et concerne tant les groupes prépositionnels que nominaux et adjectivaux. Pour le seul groupe prépositionnel, dont la complexité est la plus grande, [Fuchs, 1996] distingue en tout dix contextes syntagmatiques différents répartis en quatre groupes, selon la nature du point ou syntagme de rattachement : mixte, verbale, nominale et prépositionnelle.

Comme nous l'avons annoncé, nous limiterons notre étude aux faits dégagés par l'analyse de nos corpus et, plus particulièrement, à un sous-ensemble de ceux-ci : les faits agissant directement aux frontières des syntagmes nominaux. La phrase élémentaire et, par transitivité, le processus d'extraction, reposent, en effet, sur la formalisation du groupe nominal (mis en perspective dans la phrase élémentaire et objet principal du processus d'extraction). Par conséquent, nous devons nous assurer de sa cohérence interne afin, dans le même temps, d'assurer la pertinence syntaxique des séquences extraites.

Ainsi, nous ne discuterons pas les cas d'attachement adjectival qui sont, pour la plupart, internes au syntagme nominal et ne compromettent pas le découpage en constituants (*cf.* Exemples 68 à 70). Les syntagmes adjectivaux peuvent également dépendre de certains verbes (ex. : *Il a retrouvé la voiture endommagée*, [Fuchs, 1996 : p. 116]) mais les prédicats relevés ne présentent pas ce type d'argument.

(68) **La société de biotechnologie britannique** Celltech a fait savoir qu'elle avait racheté les droits de son traitement contre l'arthrite rhumatoïde (nom de code CDP 870) auprès de Pfizer.

(69) GFI Informatique (−0,5%) : à l'occasion d'une interview sur Boursorama, Jacques

[19] Cette grammaire est une approximation de l'apposition. Au chapitre 6, nous complèterons cette illustration.

Tordjman, président de la SSII, a exclu une fusion avec **une société de taille équivalente**, comme Sopra, Unilog ou Steria, démentant ainsi les rumeurs récurrentes qui entourent l'action en Bourse.

(70) La banque d'affaires américaine a acquis **la société de gestion privée** Croosroads, dont le siège est localisé à Dallas.

Les relatives peuvent également être sources d'ambiguïté. Au sein d'un groupe nominal, il n'est pas toujours aisé, sans avoir recours à une analyse profonde de la phrase, d'identifier le nom auquel elles se rattachent. Toutefois, cette difficulté est à nouveau interne au groupe nominal et n'interfère aucunement au niveau du découpage en constituants. De telles structures amènent néanmoins une question intéressante, la résolution d'anaphore, que nous abordons plus loin dans ce chapitre. Le pronom relatif *qui* de la phrase 71, par exemple, renvoie au groupe nominal *le groupe allemand de services aux collectivités* et non pas au nom *collectivités*.

(71) La filiale de France Télécom discuterait avec le groupe allemand de services aux collectivités **qui** pourrait lui céder les 51% qu'il détient dans Connect Austria.

Dans son acception la plus simple, le groupe nominal ne pose pas non plus de réel problème d'attachement (*cf.* Figure 3.9).

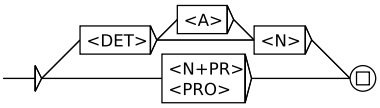


Figure 3.9 : Syntagme nominal minimal

Les ajouts gauches possibles ne présentent pas *a priori* d'ambiguïté structurelle. S'il y a ambiguïté, celle-ci est imputable à la complexité lexicale. Les problèmes d'attachement propres au *SN* proviennent des ajouts droits : adjectifs, groupes prépositionnels et relatives. Le nombre de ces ajouts n'est pas, bien entendu, récursif à l'infini, le cerveau humain ne le permet pas. Toutefois, parce qu'il ne nous est pas possible de lister *in extenso* l'ensemble des combinaisons possibles, notre reconnaisseur est récursif (*cf.* Figure 3.10, dans laquelle le syntagme prépositionnel, noté *SP*, appelle le syntagme nominal).

Le vrai problème n'est pas d'obtenir l'attachement correct mais plutôt de limiter l'extraction à cette seule solution. Comme tous les cas d'ambiguïté, l'ambiguïté d'attachement, si elle n'est pas résolue, nous permet d'éviter un choix erroné. Dans le même temps, nous sommes malheureusement confronté à une multiplication des extractions pour une seule séquence. Bien sûr, l'heuristique du plus long motif nous permet déjà de limiter la combinatoire. Mais, il nous reste souvent plus d'une analyse, dont la bonne, en sortie du processus de reconnaissance.

(72) <ESENT pred="vendre"><N0>L'assureur helvétique</N0> a vendu<N1> ses activités d'assurance-vie aux Etats-Unis</N1> à<N2> Bank One Corporation</N2><N3> pour un montant de 500 millions de dollars</N3></ESENT>

(73) <ESENT pred="racheter"><N0>le groupe</N0> a racheté<N3> pour 40,3 millions de dollars en cash</N3><N1> Ximian un fournisseur de logiciels Linux</N1></ESENT>

(74) <ESENT pred="racheter"><N0>Royal Bank Of Scotland</N0> rachète<N1> Churchill Ins</N1> au<N2> Credit Suisse</N2><N3> pour un montant supérieur à 1 milliard de livres</N3></ESENT>

(75) <ESENT pred="vendre"><N0>Vivendi Universal</N0> a vendu<N1> sa tour de bureaux à Los Angeles</N1><N3> pour 190 millions de dollars</N3></ESENT>

(76) <ESENT pred="vendre"><N0>Alstom</N0> vend<N1> ses turbines</N1> à<N2> Siemens</N2><N3> pour 1,1 milliard d'euros</N3></ESENT>

(77) <ESENT pred="céder"><N0>le groupe helvétique</N0> cède<N1> les marques</N1><N1> Dreamery, Godva et Whole Fruit</N1> à<N2> une filiale de CoolBrands International</N2></ESENT>

(78) <ESENT pred="céder"><N0>Northrop Grumman</N0> a cédé<N1> une partie de l'équipementier automobile TRW</N1> au<N2> fonds d'investissement Blackstone Group</N2></ESENT>

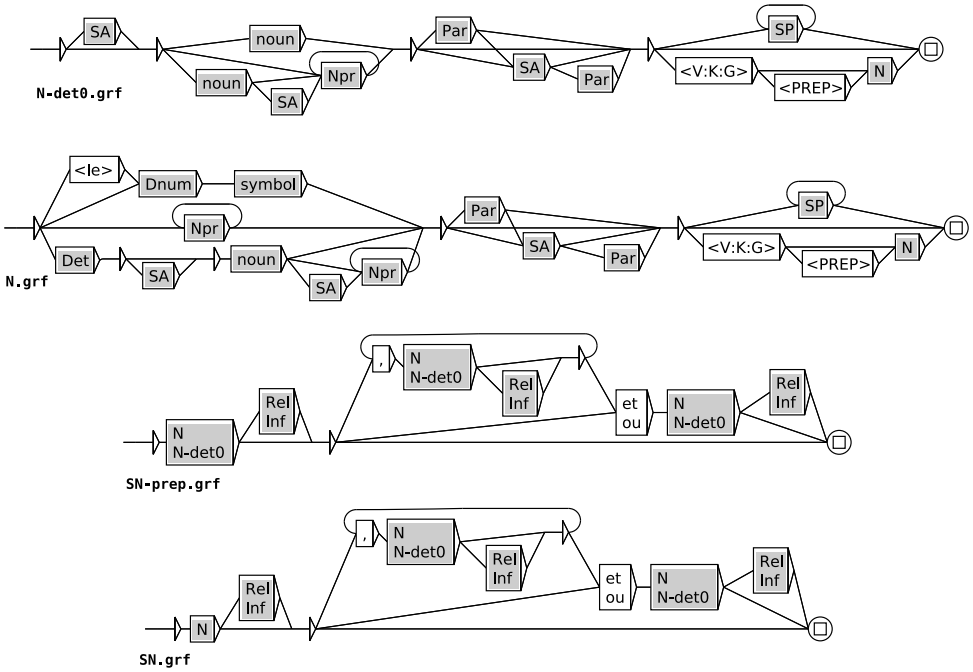


Figure 3.10 : Syntagme nominal complexe

Les exemples 72 à 78 sont tous corrects. Chacune de ces solutions était toutefois noyée dans un faisceau de résultats. Des heuristiques[20] nous permettent d'évacuer la plupart des analyses improbables mais il reste bien souvent plus d'une solution.

(79) <ESENT pred="vendre"><N0>L'assureur helvétique</N0> a vendu<N1> ses activités d'assurance-vie aux Etats-Unis à Bank One Corporation</N1><N3> pour un montant de 500 millions de dollars</N3></ESENT>

(80) <ESENT pred="vendre"><N0>L'assureur helvétique</N0> a vendu<N1> ses activités d'assurance-vie aux Etats-Unis</N1> à<N2> Bank One Corporation</N2><N3> pour un montant de 500 millions de dollars</N3></ESENT>

(81) <ESENT pred="vendre"><N0>L'assureur helvétique</N0> a vendu<N1> ses activités d'assurance-vie</N1> aux<N2> Etats-Unis à Bank One Corporation</N2><N3> pour un montant de 500 millions de dollars</N3></ESENT>

(82) <ESENT pred="vendre"><N0>L'assureur helvétique</N0> a vendu<N1> ses activités d'assurance-vie</N1> aux<N2> Etats-Unis</N2> à<N2> Bank One Corporation</N2><N3> pour un montant de 500 millions de dollars</N3></ESENT>

Il n'est pas, à proprement parler, de solution idéale ou définitive à ce problème d'attachement. Nous avons de notre côté privilégié une résolution incrémentale agissant tant au niveau lexical (*cf.* Chapitre 5) que syntaxique (*cf.* Chapitre 6) et reposant sur l'idée de regroupements successifs en segments non récursifs. Cette solution, nous le verrons, est hybride et tire parti de la statistique afin de diriger les raisonnements linguistiques.

Spécialisation syntaxique

(83) Ergo a cédé ses titres à un prix compris entre 10 et 15 euros.

(84) Ce sujet sensible, puisque la banque verte avait acquis ses titres en 2000 à un prix de revient proche de 308 euros alors qu'aujourd'hui ils ne valent qu'environ 151 euros, se heurte à divers obstacles.

(85) Novartis va acquérir 48 millions de dollars d'actions Regeneron.

[20] Nous ne croyons pas nécessaire de détailler ces heuristiques qui offrent un intérêt linguistique très limité. Nous citerons à titre d'exemple celle de *la plus grande diversité* qui, à la manière du plus long motif, vise à conserver uniquement les extractions présentant le plus grand nombre de constituants pertinents différents. Si nous considérons les exemples 79 à 82, cette heuristique nous aurait amené à éliminer la solution 79. Nous pourrions également supprimer les exemples 81 et 82 en considérant comme erronée une analyse conduisant à l'extraction de plusieurs constituants du même type (N_2, dans les exemples 79 à 82). Le lecteur intéressé pourra découvrir ces heuristiques dans le manuel technique proposé en annexe.

La table 36DT que nous avons utilisée dans cette première expérience reflète l'usage standard des prédicats qu'elle décrit. Or, nous notons, au sein de nos corpus financiers, une certaine déviance par rapport à cette norme. Nous avions déjà constaté une spécialisation lexicale. Celle-ci s'accompagne également d'une spécialisation syntaxique qui s'observe au niveau de la structure élémentaire et des contraintes qui la régissent.

3.5.3 Le plan sémantique

La pertinence des syntagmes extraits

(86) <ESENT pred="offrir"><N0>Cette solution</N0> offre à<N2> la prison</N2><N1> un grand éventail de fonctionnalités</N1></ESENT>

(87) <ESENT pred="offrir"><N0>Ces vaccins</N0> offrent<N1> une protection contre la diphtérie</N1></ESENT>

Ces exemples illustrent la problématique de l'ambiguïté sémantique que nous évoquions dans l'introduction de ce chapitre. Une mauvaise interprétation conduit à un défaut de précision. Il s'agit pourtant d'un emploi transitif dont la structure rappelle la phrase élémentaire définitoire de la table 36DT[21].

La confusion peut, selon nous, être évitée en augmentant les informations de sous-catégorisation. Ainsi, l'ajout d'un trait *+fin* (pour finance) nous permettrait de contraindre davantage les constituants essentiels (*cf.* Chapitre 4).

L'anaphore

La résolution d'anaphore est un sujet très étudié en traitement automatique des langues (*cf.* [Mitkov, 2002]). Les conférences *ACE*, et avant elles, les conférences *MUC*, ont, très rapidement, consacré une session complète à cette problématique. En effet, il est bien souvent nécessaire de dépasser le contexte d'un prédicat donné pour être à même de compléter au mieux un patron d'extraction ou de mettre en relation les informations extraites.

Si les pronoms sont les marqueurs anaphoriques les plus fréquents et les plus facilement identifiables, les quelques exemples que nous reprenons ici étendent également le terrain d'investigation aux noms propres et à certains syntagmes nominaux.

(88) ***Rhodia*** (+0,9%) : nouveau désinvestissement pour **le groupe qui** a cédé sa divi-

[21] La propriété $N_0 =: N_{hum}$ définitoire de la table 36DT ne prévoit pas le trait *non-humain* pour le sujet et ne permet donc pas d'expliquer la construction 86.

sion ignifugeants.

(89) ***Crédit Agricole*** : au deuxième trimestre, **la banque verte** a réalisé un bénéfice brut d'exploitation de 1,1 milliard d'euros en hausse de 23% sur un an pour un produit net bancaire de 3,36 milliards d'euros, en croissance de 10,3%.

(90) L'usine du Pas-de-Calais de la Française de Mécanique – la société de production de moteurs détenue à parité par Renault et ***PSA Peugeot Citroën*** – voit son rythme de production augmenter progressivement, avec la montée en puissance de son dernier né, l'"ET3". La production atteint déjà 500 moteurs par jour et le rythme de 1.300 est prévu prochainement selon son commanditaire, **PSA**. **La firme au lion** envisage en effet de compléter sa gamme de moteurs intermédiaires, à mi-chemin entre le TUI (60 chevaux) et le TU 5JP4 (110 chevaux).

Il est évident qu'une extraction limitée aux constituants anaphoriques serait incomplète et demanderait à l'utilisateur un retour au texte afin d'apprécier le contenu informationnel du patron d'extraction.

Nous nous proposons d'envisager ici, en guise d'illustration, les cas d'anaphores les plus évidents. Ces derniers sont assez simples à identifier, les quelques introducteurs pouvant être recensés exhaustivement. Il est entendu que la résolution complète de l'anaphore dépasse largement le cadre de ce travail. Notre solution est par conséquent partielle mais, parce qu'elle s'intéresse aux cas les plus fréquents, nous permettra néanmoins, une fois intégrée au processus d'extraction, de résoudre un grand nombre d'entre eux.

Les algorithmes traitant l'anaphore ne demandent pas, pour la plupart, de connaissances linguistiques importantes. Les éléments qui sous-tendent sa résolution sont assez limités.

– La compatibilité des traits syntaxiques et sémantiques. Nous touchons ici du doigt la réelle difficulté qui consiste à extraire le maximum d'informations concernant le référent (le genre et le nombre, par exemple). Comme nous le verrons, le pronom anaphorique ne suffit pas toujours. Il est donc nécessaire d'étendre notre analyse à son contexte proche.
– La distance. Bien que certains référents puissent apparaître plusieurs phrases ou paragraphes en amont du marqueur anaphorique, la toute grande majorité s'observe dans une fenêtre assez restreinte, la ou les deux phrases précédentes.

En marge de ces deux critères, nous pouvons également ajouter la fonction syntaxique. Certaines études (*cf.* [Boudreau et Kittredge, 2005]) ont montré que le sujet, par exemple, était un candidat référent particulièrement pertinent. Pour notre part, nous ne prenons pas cet élément en compte parce qu'il nécessite une analyse syntaxique, fût-elle partielle, de la phrase.

Pratiquement, notre processus de résolution opère en quatre temps :

1. identification du marqueur anaphorique ;
2. circonscription du segment textuel référent ;
3. extraction des syntagmes nominaux candidats ;
4. sélection du référent le plus pertinent au regard des critères retenus.

Figure 3.11 : Graphe principal d'identification des anaphores

Nous avons limité nos recherches aux pronoms relatifs, démonstratifs et personnels. Nous ne traiterons donc pas le cas des déterminants. Toutefois, la solution dégagée peut facilement être transposée à ce dernier cas et devrait amener des résultats pour le moins équivalents.

Les pronoms relatifs

(91) Actualité chargée également du côté de ***Rhodia***, **qui** a cédé sa branche alimentaire à Danisco et du côté d'Alstom, après son avertissement de la veille.

(92) ***Cyber Press Publishing*** **qui** a repris les magazines musicaux du groupe Ixo gagne 3% à 12,4 Euros.

(93) Dans un communiqué commun, Axa et ***Mony***, l'assureur **que** le groupe français a prévu de racheter, annoncent qu'ils ont amendé l'offre initiale en proposant de verser un dividende additionnel de 10 cents par action aux actionnaires en cas de succès du rapprochement.

Les pronoms relatifs *qui* et *que* ne nous permettent pas d'inférer le genre et le nombre du référent. Par conséquent, nous devons étendre le champ de notre analyse au reste de la proposition relative et, plus particulièrement, à son verbe principal.

Selon le temps du verbe régissant la relative, on peut ou non caractériser le référent. Ainsi, pour le pronom sujet *qui* (*cf.* Figure 3.12), le verbe de la relative nous permettra toujours de déduire le nombre du référent. Le genre pourra être observé pour les seuls verbes présentant une forme passive.

Le pronom *que* (*cf.* Figure 3.13), objet de la relative, est, quant à lui, beaucoup plus ambigu. Seuls les verbes conjugués à un temps composé au moyen de l'auxiliaire *avoir* conduiront à l'identification du genre et du nombre du référent.

Dans les faits, les pronoms relatifs ne sont pas réellement problématiques. Ils sont d'ailleurs intégrés à notre automate d'extraction (*cf.* Figure 3.6).

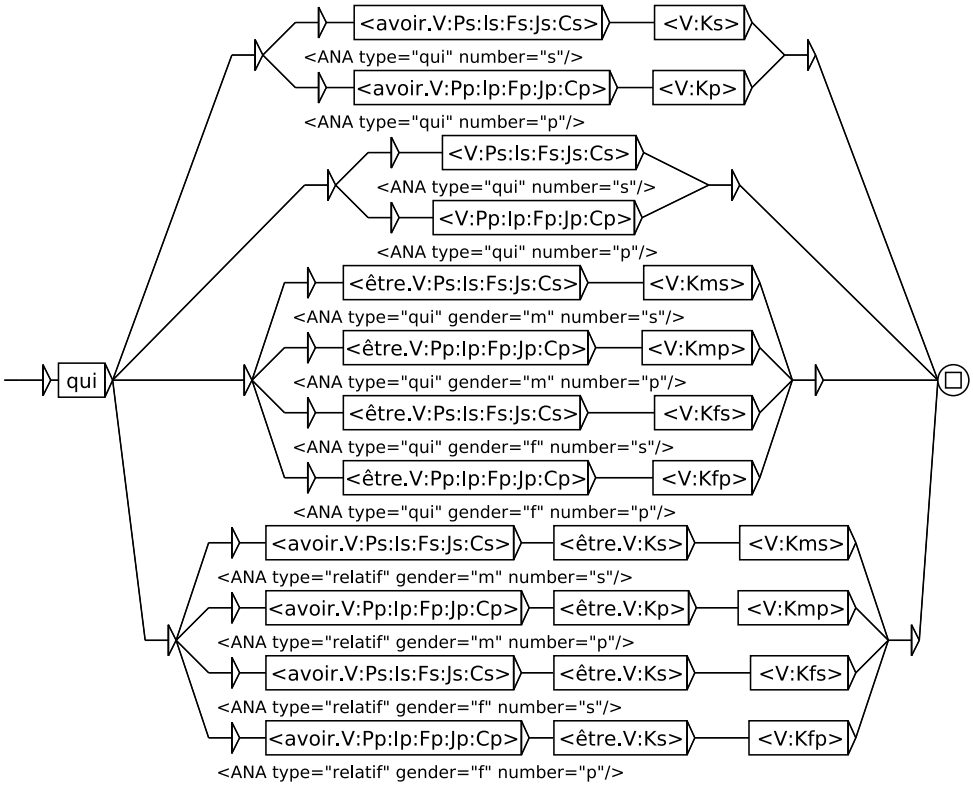


Figure 3.12 : Anaphore introduite par qui

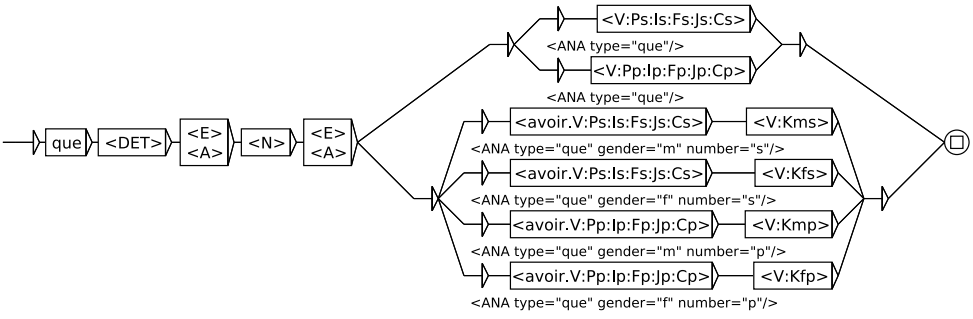


Figure 3.13 : Anaphore introduite par que

Les pronoms démonstratifs

(94) L'Autorité des Marchés Financiers a autorisé la garantie de cours visant les actions de la société Tour Eiffel présentée par Petercam SA, agissant pour le compte d'***Osiris Patrimoine***. **Cette dernière** avait acquis auprès d'un actionnaire un bloc de titres représentant 22% du capital, qui lui confère plus de la moitié du capital et des droits de vote de Tour Eiffel.

(95) La société Norcan a déposé un projet d'offre publique de rachat de ses propres actions auprès de ***l'Autorité des Marchés Financiers***. **Cette dernière** propose de racheter un maximum de 169 228 actions, soit 28,68% du capital, c'est-à-dire la totalité des actions non détenues par les actionnaires familiaux fondateurs, au prix unitaire de 26,50 euros en vue de les annuler.

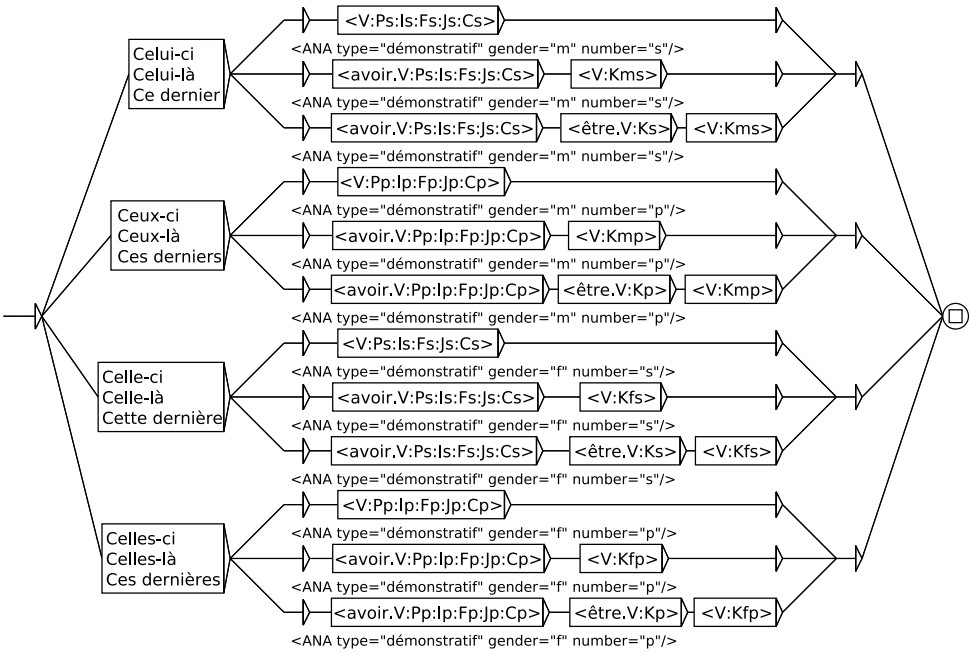


Figure 3.14 : Anaphore introduite par un pronom démonstratif

Contrairement aux pronoms relatifs, les démonstratifs présentent les indications de genre et de nombre nécessaires à l'extraction du référent.

Les pronoms personnels

(96) ***Agta Record*** a annoncé qu'**elle** avait acquis 100% du capital de son distributeur historique, la société Elemat, basée à Barcelone.

(97) Lors de la publication de ses résultats du quatrième trimestre 2003, ***Alcatel*** avait clairement annoncé qu'**il** chercherait à céder le contrôle de son activité téléphone

mais qu'**il** souhaitait conserver un intérêt minoritaire dans une nouvelle structure.

(98) Le dossier ***Alcatel*** flambe désormais de près de 8% à 13,39 euros à Paris dans de très gros volumes, toujours soutenu par la perspective d'une reprise du secteur qui porte le titre depuis maintenant près d'une semaine. **Il** cède désormais 0,5% dans un marché toujours haussier à 35,75 euros, dans le sillage du titre de l'allemand SAP qui a déçu certains analystes sur la mise à jour de ses prévisions 2004.

De même que les démonstratifs, les pronoms personnels de la troisième personne nous renseignent quant au genre et au nombre du référent. Afin de résoudre l'ambiguïté des pronoms de la première et de la deuxième personne, nous opérons de la même manière que pour le pronom sujet *qui*.

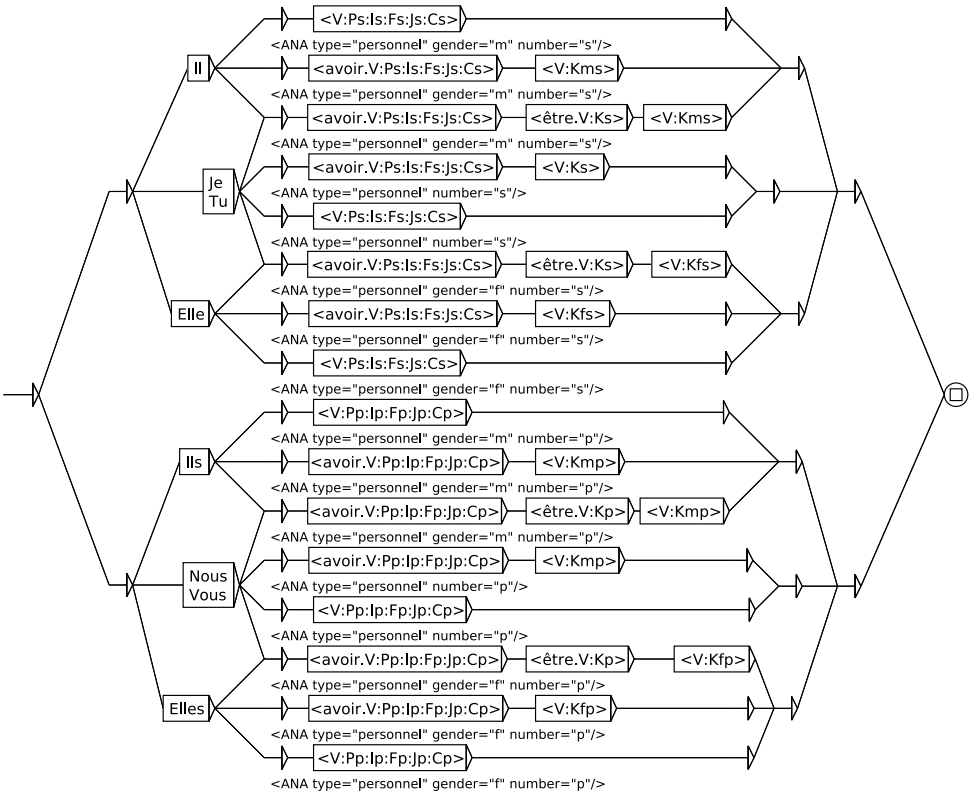


Figure 3.15 : Anaphore introduite par un pronom personnel

Extraction des candidats référents

Appliqués à un texte, les transducteurs présentés aux points précédents vont nous permettre d'insérer autant de balises <ANA> qu'il existe d'anaphores. Considérons ces quatre exemples[22] :

– La peinture, que le directeur a achetée, est accrochée dans le bureau.
– L'homme est assis dans le petit salon qui a été repeint l'an dernier.

[22]Contrairement à la plupart des exemples que nous proposons, ceux-ci sont fictifs. Le groupe nominal simplifié que nous utilisons pour cette expérience ne couvre pas la réalité de nos corpus.

– Le directeur a acheté des tableaux. Ils ont coûté très cher.
– L'homme est assis dans le petit salon. Celui-ci a été repeint l'an dernier.

Nous obtiendrons en sortie le résultat suivant :

– La peinture, que le directeur <ANA type="relatif" gender="f" number="s"/> a achetée, est accrochée dans le bureau.
– L'homme est assis dans le petit salon qui <ANA type="relatif" gender="m" number="s"/> a été repeint l'an dernier.
– Ils <ANA type="personnel" gender="m" number="p"/> ont coûté très cher.
– Celui-ci <ANA type="démonstratif" gender="m" number="s"/> a été repeint l'an dernier.

Selon le cas, nous isolons ensuite tantôt la phrase précédente, tantôt le début de la phrase courante.

– La peinture,
– L'homme est assis dans le petit salon
– Le directeur a acheté des tableaux.
– L'homme est assis dans le petit salon.

Enfin, afin d'extraire les candidats référents, nous avons développé une version du groupe nominal (cf. Figure 3.16). Cette version très simplifiée nous permet d'identifier, au sein des séquences isolées (mode *All matches* d'*Unitex*), l'ensemble des groupes nominaux potentiels. Nous associons, en cours de reconnaissance, les informations de genre et de nombre qui caractérisent chacun de ces groupes (balise XML <N>). Ces informations nous permettent ensuite d'identifier le groupe pertinent au regard des attributs de la balise <ANA>.

Les résultats que nous obtenons sont limités, à l'image de notre solution. La compatibilité en genre et en nombre du référent et du complément anaphorique nous permet d'évacuer beaucoup de faux candidats. Néanmoins, elle n'est pas toujours suffisamment contraignante quand elle n'est pas simplement absente.

Les dictionnaires associés à *Unitex* présentent un raffinement méta-linguistique qui dépasse nettement ces deux codes flexionnels. En étendant le jugement de compatibilité aux éléments de sous-catégorisation, nous pourrions contraindre plus sérieusement l'identification du référent. Malheureusement, il nous faudrait alors exprimer, au sein de nos automates, toutes les combinaisons possibles et, par conséquent, augmenter les chemins de manière exponentielle. Les structures de traits et l'unification intégrées à l'analyseur syntaxique d'*Outilex* nous semblent

plus à même de répondre à ce problème (*cf.* Section 1.3.3).

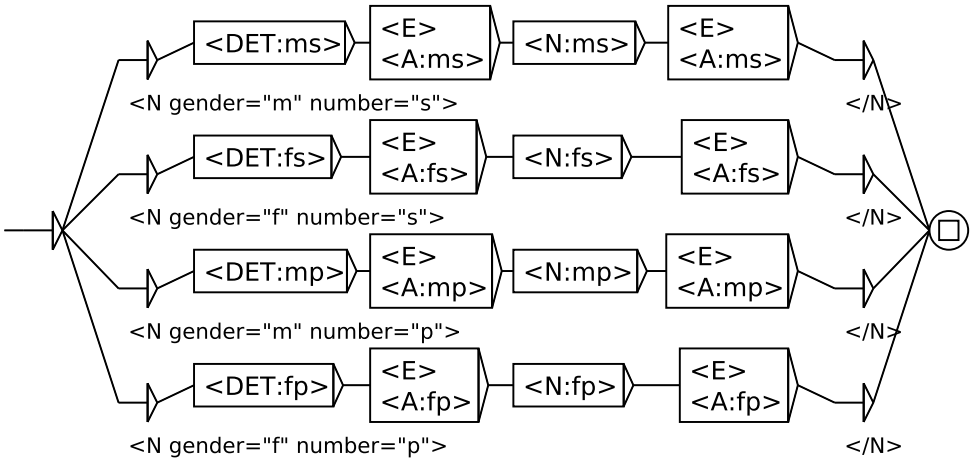


Figure 3.16 : Groupe nominal simplifié

3.6 Conclusion

Partant d'une solution naïve, ou pour le moins immédiate, nous avons pu dégager différents points circonscrivant la réelle difficulté d'un système linguistique d'extraction de même que les avantages de notre méthode.

La phrase élémentaire nous amène à envisager l'extraction sous un nouvel angle rappelant toutefois les solutions de [Grishman, 1995] et [Appelt *et al.*, 1993]. Nous avons pu constater qu'il s'agit là d'un outil linguistique puissant nous permettant tout à la fois de conférer à l'analyse syntaxique une dimension informationnelle, sinon sémantique, et de conserver les relations inter-constituants, indispensables au remplissage du patron d'extraction.

Nous avons isolé deux problèmes extérieurs au formalisme : la qualité graphique des documents et la résolution d'anaphore. Comme nous l'avions précisé en début de chapitre, ces points ont été décrits très brièvement. Nous avons proposé des solutions partielles, points de départ de développements ultérieurs.

Les erreurs restantes peuvent être réparties en trois catégories :

– la complexité lexicale et syntaxique,
– la spécialisation linguistique et
– l'attachement (prépositionnel ou adjectival).

La complexité lexicale et syntaxique recouvre deux problèmes : l'ambiguïté catégorielle et les transformations binaires (structures appositives et coordonnées). L'utilisation d'*Outilex* et du principe d'unification devrait nous permettre de résoudre, en partie, le premier problème en vé-

rifiant lors de l'analyse l'adéquation (en genre et en nombre par exemple) des différents constituants de la phrase. La seconde source d'erreur, la complexité syntaxique, n'est pas réellement problématique. Elle traduit un manque au sein de nos grammaires qui doit simplement être comblé. Nous aborderons ces deux points au chapitre 6.

La spécialisation qui caractérise un domaine d'extraction s'observe au niveau lexical (couverture des dictionnaires) mais également au niveau syntaxique (structures spécialisées). Le lexique-grammaire, parce qu'il décrit le fonctionnement standard de la langue, ne considère pas ces phénomènes. Nous envisagerons donc, au chapitre suivant, l'idée d'un lexique grammaire de spécialité au travers de la notion de sous-langage telle qu'elle est développée par [Harris, 1971].

L'attachement est la principale source d'erreur de précision. La récursivité du syntagme nominal et le manque d'information de sous-catégorisation propre au domaine d'extraction sont deux causes possibles à ce problème. Au chapitre suivant, nous étudierons ce problème plus en profondeur et proposerons, à terme, une solution d'analyse qui sera détaillée dans les chapitres 5 et 6.

Troisième partie

Une approche hybride et spécialisée

CHAPITRE 4

VERS UN LEXIQUE-GRAMMAIRE DE SPÉCIALITÉ

4.1 Introduction

Nous pouvons organiser les connaissances linguistiques nécessaires au déploiement d'une application en traitement automatique de la langue en quatre champs distincts : le lexique, la syntaxe, la sémantique et le discours. Dans le contexte d'une langue naturelle, il nous paraît difficile de tendre vers une description exhaustive. En effet, à ce jour et malgré des développements toujours plus nombreux, personne n'a pu formaliser toute la complexité et la puissance de l'outil linguistique. Par ailleurs, au-delà de la formalisation, il reste encore à gérer l'ambiguïté inhérente à la richesse lexicale et syntaxique. C'est pourquoi beaucoup d'applications de traitement automatique, et particulièrement celles liées à l'extraction d'information, sont construites autour d'une *thématique* particulière définissant un *domaine de spécialité*.

Depuis [Harris, 1971], nous savons que chaque domaine circonscrit un système linguistique cohérent, homogène et borné : le *sous-langage*. La *propriété de fermeture* d'un sous-langage a une incidence directe sur l'opération de formalisation. En effet, il ne s'agit plus ici de décrire le système dans son entier mais bien de limiter la description aux seules structures du domaine traité. Notons que cette limitation du focus linguistique entraîne également une précision des propriétés ditributionnelles et transformationnelles qui nous permet d'évacuer une partie, parfois conséquente, de l'ambiguïté lexico-syntaxique.

Pour qui travaille dans le cadre formel du lexique-grammaire, l'idée d'une spécialisation linguistique ne s'impose pas nécessairement. L'objectif théorique de [Gross, 1975] et [Boons *et al.*, 1976], pour ne citer qu'eux, est précisément l'étude systématique et exhaustive de la *langue standard*. Il est évident que nous ne remettons pas en cause cet objectif mais, dans un contexte applicatif, il nous semble intéressant sinon nécessaire de limiter la formalisation aux besoins de l'application. C'est en ce sens que nous avons abordé la notion de sous-langage[1].

[1] Rappelons que les classes d'objets envisagent déjà une certaine spécialisation des données (*cf.* Section 2.4).

Depuis la fin des années soixante, différents projets ont intégré cette notion de manière plus ou moins aboutie (*cf.* 4.3). Toutefois, et malgré la pertinence des résultats, la linguistique computationnelle ne semble pas encore unanime. [Kittredge, 1982] expliquait déjà cette situation, toujours d'actualité, en avançant l'énorme investissement que demande la description linguistique même quand il s'agit d'un sous-ensemble de la langue. Ce constat n'est pas sans poser problème dans un contexte d'extraction d'information. Un des critères d'évaluation de tout moteur d'extraction consiste précisément en son adaptabilité (ce qui peut être difficile s'il faut, à chaque changement de sujet d'extraction, développer de nouvelles ressources). Nous tenterons donc de répondre à ce critère par l'automatisation de certaines procédures et par la mise au point d'une structure modulaire afin d'accélérer de manière significative le processus descriptif.

Avant toutefois de nous lancer dans la description linguistique proprement dite, il nous semble pertinent de nous arrêter sur la définition du sous-langage qui peut prêter à confusion au regard de notions telles que le *registre*, la *langue de spécialité* et la *langue contrôlée*.

4.2 La notion de sous-langage

4.2.1 Définition

> Certains **sous-ensembles propres de phrases** d'un langage peuvent être **fermés** pour certaines (ou toutes les) opérations définies dans le langage, et constituer ainsi un sous-langage de ce langage. [Harris, 1971 : pp. 170-1]

Harris a, le premier, défini le sous-langage dans un cadre linguistique. Sa définition induit un certain nombre de propriétés particulièrement pertinentes pour le traitement automatique des langues.

Propriété de fermeture

La propriété de fermeture nous amène à considérer deux corollaires des plus intéressants. D'une part, si un sous-langage possède tout à la fois une syntaxe et un lexique clos, il est possible de développer un jeu fini de règles définissant la grammaire de ce sous-langage. D'autre part, la propriété de fermeture nous permet d'envisager la compilation d'un corpus renfermant la totalité du lexique et de la grammaire d'un sous-langage donné.

Désambiguïsation morpho-syntaxique

La spécialisation sémantique d'un sous-langage nous permet d'évacuer, pour les entrées lexicales polysémiques, certains sens non pertinents parce que ne relevant pas du domaine spécialisé

(*e.g. par*, «nombre de coups nécessaires pour un trou au golf» [Rey-Debove et Rey, 1994]).

> A parsing grammar that contains most of the constructions found in English sentences, plus reduced sentence forms, is very likely to produce multiple analysis of an input string. To constrain the number of analysis and, hopefully, arrive at the intended one, the grammar must be further restricted, and this is the primary role of sublanguage in parsing. [Sager et Nhàn, 2002 : p. 99]

Cette simplification s'observe également au niveau morpho-syntaxique : chaque suppression de sens réduit la complexité lexicale du dictionnaire. Ainsi, par exemple, *par* ne serait pas considéré comme nom en dehors du discours sportif. Par conséquent, le résultat de l'étiquetage s'en trouve simplifié, les hypothèses d'analyses étant plus restreintes.

Composante sémantique

Dans le cadre de la langue générale, les liens entre forme et sens, bien qu'indéniables, ne sont pas toujours évidents à appréhender et sont généralement assez limités, d'autant, nous l'avons vu, qu'une même forme peut avoir différentes significations.

Le contexte local de la phrase élémentaire, unité minimale de sens dans notre approche (*cf.* Chapitre 1), nous permet, déjà, d'apporter certaines précisions sémantiques au niveau des constituants essentiels. Ainsi, les définitions des classes lexico-syntaxiques du lexique-grammaire actualisent des notions telles que *concret vs abstrait*, *humain vs non humain*, *etc.* (*cf.* Tableau 4.1 [Guillet et Leclère, 1992])

TABLE	PHRASE ÉLÉMENTAIRE
36DT	N_0*hum V* N_1 *à* N_2*hum*
36SL	N_0 *V* N_1 *(Loc, avec, et)* N_2
37E	N_0 *V* N_1*source de* N_2*arg*
37M1	N_0 *V* N_1*dest de* N_2*arg* "non locatif"
37M2	N_0 *V* N_1*dest de* N_2*arg* "trace"
37M3	N_0 *V* N_1*dest de* N_2*arg* "déformation"
37M4	N_0 *V* N_1*dest de* N_2*arg* résiduel
37M5	N_0 *V* N_1*dest de (N de V-n)*$_2$
37M6	N_0 *V* N_1*dest de (V-n de N)*$_2$
38L	N_0 *V* N_1*arg de Nsource Loc Ndest*
38LD	N_0 *V* N_1*arg Loc* N_2*dest*
38LS	N_0 *V* N_1*arg de* N_2*source*
38LH	N_0 *V* N_1*hum (de, Loc)* N_2
38LR	N_0 *V* N_1*arg Loc* N_2
38L0	N_0*lieu V* N_1*arg*
38L1	N_0*arg V* N_1*lieu*

Tableau 4.1 : Structures transitives locatives

Si, maintenant, en plus du contexte local, nous considérons un contexte discursif particulier,

nous constatons que ces notions sémantiques se précisent. La phrase élémentaire nous permet alors d'envisager les relations qui lient les différents constituants au travers du prédicat.

> In a sublanguage, selectional word classes have relatively sharp boundaries, reflecting the division of real word objects into classes[2] that are sharply differentiated in the domain. Refined syntactic formulas stated in terms of these classes reflect the types of relations the object named in these classes can have to each other and thereby provide a semantic characterization of the discourse in the given domain, using grammatical methods of description. [Sager, 1986 : p. 3]

Le prototype que nous décrivons au chapitre 3 repose directement sur cette propriété. Si nous ne l'avons pas ancré dans un contexte de sous-langage, il n'en reste pas moins que nous avons associé, à chaque lien syntaxique, un lien informationnel[3] et organisé de ce fait les différents constituants en classes lexicales (ou classes d'objet).

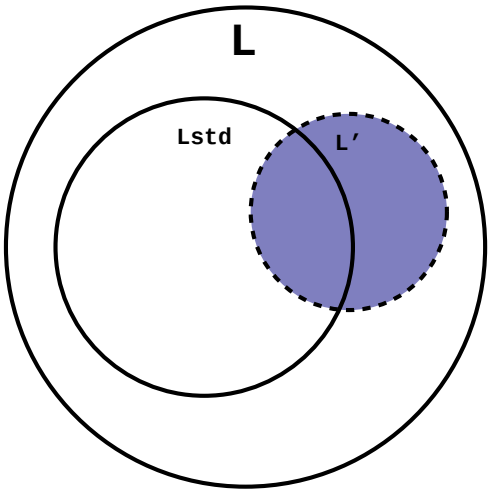


Figure 4.1 : Langue naturelle (L), générale Lstd et spécialisée (L')

Déviance grammaticale

Harris considère le sous-langage comme un sous-ensemble linguistique clos. Notons néanmoins qu'il ne s'agit pas à proprement parler d'un sous-ensemble de la langue générale.

> A specialized variety of a language is not always just a restricted selection from the

[2]Nous attirons l'attention du lecteur sur cette division des objets linguistiques en classes qui rapelle le travail de [G. Gross, 1994] que nous évoquions dans l'introduction de ce chapitre, confirmant ainsi notre intuition d'une correspondance entre les classes d'objets et le concept de sous-langage

[3]Bien souvent, en traitement automatique des langues, le terme *sémantique* effraie. Dans ce travail, quand nous parlons de sémantique pour un mot donné, nous désignons plus sa *charge informationnelle* (*i.e.* l'entité à laquelle il correspond au sein du formulaire d'extraction) que son sens proprement dit. Nous ne cherchons pas à définir le sens d'un mot mais plutôt ce à quoi il fait référence.

> full set of possible sentences of the language. Its rules do not necessarily form a proper subset of the rules for the language as a whole, contained precisely within the general rules of the grammar of the language. As well as lacking some of the features of the general language, the variety may have some of its own rules, which are not found in the general rule set. [Barnbrook et Sinclair, 2001 : p. 242]

S'il est évident que les deux concepts – langue générale et sous-langage – partagent des éléments tant syntaxiques que lexicaux, un sous-langage peut connaître des tours syntaxiques que la langue générale invaliderait et des termes spécifiques inconnus du lexique standard. Dès lors que la question de grammaticalité se pose différemment, nous parlerons plus volontiers d'intersection que d'inclusion (*cf.* Figure 4.1 [Lehrberger, 1986]).

Il convient de distinguer deux types de déviances selon que l'emploi syntaxique est considéré comme agrammatical par la langue générale (*structures syntaxiques spécialisées*) ou simplement qu'il n'existe pas dans cette dernière (*structures syntaxiques de spécialité*) :

– **Structures syntaxiques spécialisées** : l'agrammaticalité s'explique le plus souvent par la structuration textuelle propre au sous-langage étudié[4]. Un rapport médical, par exemple, contient une très large majorité de structures syntaxiques télégraphiques que la grammaire standard[5] ne permet pas de valider (*cf.* Exemples 99 à 102). Les dépêches financières, quant à elles, peuvent présenter des phrases qui, bien que proches de la langue standard, n'en actualisent pas toutes les particularités : absence d'articles, structures verbales incomplètes (participes, nominalisation, *etc.*), suppression des prépositions dans l'expression des dates, *etc.* (*cf.* Exemples 103 à 106).

(99) Fracture du col fémoral droit opérée en 08/02.

(100) Pas d'argument pour une insuffisance hypophysaire ni gonadique primaire.

(101) Insuffisance respiratoire majeure ayant entraîné le décès.

(102) Conclusion : Vésicule micro-lithiasique sans complication inflammatoire et/ou infectieuse mais symptomatique.

(103) Bouygues : frémissement à 30,50E ?

(104) Vinci : bons résultats annuels, 2004 vu en croissance.

(105) Coface : bondit en tête du SRD, résultats 2003 records.

[4]Notons que si un sous-langage est avant tout caractérisé par une thématique spécifique, il est également conditionné par son mode de transmission (presse, article scientifique, rapport,...). Tant sa thématique que son support influent sur sa grammaire. Nous ne pouvons donc, lors de l'établissement d'un sous-langage, négliger l'une ou l'autre de ces notions.

[5]La notion de grammaire ou langue standard est très fluctuante. Nous pourrions admettre la grammaticalité de certains exemples proposés ici. Il serait donc plus raisonnable, dans ce cas, de parler d'un usage moins *canonique*.

(106) Havas : inchangé, Euro RSCG retenue par FT pour Wanadoo.

– **Structures syntaxiques de spécialité** : le second type de déviance concerne les prédicats dont le fonctionnement syntaxique ne rencontre pas le fonctionnement standard. Nous pourrions noter ces divergences comme autant d'exceptions à la norme. Toutefois, leur régularité au sein de la langue de spécialité nous pousse à y voir un emploi alternatif. Il ne s'agirait donc pas d'un prédicat dont l'emploi aurait été spécialisé mais plutôt d'un nouveau prédicat, propre au domaine étudié (*cf.* Exemples 107 à 109).

(107) Alstom sort d'une longue série de baisses pour reprendre 2% à 1,45 Euro, plus forte hausse du SRD ce midi.

(108) A la bourse de Paris, l'action Wanadoo corrige de 2,1% à 7,45E.

(109) Zurich Financial Services cède 0,6% à 208,25 CHF à Zurich.

4.2.2 Débat terminologique

La métalangue a nourri de nombreux débats entre linguistes de confessions différentes. Ainsi, la terminologie linguistique varie parfois considérablement selon que l'on évolue en sociolinguistique, en linguistique computationnelle, en didactique des langues, *etc.* La notion de sous-langage n'a pas manqué également d'alimenter certaines discussions terminologiques. Si, naturellement, notre idée du sous-langage rencontrera plus volontiers celle développée dans le cadre de la linguistique computationnelle ou mathématique, nous ne refusons pas la pertinence d'autres cadres terminologiques. Plus encore, nous pensons que leur exploitation peut nous permettre de situer plus précisément la notion.

Sous-langage vs registre

[Biber et Finegan, 1994] définit le registre[6] comme suit : «A communication situation that recurs regularly in a society (in terms of participant, setting, communicative functions, and so forth) will tend over time to develop identifying markers of language structure and language use, different from the language of other communication situations» (p. 20). Par conséquent, une étude de registre opère sur trois niveaux : «description of the situational characteristic of a register, description of the linguistic characteristics, and analysis of the functional or conventional associations between the situational and linguistic features» (p. 33).

[6] L'ambiguïté terminologique de la notion de registre dépasse très largement ce que nous en présentons. Nous pourrions très facilement trouver une définition du registre qui infirme le point de vue que nous développons. Nous avons, en effet, délibérément limité notre investigation aux articles de Biber dont l'expertise dans ce domaine de recherche ne peut être mise en doute.

L'aspect situationnel de la description relève principalement d'une idée sociolinguistique de la langue. Toutefois, si nous dépassons la définition de Harris et que nous nous concentrons sur le locuteur, un sous-langage peut se comprendre comme la langue partagée par un groupe de personnes rassemblées autour d'une situation de communication spécifique. Ainsi présentée, la notion de sous-langage propre à la linguistique mathématique rejoint celle de registre telle qu'énoncée par la sociolinguistique. Le lexique et les structures syntaxiques propres à un sous-langage donné sont le reflet des habitudes linguistiques propres à une situation d'énonciation particulière (*i.e.* le rapport médical que nous avons évoqué ci-dessus).

Dès lors, nous pourrions temporairement conclure notre débat terminologique de la manière suivante : tant le registre que le sous-langage désignent un système linguistique plus ou moins clos et structuré par une grammaire distincte de la grammaire générale. Toutefois, quand un sous-langage limite son appareillage descriptif aux propriétés internes (*linguistic characteristics*), le registre ajoute une dimension externe, la situation d'énonciation (*situational characteristics*). Par ailleurs, le sous-langage intègre également une dimension thématique (le domaine de spécialité) que nous ne retrouvons pas dans la notion de registre.

Sous-langage vs langage contrôlé

De la même manière que le sous-langage, le langage contrôlé consiste en un ensemble de structures syntaxiques et d'entrées lexicales propres à un domaine particulier. Ce formalisme est également lié au traitement automatique de la langue puisqu'il a été développé afin de faciliter le processus de traduction automatique.

> A "restricted or controlled language refers to a system that limits language to a set number of core vocabulary words, and usually, a set of writing guidelines for grammar, mechanics and style, [...] A controlled language attempts to reduce ambiguities, colloquialisms, and synonyms". [Spaggiari *et al.*, 2005 : p. 107]

Cependant, bien que le langage contrôlé soit effectivement un système linguistique clos, nous ne pouvons le considérer comme équivalent au sous-langage. La définition d'un langage contrôlé repose sur une *convention*, un ensemble de structures syntaxiques et un lexique prédéfini et immuable. Il s'agit donc bien plus d'un langage construit, presque formel, que d'un langage naturel. Au contraire, un sous-langage circonscrit les habitudes langagières d'une communauté homogène. La fermeture n'est donc pas, dans ce cas, une propriété artificielle.

Sous-langage vs langage de spécialité

> In applied linguistic the designation 'language for special/specific purposes'[7] has been established as the generic term for the concept of 'communication among specialists'. [Asher, 1993 : p. 2010]

L'idée développée dans ce cadre est de mettre l'analyse linguistique au service de l'enseignement des langues. Très nettement tourné vers le monde professionnel, ce type d'enseignement permet à l'étudiant de se mettre en contact avec la langue qu'il devra manipuler et, par conséquent, de développer des compétences adaptées à ses besoins directs.

> The morphological and syntactic properties of LSP[8], by and large, correspond to those of the general language, but may differ strikingly in frequency of occurrence. Thus, particular lexical morphemes or certain economical structures may be given greater preference than others. On the textual level, however, LSP has developed traditional text forms (or genres) of its own, whose compositional patterns may be highly conventionalized (e.g., experimental reports ; case studies ; conference abstracts and abstracts in reference journals ; patent applications and lexicon articles). [Asher, 1993 : p. 2011]

D'un point de vue formel, la langue de spécialité, telle qu'elle est définie ici, présente les mêmes attributs que ceux observés pour le sous-langage. Il nous semble donc correct d'utiliser indifféremment les deux termes comme synonymes d'une même notion.

4.3 Domaines d'application

Nous avons délibérément restreint notre étude du domaine d'application des sous-langages aux seules disciplines ressortissant au traitement automatique de la langue. Cette notion est également utilisée en acquisition des langues (*cf.* [Hunston et Francis, 2000]), par exemple, mais cela sort du cadre de notre recherche.

La formalisation des sous-langages a déjà conduit à des applications pertinentes en extraction d'information, mais également en traduction automatique et en classification textuelle. Nous décrivons dans les paragraphes suivants certaines de ces applications afin de donner un aperçu représentatif des différents modes d'exploitation des facilités qu'offrent les particularités linguistiques des sous-langages.

[7] Nous considérons le terme de *langue de spécialité* comme la traduction du terme anglais *language for special/specific purposes*.

[8] *Language for special/specific purposes*.

4.3.1 Sous-langage et recherche d'information

À la fin des années soixante, le *Linguistic String Project*[9] a développé, dans le cadre d'un projet en recherche d'information, plusieurs grammaires décrivant les restrictions linguistiques propres à certains domaines médicaux et pharmaceutiques. Ces grammaires permettent d'évacuer partiellement l'ambiguïté générée au préalable par une analyse syntaxique générique. Le jeu de contraintes doit être considéré ici comme un filtre dégageant de l'ensemble des analyses possibles celles pertinentes au regard de la thématique étudiée.

L'analyse syntaxique développée dans le cadre du *Linguistic String Project* opère en quatre temps correspondant à quatre niveaux successifs de raffinement grammatical.

1. **Étiquetage morpho-syntaxique.** Un lexique associe un code grammatical à chaque entrée lexicale de la phrase en cours d'analyse. En marge de cette information grammaticale, le lexique renseigne un code sémantique qui renvoie à l'une ou l'autre classe lexicale spécialisée (patient, symptôme, *etc.*).
2. **Analyse syntaxique.** Une série de règles syntaxiques met en évidence les relations SVO, les modifieurs, *etc.* Après application de ces règles, nous obtenons une, ou, le cas échéant, plusieurs analyses possibles de la phrase d'entrée. À ce stade, nous pouvons observer une première désambiguïsation. L'application du lexique introduit inévitablement de l'ambiguïté, une même forme pouvant actualiser différentes catégories grammaticales. Toutefois, la langue ne permet pas n'importe quelle suite syntaxique. Dès lors, l'application des règles conduit nécessairement à la réduction du champ des possibles.
3. **Normalisation transformationnelle.** Un ensemble de règles transformationnelles décompose chaque phrase complexe en une suite de phrases simples, facilitant ainsi l'accès à l'information.
4. **Analyse sémantique.** Les constituants de chaque phrase sont projetés, par le biais de règles propres au domaine, sur le patron informationnel, résultat final de l'analyse (*cf.* Tableau 4.3.1 [Hirschman et Sager, 1982]).

	PATIENT	PT-STATUS		
		V-PT	NEG	SIGN-SYMPTOM
1.	patient	have	no	cold
	patient	have	no	fever
2.	patient	be		sick

Tableau 4.2 : Patron informationnel médical simplifié

Cette approche a eu un succès très important. Les développements, initiés en 1965 pour l'an-

[9] http://www.cs.nyu.edu/cs/projects/lsp/

glais, ont également été étendus au français, à l'allemand et au néerlandais. La dernière publication dont nous avons connaissance date de 1998 (*cf.* [Spyns *et al.*, 1998]) ce qui nous amène à une vitalité effective de 33 ans minimum. Notons qu'actuellement, le projet est toujours existant et recensé par l'Université de New York même s'il semble moins dynamique.

4.3.2 Sous-langage et extraction d'information

Un problème fondamental en extraction d'information est l'adaptabilité des systèmes (*cf.* Section 2.3). Un moteur d'extraction doit pouvoir être «porté» d'un domaine à un autre sans nécessiter de trop longs développements. Or, tout le monde s'accorde à dire qu'une réponse linguistique au traitement automatique de la langue requiert un temps de formalisation pour le moins conséquent. Il n'existe pas à ce jour de méthode d'acquisition ou d'apprentissage entièrement automatique de connaissances linguistiques exactes[10]. Néanmoins, nous devons constater que l'idée de sous-langage offre certaines perspectives qui nous permettent d'imaginer une automatisation, sinon complète, au moins partielle de la procédure de formalisation. La propriété de fermeture de même que l'homogénéité sémantique de documents relevant d'un sous-langage particulier laissent, en effet, supposer une certaine efficacité des méthodes d'apprentissage. [Riloff et Jones, 1999] et [Grishman, 2001] ont exploré cette hypothèse et montrent d'ailleurs des résultats encourageants.

> We explore the idea of learning both a dictionary of extraction patterns and domain specific semantic lexicon simultaneously. Furthermore, our technique requires special training resources. The input to our algorithm is a set of unannotated training texts and a handful of "seed" words for the semantic category of interest. [Riloff et Jones, 1999 : p. 1]

À aucun moment, [Riloff et Jones, 1999] ne fait allusion à la notion de sous-langage. Néanmoins, tant la méthode (le distributionnalisme) que le contexte (le discours spécialisé) évoquent le travail de Harris et particulièrement ses développements en traitement de l'information dont le *Linguistic String Project*, évoqué plus avant, est une implémentation fidèle. Le travail présenté dans cette étude repose sur deux corpus de spécialité : un ensemble de pages Internet de présentation d'entreprises et une compilation de dépêches décrivant des attaques terroristes. La procédure développée opère en trois temps. Dans un premier temps, sur base d'un petit lexique de termes propres au domaine, une procédure extrait la totalité des syntagmes nominaux les contenant. Dans le même temps, cette procédure applique une série d'heuristiques afin d'abstraire des motifs linguistiques généralisant les syntagmes extraits (*cf.* Section 2.3.1). Ensuite, ces motifs sont proposés au corpus d'entraînement et les plus productifs sont retenus comme motifs

[10] Nous pouvons cependant envisager une intervention humaine ou même tolérer une certaine approximation. C'est d'ailleurs en ce sens que nous avons travaillé au chapitre 5.

d'extraction (*cf.* Tableau 4.3.2). Finalement, les données extraites sont utilisées pour alimenter le lexique sémantique et servent à identifier de nouveaux syntagmes.

FINANCE (sociétés)	TERRORISME (armes)
owned by <x>	<x> exploded
both as <x>	threw <x>
<x> employed	bringing <x>
<x> is distributor	seized <x>
<x> positioning	quantity of <x>
marks of <x>	surrender <x>
motivated <x>	search for <x>
<x> trust company	rocket <x>
sold to <x>	<x> parked
devoted to <x>	hurled <x>
<x> consolidated stmts.	clips for <x>
<x> thrive	defused <x>
message to <x>	million in <x>
<x> is obligations	confiscated <x>
<x> request information	<x> was hurled
<x> is foundation	placed <x>
<x> has positions	rounds for <x>
incorporated as <x>	consisted of <x>
offices of <x>	firing <x>
<x> required to meet	explosion of <x>

Tableau 4.3 : Top 20 des motifs d'extraction extraits par [Riloff et Jones, 1999]

La méthode présentée ici est entièrement automatique et repose sur l'idée d'un déclenchement mutuel inspiré du distributionnalisme harrissien.

> In the case of sublanguage, through syntactic regularization and decomposition, one could reduce complex sentences to simple canonical syntactic structures (e.g., simple active declarative clauses ; what Harris terms *kernel sentences*). These structures would then exhibit repetitive patterns of word choice, which could be captured in terms of sublanguage sentence structures using these classes. [Grishman, 2001]

Le parallèle entre les études de [Riloff et Jones, 1999] et [Grishman, 2001] est indéniable. Tous deux construisent des classes de mots au départ de l'étude distributionnelle d'un corpus de spécialité. Cependant, [Grishman, 2001] ne limite pas les motifs d'extraction à un dictionnaire de syntagmes nominaux ou verbaux mais étend son formalisme à la phrase élémentaire[11]. Dès lors, le processus d'apprentissage se complexifie quelque peu. Grishman note, à ce propos,

[11] Au contraire de [Riloff et Jones, 1999], [Grishman, 2001] ne donne pas d'exemples et nous ne pouvons donc pas en rapporter ici. Toutefois, l'article laisse supposer un travail proche du nôtre. Par conséquent, nous croyons pouvoir présenter les motifs d'extraction que nous détaillons dans ce travail comme autant d'illustrations potentielles du travail de Grishman.

différents éléments perturbateurs : d'une part l'analyse automatique du corpus d'apprentissage nécessite, au préalable, l'annotation des entités nommées et, d'autre part, afin d'obtenir une phrase élémentaire à partir d'une phrase brute, nous devons nécessairement disposer d'un analyseur syntaxique robuste.

4.3.3 Sous-langage et traduction automatique

Comme le souligne [Lehrberger, 1982], un système de traduction automatique ne peut pas se résumer à un ensemble de règles de transfert (afin d'obtenir une traduction adaptée des entrées lexicales) et à quelques règles syntaxiques (réorganisant correctement la suite d'entrées ainsi obtenue). Les expériences menées dans le cadre du projet TAUM[12] ont démontré qu'une description linguistique fine des langues sources et cibles demeurait nécessaire. Dès lors, se pose une nouvelle fois la question de la formalisation de la langue.

> The authors of maintenance manuals, cook books, articles in scientific journals, etc. are generally guided by norms for writing in their particular fields. [...] These norms do not themselves constitute a grammar - that can only be determined by examining texts. But they do indicate certain regularities not present throughout the whole language, thus simplifying the task of writing formal grammars for texts in specialized fields. [Lehrberger, 1982 : p. 99]

Ces conventions d'écriture, associées aux propriétés inhérentes à tout sous-langage (lexique limité, polysémie restreinte et contraintes syntaxiques fortes), permettent d'envisager plus sereinement l'écriture de grammaires formelles. Cette idée a largement été exploitée par l'Université de Montréal dans le cadre d'un projet visant la traduction anglais/français de bulletins météorologiques. Le système TAUM analyse, dans un premier temps, les phrases en langue source et leur associe une suite de structures normalisées, des arbres dont les nœuds sont composés des entrées lexicales des phrases analysées et de leurs propriétés syntaxiques et sémantiques (spécialisées). Un jeu de règles assure ensuite le transfert de ces entrées de la langue source à la langue cible. Enfin, un processus de génération organise les structures normalisées en phrases de la langue cible.

4.4 Analyse linguistique d'un sous-langage

Dans les différents point présentés ci-après, nous évoquerons les étapes essentielles à la description linguistique d'un sous-langage ainsi que les méthodes que nous avons mises au point

[12] Le projet de traduction automatique de l'Université de Montréal (TAUM) a envisagé la traduction de langues très spécialisées comme les bulletins météorologiques (TAUM-MÉTÉO) ou les manuels d'entretien d'avions (TAUM-AVIATION) (*cf.* [Isabelle, 1987]).

Texte en langue source
Un anticyclone centré sur nos régions détermine un temps calme.
Structure normalisée
Un anticyclone centrer sur notre région déterminer un temps calme.
Transfert en langue cible
An anticyclone center on our region bring fair weather.
Génération de l'énoncé
An anticyclone centered on our regions brings fair weather.

*Tableau 4.4 : Illustration du processus de traduction (*TAUM*)*

afin d'automatiser, autant que possible, notre procédure[13]. Ces étapes sont autant d'éléments que nous couvrirons ensuite dans notre brève analyse du sous-langage de la finance et plus particulièrement celui de la vente et de l'achat d'entreprises tel qu'on peut l'observer dans la presse spécialisée.

> D'un point de vue axiomatique, une grammaire construite pour un langage (un ensemble de phrases) consiste en un ensemble de classes (et de sous-classes) de mots et de morphèmes, en un ensemble de séquences bien formées de ces classes (structures de phrases élémentaires) et en un ensemble de règles transformationnelles qui dérivent une structure de phrase d'une autre. [Harris, 1971 : p. 171]

Partant du lexique-grammaire, notre approche de la grammaire s'inscrit naturellement dans le cadre théorique et formel défini par Harris, c'est-à-dire celui du distributionnalisme et du transformationnalisme. En ce sens, nous serons amené à considérer deux composantes principales. La première, *lexicale*, consiste en un ensemble de mots (simples et composés) circonscrivant le lexique de la langue tandis que la seconde, *syntaxique*, se compose de l'ensemble des structures de phrases simples, de leurs propriétés distributionnelles et des opérations de transformation auxquelles elles se soumettent.

Notons que ces deux composantes constituent la grammaire de la langue générale ou standard, au sens harrissien bien entendu. Le contexte de notre dissertation concerne, quant à lui, l'étude d'un sous-langage qui nous permet d'ajouter au formalisme général une donnée supplémentaire : la caractérisation informationnelle des constituants essentiels.

4.4.1 Corpus de spécialité

> The beginning of any corpus study is the creation of the corpus itself. The decisions that are taken about what is to be in the corpus, and how the selection is to be

[13] Tout comme l'ensemble des applications suggérées dans notre travail, cette procédure a donné lieu à la création d'un programme informatique que nous décrivons en annexe.

> organized, control almost everything that happens subsequently. The results are only as good as the corpus. [Sinclair, 1997 : p. 13]

Définition d'un corpus de dépêches financières

Nous ne discuterons plus l'importance du recours au corpus dans une étude linguistique. Nous avons suffisamment débattu la question dans le chapitre 1. Il ressort de notre étude qu'il n'est pas réaliste d'envisager la formalisation d'une langue, générale ou spécialisée, au départ de l'intuition uniquement. Cette dernière doit nécessairement se combiner à l'exploration des faits linguistiques en contexte d'énonciation réel. Cette remarque, triviale depuis l'avènement de la linguistique de corpus, soulève pourtant une question essentielle : la *représentativité* des données. En effet, si nos résultats, et par conséquent la qualité de notre formalisation, reposent sur l'étude, exclusive (études *corpus-driven*) ou non (études *corpus-based*), d'occurrences, nous devons nous assurer que nos sources nous mettent en contact de TOUTE la complexité du langage étudié.

[Sinclair, 1997] et [Biber, 1993a] ont largement traité le problème de la compilation de corpus et nous ont permis d'identifier quatre points fondamentaux.

1. **Identification des besoins.** Selon que l'on cherche à décrire la langue standard ou à répondre aux besoins spécifiques d'une application particulière, les besoins sont évidemment différents même si la procédure générale de compilation reste identique. Ainsi, un corpus de référence, prétendant couvrir la langue dans son ensemble, doit refléter tout à la fois la langue écrite et orale, littéraire et scientifique, *etc.* De plus, comme on ne peut pas enfermer toute la langue dans un corpus, se pose également la question de l'échantillonnage[14]. De notre côté, nous visons l'étude distributionnelle de dépêches financières. Par conséquent, nous limitons la pertinence de nos corpus à un registre (journalistique) et un domaine (financier) particuliers.
2. **Sélection des sources.** L'objectif inhérent à notre démarche et, plus généralement, aux démarches en linguistique descriptive, est la généralisation des observations à la langue ou au sous-langage dans son ensemble. Il s'agit donc de refléter au mieux la population cible au sein de nos données en assurant la multiplicité des sources.
3. **Taille du corpus.** Ce point est à rapprocher du précédent en ce qu'il participe aussi de la notion de «validité externe» (principe d'adéquation du modèle à la réalité). Ce sont ces deux points qui définissent la représentativité.

> Within social science, these considerations are associated with the two main

[14] Notons que, dans le cadre d'un sous-langage – système linguistique clos – cette contrainte peut théoriquement être éliminée (*cf.* 4.2.1, *Propriété de fermeture*).

> kinds of error that can threaten 'external validity' (the extent to which it is possible to generalize from a sample to a larger target population) : 'random error' and 'bias error'. Random error occurs when a sample is not large enough to accurately estimate the true population ; bias error occurs when the selection of a sample is systematically different from the target population it is intended to represent. Both kinds of error must be minimized to achieve a representative corpus. [Biber, 1993b : p. 219–220]

Tout le monde ne s'accorde pas à donner à la taille du corpus une importance capitale (*cf.* [Biber, 1993a]). S'il est entendu que la définition de la population cible est un, sinon *le*, point central dans la compilation d'un corpus, la pertinence de certaines applications, tel que l'apprentissage automatique de terminologie (*cf.* Chapitre 5), repose directement sur la quantité de données traitées. Sans entrer dans ce débat, il nous semble préférable d'asseoir notre description sur une masse de données aussi importante que possible (*cf.* [Sinclair, 1997 : p. 18]). Au-delà de la taille générale du corpus, se pose également la question de la taille de ses différentes articulations. Un corpus visant la représentation globale d'une langue doit non seulement assurer la représentativité de l'ensemble mais également celle des différentes parties (registres) qui le composent.

4. **Maintenance des données.** La langue évolue chaque jour, particulièrement dans certains domaines scientifiques comme l'informatique, par exemple. Nous ne pouvons donc borner un corpus si nous envisageons la description du langage actuel. Or, les développements technologiques (capacité de stockage, accès à l'Internet) rendent possible la création de corpus dynamiques reflétant ces changements potentiels en évoluant parallèlement à la langue (*cf.* [Renouf, 1993; Fairon, 1999, 2006]).

Sur base de ces quatre points, nous avons compilé un corpus de dépêches financières dont nous résumons les propriétés au sein du tableau 4.4.1.

BESOINS	*registre* : dépêches journalistiques *domaine* : sous-langage de la finance *langue* : français
SOURCES	http://boursier.com [B] http://cerclefinance.com [CF] http://firstinvest.com [FI]
TAILLE	12 493 333 mots - 51 208 dépêches [B : 1 076 851 mots - 6 242 dépêches] [CF : 3 485 163 mots - 14 966 dépêches] [FI : 7 931 319 mots - 30 000 dépêches]
PÉRIODE	B : 01/12/03 → 13/04/04 CF : 16/02/04 → 26/10/04 FI : 22/09/99 → 13/04/04

Tableau 4.5 : Caractéristiques du corpus de dépêches financières

Les trois sites – boursier.com, cerclefinance.com et firstinvest.com – proposent une gamme assez complète d'informations, analyses, statistiques et conseils boursiers. Tous trois développent également une formule d'abonnement qui donne accès à différents services comme la gestion de portefeuille, l'avis des *brokers*, *etc.* Ces trois portails disposent d'un ensemble de rédacteurs, spécialistes du domaine, qui résument l'information au fur et à mesure qu'elle se crée, depuis l'ouverture jusqu'à la clôture des marchés. La quantité d'information diffusée par jour fluctue bien sûr en fonction de l'actualité. Toutefois, les chiffres que nous proposons dans le tableau 4.4.1 montrent une activité conséquente, d'autant que nous avons limité notre téléchargement à la rubrique *Actualités* de ces trois sites. Ce constat ne suffit pas à faire de ces sources des références. Cependant, leur vitalité (approximativement dix années d'existence pour chacun d'eux) laisse supposer une certaine stabilité que nous avons interprétée comme un gage de qualité.

Chaque dépêche est balisée selon un format XML (*cf.* Tableau 4.6). La balise <CORPUS> est l'élément racine de nos corpus financiers. Chaque corpus contient un certain nombre de dépêches balisées <NEWS>. Cette étiquette possède un attribut – nbr – qui renseigne l'identifiant du texte au sein du site dont il est extrait. Il n'est donc pas nécessairement unique dans notre base de données et nous sert seulement à pouvoir localiser la source originale.

Une balise <NEWS> contient les informations relatives aux titre, url, date, catégorie et le texte proprement dit. Notons que la balise <CATEGORY> est dépendante de chaque source, nous n'avons pas défini de catégories canoniques. Par ailleurs, firstinvest.com est le seul à offrir cette information.

<CORPUS>
<NEWS nbr="58410">
<URL href="http ://firstinvest.com/actualite/Default.asp"/>
<TITLE>Arcelor - Mittal : réouverture de l'offre le 27 juillet ?**</TITLE>**
<DATE>17/07/2006 10 :14**</DATE>**
<CATEGORY>ACTU FRANCE**</CATEGORY>**
<TXT>
L'offre lancée par Mittal sur Arcelor (LOR) (LORB) s'est terminée jeudi soir à minuit. Le résultat n'en sera connu que le 26 juillet après centralisation de l'ensemble des titres, le règlement et livraison des titres intervenant le 1er août.
Selon des chiffres provisoires, l'Indo-néerlandais détenait 17,63% des actions avec droits de vote mercredi soir dernier. Si Mittal obtenait plus du tiers du capital d'Arcelor (LOR) le 26 juillet prochain, l'offre pourrait redémarrer dès le lendemain. En effet, le droit luxembourgeois oblige le prolongement de l'offre de 15 jours lorsque l'acquéreur obtient plus de 33% du capital.
Cependant, la période de prolongement pourrait être plus longue que la période légale, celle-ci intervenant en pleine période estivale.
</TXT>
</NEWS>
</CORPUS>

Tableau 4.6 : Exemple d'organisation d'une dépêche

Acquisition des données

Nous avons souligné l'importance de la mise à jour des données. Pour satisfaire à cette propriété, nous avons implémenté une méthode d'acquisition automatique triviale mais puissante et très facilement adaptable à de nouvelles sources.

Le travail est grandement facilité par le sujet traité, très présent sur l'Internet. De nombreux sites sont dédiés à la finance et proposent un large éventail de rubriques spécialisées. Ces sites présentent l'information de manière structurée tant au niveau de la présentation graphique qu'au niveau du code HTML associé à chaque page. Cette structuration du code est inhérente à l'architecture de ce type de portails. Ils reposent, pour la très grande majorité, sur une base de données et un squelette de code unique. Dans un premier temps, la base de données est interrogée, au départ d'une requête de l'utilisateur ; cette requête peut être intentionnelle (la saisie d'un mot-clef dans l'interface de recherche) ou non (l'activation d'un lien menant, par exemple, vers une rubrique ou une dépêche particulière). Le résultat de la requête est ensuite organisé automatiquement au sein du squelette de code. Par conséquent, le code HTML est immuable et offre de bons repères permettant d'identifier efficacement le contenu textuel (*cf.* Figure 4.2).

Figure 4.2 : Méthode d'acquisition de ressources textuelles

Nous constatons que l'hypothèse distributionnelle qui confère au contexte une propriété de désambiguïsation, s'observe également dans notre processus d'acquisition. En effet, pour identifier la zone de dépêche, de même que pour cibler les informations qui lui sont propres (date, titre, *etc.*), nous définissons un jeu de frontières, initiale et finale, sous la forme d'un bloc de

code HTML. Ces frontières forment donc le contexte d'apparition du segment textuel que nous voulons extraire[15].

Notre processus prend, en entrée, la carte d'identité d'un site (*cf.* Tableau 4.7). Cette carte d'identité est composée des règles de construction des requêtes permettant d'accéder au contenu que nous cherchons à extraire et de l'ensemble des frontières que nous avons définies.

```
self.name = "firstinvest"
self.count = {"begin" : 34, "end" : 41145}
self.url = "http ://www.firstinvest.com/actualite/Default.asp"
self.id = "numero"
self.get1 = None
self.article = {"begin" : '<!- Texte de l'article ->', "end" : '<br></p>'}
self.title = {"begin" : '<!- titre de l'article ->', "end" : '</h2>'}
self.date = {"begin" : '<!- Date de l'article ->', "end" : '</p>'}
self.category = {"begin" : '<p class="histoLink">', "end" : '</p>'}
```

Tableau 4.7 : Carte d'identité de `firstinvest.com`

[Fairon, 2006] décrit une méthode plus simple encore, reposant sur l'acquisition de fils RSS, structure XML organisant l'information en champs distincts directement accessibles. Nous croyons que la combinaison des deux méthodes nous permettrait de couvrir aisément et largement un domaine particulier afin, par exemple, de proposer un moteur de recherche spécialisé.

4.4.2 Lexique-grammaire, sous-langage et analyse syntaxique hybride

Au chapitre 3, nous avons identifié différents problèmes à l'origine des erreurs mises en avant par l'évaluation de notre prototype. Ces erreurs relèvent tout à la fois des plans lexical, syntaxique et informationnel[16].

– Plan lexical : couverture lexicale et ambiguïté catégorielle.
– Plan syntaxique : complexité structurale, attachement (prépositionnel et adjectival) et spécialisation syntaxique.
– Plan informationnel : pertinence *sémantique* des syntagmes extraits et anaphore.

En conclusion, nous avancions l'hypothèse d'une nécessaire spécialisation des ressources linguistiques en fonction du domaine d'extraction. Dans ce chapitre, nous avons identifié les différents corollaires que suppose la notion de sous-langage. Nous nous proposons maintenant de les mettre en perspective dans une nouvelle procédure d'analyse spécialisée et de voir dans quelle mesure ils nous permettent de répondre aux problèmes recensés ci-dessus.

[15] L'architecture d'un site peut bien évidemment changer au cours du temps. Il est donc nécessaire d'assurer une certaine maintenance des définitions de frontières.

[16] Au problème posé par l'anaphore, nous avons proposé une solution qui, si elle reste très partielle, donne une bonne assise pour des développements ultérieurs. Il s'agit d'un problème extrêmement complexe que nous ne pourrions résoudre sans nous éloigner de notre objectif initial. Nous ne reviendrons donc plus sur ce point.

Nous sommes parti d'un constat : l'analyse syntaxique automatique s'apparente au processus distributionnel développé dans [Harris, 1974]. En ce sens, il se compose d'une phase de **segmentation** et d'une phase d'**attachement**[17].

La fonction de la **segmentation** est de transformer une séquence textuelle en une suite de segments uniques ayant un environnement unique également[18]. Ces segments sont rassemblés en classes d'équivalences (*similarity groupings*, [Harris, 1981b : p. 16]) qui correspondent aux différentes catégories grammaticales. Ce sont ces catégories qui servent ensuite d'entrées à la seconde phase. Nous ne pourrions, en effet, généraliser l'observation distributionnelle à la langue (ou au sous-langage) en travaillant au départ d'éléments discrets.

> The first distributional fact is that it is possible to divide (to segment) any flow of speech into parts, in such a way that we can find some regularities in the occurrence of one part relative to others in the flow of speech. These parts are the discrete elements which have a certain distribution (set of relative locations) in the flow of speech ; and each bit of speech is a particular combination of elements. The first operation is purely segmenting, arbitrary if need be. [Harris, 1981b : p. 15]
> Another essential distributional fact is that some elements are similar to others in terms of certain tests ; or similar in the sense that if we group these similar elements into sets ('similarity groupings'), the distribution of all members of a set (in respect to other sets) will be the same as far as we can discover. [Harris, 1981b : p. 16]

L'**attachement** vise quant à lui le rassemblement des classes lexicales en constituants de manière à obtenir une suite de séquences homogènes et indépendantes. Ainsi, il ne reste plus d'éléments extérieurs à tout rattachement sinon les éléments du niveau supérieur.

> To obtain a least set of elements sufficient for description, we join any elements which are completely dependent : if A is a set of similar elements (a similarity grouping) and so is B, and (in a particular type of environment) only AB occurs (not necessarily contiguously), never A or B alone, then we set up AB as a single element (a single set of similar elements). [Harris, 1981b : p. 16]

Le formalisme du lexique-grammaire, qui nous préoccupe plus particulièrement, repose lui-même sur une analyse distributionnelle. L'analyse automatique qui en découle consiste donc en la «simple» reconnaissance des classes et propriétés distributionnelles illustrées par l'analyse

[17] Que l'analyse s'applique à un langage naturel ou formel (*cf.* [Aho *et al.*, 2000]), nous retrouvons ces deux articulations qui correspondent aux niveaux lexical et syntaxique. Si nous pouvons constater des différences, parfois fondamentales, entre deux solutions d'analyse, il s'agit pour l'essentiel de différences d'appréciation des deux composantes.

[18] Harris désigne ces éléments comme les *morphèmes* de la langue ; aujourd'hui, nous préfèrerons l'appellation d'*entrées lexicales.*

linguistique[19]. Nous retrouvons, comme attendu, les deux étapes de l'analyse :

– segmentation et regroupement en classes d'équivalences par application des dictionnaires de formes simples et composées,
– attachement en constituants par reconnaissance des structures élémentaires ou des combinaisons de structures élémentaires.

L'avantage de cette procédure est double. D'une part, il s'agit d'un processus entièrement distributionnel et, par conséquent, particulièrement pertinent pour l'analyse automatique. Il repose, en effet, uniquement sur les occurrences d'entrées lexicales et de constituants et ne requiert aucune information supplémentaire ([Harris, 1974]). D'autre part, elle conduit à la désambiguïsation du prédicat. La caractérisation du contexte de chaque prédicat nous permet d'évacuer l'ambiguïté lexicale inhérente à la toute grande majorité d'entre-eux. De cette manière, il nous est possible d'associer à l'analyse syntaxique une dimension, sinon sémantique, au moins informationnelle.

Parallèlement à ces deux avantages, nous avons dû constater que l'attachement prépositionnel ou adjectival reste très délicat en ce qu'il conserve souvent un forte ambiguïté et génère donc du bruit au niveau des résultats (*cf.* Chapitre 3). Nous avons isolé deux causes à ce problème : la récursivité du groupe nominal et la nécessaire généralisation des propriétés distributionnelles.

1. **Récursivité du** *GN***.** L'attachement proposé par Harris repose sur la notion de dépendance. Cette dépendance se caractérise, au niveau syntaxique, selon le «domaine»[20] auquel l'attachement ressortit (l'attachement au nom relève du domaine syntagmatique tandis que l'attachement au verbe relève du domaine phrastique, *cf.* [Harris, 1981b]). Dès lors, la récursion que nos modèles formels intègrent au groupe nominal (*cf.* 3.10) est en partie responsable du phénomène d'ambiguïté, nos reconnaisseurs ne pouvant pas décider quelle solution choisir.
2. **Propriétés distributionnnelles.** Le lexique-grammaire offre une description des propriétés linguistiques de la langue générale et se limite aux éléments transversaux à tout énoncé indépendamment du contexte discursif. Par conséquent, la sélection lexicale qui repose sur ces propriétés pour diriger l'attachement peut ne pas être suffisante.

À ces deux problèmes, nous avons envisagé deux solutions : l'*évacuation de la récursivité* de nos modèles formels et la *spécialisation lexico-syntaxique* des faits linguistiques. Les paragraphes suivants décrivent brièvement la solution que nous avons mise au point. Les détails

[19]Ce point est, d'ailleurs, l'hypothèse qui sous-tend le travail de [Paumier, 2003] : *"De la reconnaissance de formes linguistiques à l'analyse syntaxique*.

[20]Le domaine dont il est question ici n'est pas à mettre en parallèle avec le *domaine de spécialité*. Il s'agit du *domaine formel* ou *syntaxique*.

seront abordés dans le chapitre 5 pour le côté lexical et dans le chapitre 6 pour le côté syntaxique.

[Abney, 1991] suggère une solution d'analyse incrémentale reposant sur la notion de *chunks* ou *tronçons syntaxiques*.

> Roughly speaking, 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. [Abney, 1996 : p. 1]

Il s'agit d'un constituant syntaxique inférieur au syntagme, une couche d'abstraction intermédiaire entre le plan lexical et le plan syntagmatique. Pratiquement, cette méthode d'analyse décompose le processus d'attachement syntagmatique en deux phases successives : *segmentation en chunks* et *attachement en syntagmes*.

Le problème de l'attachement n'est donc pas évacué mais envisagé à un niveau supérieur, moins ambigu. Chaque chunk s'organise en effet autour d'une tête lexicale dont il hérite les propriétés lexicales et la catégorie grammaticale. Dès lors, les regroupements syntagmatiques sont calculés au départ d'un ensemble plus restreint et moins complexe.

Notons que le chunk est une unité purement syntaxique indépendante du contexte d'énonciation. Cette particularité a le double avantage de permettre la formalisation à partir d'une simple grammaire context-free et la réutilisation des règles qu'elle définit quel que soit le sous-langage étudié. Par conséquent, elle nous permet de tendre vers le principe de *modularité* que nous évoquions au chapitre 2, en séparant structures générales (ou standards) et spécialisées.

L'attachement relève quant à lui, nous l'avons vu, d'une procédure mixte mêlant *patrons syntaxiques* et *sélection lexicale*. C'est à ce niveau de l'analyse qu'interviennent les contraintes propres au sous-langage.

Patrons syntaxiques

Dans un lexique-grammaire, chaque prédicat actualise une structure élémentaire particulière définissant un patron syntaxique au sein duquel les constituants essentiels sont représentés. Les compléments prépositionnels se présentent de la même manière que tout autre constituant sinon que leur préposition introductrice est lexicalisée. Chacune de ces prépositions peut servir de marqueur d'attachement au verbe. Une première solution, évidente, consiste à pondérer ces indices afin de «forcer» l'attachement (*cf.* Chapitre 6).

Cette solution n'est toutefois pas optimale. Le patron syntaxique de la phrase élémentaire ne nous semble pas suffisamment contraignant pour départager attachement aux noms et aux

verbes, particulièrement dans le cas des prépositions *de* et *à*, très fréquentes et très ambiguës.

Pour pallier ce problème, il nous semble intéressant d'envisager certains regroupements au niveau lexical afin d'évacuer une partie plus ou moins grande de l'ambiguïté d'attachement. En ce sens, la collocation ouvre des perspectives intéressantes. Bien que l'empreinte de ces regroupements soit particulièrement forte dans les sous-langages, il s'agit d'un phénomène linguistique généralisé à l'ensemble des langues naturelles, spécialisées ou non. Une collocation désigne une combinaison de mots **récurrente** définie syntaxiquement et **co-occurrant** de manière significative. Considérée comme telle, elle forme une unité distributionnelle.

> The chief novelty of the procedure which is offered here is the extension of substitution classes to include sequences of morphemes, not merely single morphemes. We now ask not only if *A* and *B* each occur in the environment *C – D* but also if *AE* together, or *FGH*, also occur in that environment. If they do, the *A*, *B*, *AE*, *FGH* are all substitutable for each other. We may say that they are all members of one substitution class, which is nox not merely a class of morphemes but a class of morpheme sequences. The single morphemes in the class are merely the special cases where the sequence consists of one. [Harris, 1964 : p. 164]

Par conséquent, la collocation[21], que nous considérons ici comme un *chunk lexical*, nous permet de subdiviser le niveau lexical en un double processus de segmentation (par application des dictionnaires) et d'attachement (par reconnaissance des structures lexicales complexes), de même que le *chunk syntaxique* au niveau syntagmatique.

Sélection lexicale

Le lexique-grammaire, parce qu'il formalise la langue standard, n'offre pas de contraintes sémantiques fortes. La spécification linguistique d'un sous-langage, par contre, limite le focus sémantique à un domaine particulier.

> It remains to be shown as a matter of empirical fact that such formal correlations do indeed exist, that the discourses of a particular person, social group, style, or subject-matter exhibit not only particular meanings (in their selection of morphemes) but also characteristic formal features. [Harris, 1964 : p. 3–4]

Le sous-langage offre la possibilité de caractériser davantage les constituants en augmentant sensiblement leurs propriétés distributionnelles. Dès lors que nous contraignons davantage les

[21] Nous pourrions nous demander si la collocation n'est pas simplement une notion rejoignant celle de *nom composé* telle qu'elle est proposée par [G. Gross, 1996]. Nous verrons au chapitre 5 que ces deux notions se rejoignent effectivement mais que la collocation recouvre une gamme plus large de structures qui ne sont pas nécessairement figées.

compléments essentiels, nous pouvons envisager d'affiner la sélection lexicale et d'ainsi contraindre l'attachement. Ce concept nécessite bien entendu la compilation d'un lexique spécialisé propre au domaine.

4.5 Conclusion

Nous avons défini brièvement un processus syntaxique **incrémental** et **hybride** tirant parti des spécificités du **sous-langage**. Ce processus est entièrement distributionnel et vise à exploiter au mieux la théorie linguistique développée par Harris.

> In contrast with this, it is possible to describe the occurrence of each element indirectly, by successive groupings of elements into sets and the relative occurrence of the sets are fewer and simpler than the total statements about the relative occurrence of each element directly. [Harris, 1981b : p. 4]

D'un point de vue pratique, la subdivision des niveaux lexical et syntagmatique en deux procédures successives de segmentation et d'attachement nous amène à considérer deux couches d'abstraction supplémentaires : le *chunk lexical*, ou collocation, et le *chunk syntaxique*. Dans une certaine mesure, ces regroupements relèvent de l'**étiquetage** puisqu'une étiquette grammaticale, celle de la tête de la séquence, sera attribuée à la séquence (réduisant ainsi sa complexité). C'est pourquoi nous les considérons comme intermédiaires aux deux plans d'analyse (*cf.* Figure 4.3).

Par ailleurs, cette double subdivision nous amène à considérer l'impératif d'adaptabilité inhérent à l'extraction d'information. La séparation des entrées lexicales simples et des entrées complexes, d'une part, et des structures syntaxiques minimales et des structures de phrases élémentaires, d'autre part, isole implicitement les ressources linguistiques propres à la langue standard de celles propres au sous-langage. De cette manière, le processus d'adaptation à un nouveau domaine s'applique à un sous-ensemble des ressources linguistiques.

D'un point de vue plus général, en plus de résoudre les problèmes isolés au chapitre 3, cette procédure répond aux deux questions que nous posions en fin de chapitre 2 :

Solution linguistique ou statistique ?

Nous avons déjà abordé cette question en traitant la problématique de l'étiquetage (*cf.* Section 3.2.3). Dans ce cadre précis, nous avons démontré qu'une désambiguïsation statistique préalable à l'analyse syntaxique n'offrait pas de réelle plus-value en termes de performance. Dans les chapitres suivants, nous poursuivrons cette réflexion en envisageant l'apport statistique au niveau lexical, par la mise au point d'une procédure d'apprentissage de collocations spéciali-

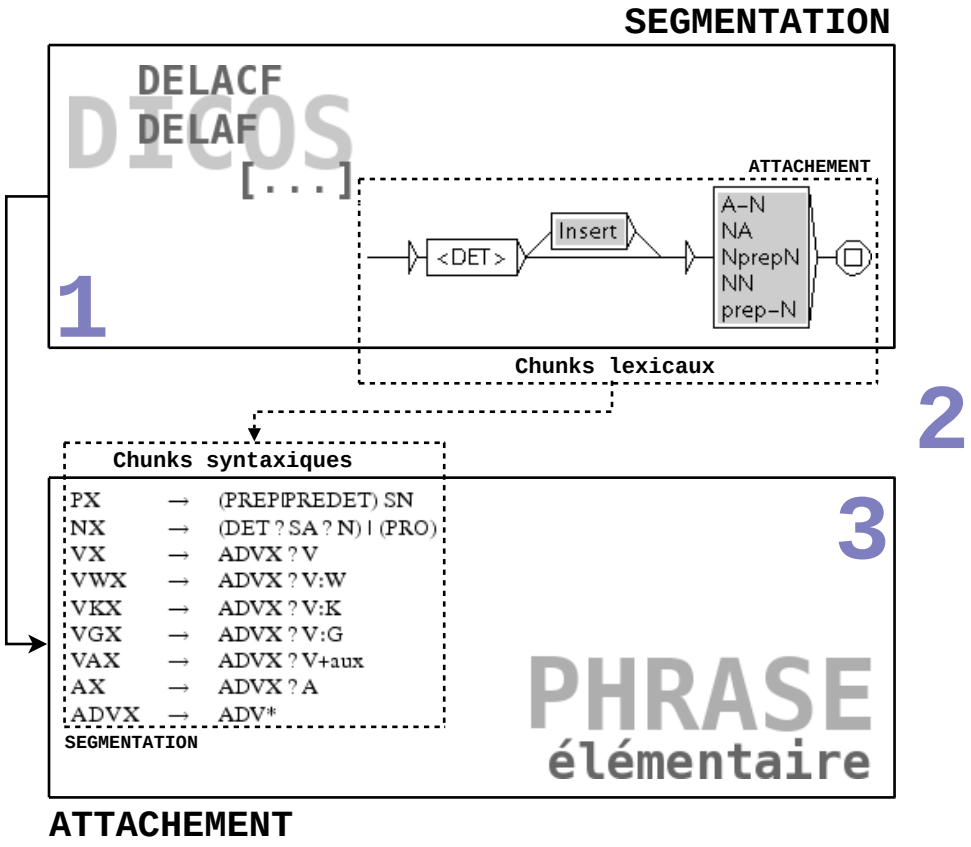


Figure 4.3 : Procédure incrémentale de segmentation/attachement

sées, et au niveau syntaxique par l'introduction de la pondération.

Analyse profonde ou partielle ?

L'analyse suggérée au chapitre 3 relève bien d'une procédure syntaxique profonde même si nos reconnaisseurs limitent l'analyse aux prédicats du domaine et ne couvrent pas nécessairement toute la complexité de la phrase. Au contraire, les chunks lexicaux et syntaxiques réduisent successivement la complexité linguistique aux têtes des structures qu'ils définissent. Par conséquent, cette procédure relève du paradigme de l'analyse de surface.

Nous explorons dans les prochains chapitres les hypothèses que nous avons formulées ici. Le chapitre 5 s'intéressera au niveau lexical tandis que le chapitre 6 étudiera le niveau syntaxique. Dans la mesure où l'intérêt d'une procédure dépend directement des résultats qu'elle obtient, nous nous efforcerons d'estimer l'apport des solutions que nous avons définies au regard du processus général d'extraction.

CHAPITRE 5

APPRENTISSAGE LEXICAL ET SOUS-LANGAGE

5.1 Introduction

Intuitivement évidente, la question de l'analyse lexicale est nettement plus complexe qu'il n'y paraît. En effet, il ne suffit pas d'établir le lexique des formes simples pour prétendre analyser une langue. Les mots composés de même que les noms propres demandent également toute notre attention. Par ailleurs, si l'on envisage l'analyse d'une langue de spécialité, il nous faudra encore prendre en compte un ensemble de termes qui peuvent ne pas appartenir au lexique standard (les termes de spécialité) ou qui, utilisés dans un contexte spécifique, sont sémantiquement altérés (les termes spécialisés).

Il est évident que nous ne pouvions pas couvrir ces trois articulations dans notre travail. Nous limiterons donc notre étude aux structures lexicales complexes, propres au domaine de la finance, afin de vérifier l'hypothèse d'un attachement au niveau lexical que nous avancions en fin de chapitre précédent. Dans le même temps, nous envisagerons la mise au point d'une procédure d'acquisition terminologique automatique qui nous permettra, d'une part, d'*apprendre* un lexique de termes propres à un domaine de spécialité et, d'autre part, de préciser leurs traits sémantiques. Ensuite, parce que leur méconnaissance peut lourdement pénaliser le processus d'analyse, nous aborderons la question des noms propres au travers de la compilation d'un dictionnaire d'entreprises et d'une formalisation lexicale générique[1].

[1]Précisons que de nombreux travaux, avant le nôtre, abordent cette idée d'une réduction de la complexité lexicale par la prise en compte de structures complexes. L'ouvrage le plus intéressant à ce sujet est, selon nous, [G. Gross, 1996]. Toutefois, parce qu'il travaille dans un contexte général (la langue standard) et parce qu'il limite son étude aux noms composés, la couverture qu'il obtient n'est pas comparable à la nôtre. En étendant le paradigme des structures complexes à la collocation, moins figée, nous sommes à même de multiplier les entrées de manière conséquente. De plus, la prise en compte du contexte d'énonciation (le sous-langage), nous permet de considérer des entrées qui, hors du domaine traité, n'auraient pas nécessairement un comportement collocationnel. L'analyse qui en ressort est donc dépendante du domaine de spécialité et ne s'inscrit pas dans une démarche générale.

5.2 Structures lexicales complexes

L'hypothèse que nous défendons ici est la suivante : la collocation peut s'apprécier en tant qu'*unité distributionnelle* et, par conséquent, participer, au niveau lexical, de l'idée harrissienne d'une analyse syntagmatique par regroupements successifs (*cf.* [Harris, 1964]). En d'autres termes, nous voulons partir des unités lexicales simples pour obtenir un ensemble de structures complexes, les collocations. Ces structures, résultats d'un attachement au niveau lexical, sont ensuite utilisées, au niveau syntaxique, pour atteindre l'articulation en constituants. Notre objectif est donc de simplifier l'attachement syntagmatique en évacuant une partie de sa complexité au niveau lexical.

Cette idée nous amène à définir un niveau d'abstraction intermédiaire aux plans lexical et syntaxique de l'analyse syntagmatique. En ce sens, elle vient compléter la notion de *chunk* développée par [Abney, 1991] en nous permettant de travailler, au niveau syntaxique, depuis un ensemble réduit, et donc moins ambigu, d'éléments terminaux. Par conséquent, notre procédure se pose dans la continuité de l'étiquetage et permet, dans le cas d'une annotation par dictionnaire, d'évacuer une partie conséquente des ambiguïtés lexicale et structurale sur base du contexte local (tout comme le ferait un étiqueteur probabiliste). Nous réduisons en effet une séquence d'entrées potentiellement ambiguës en une unique entrée désambiguïsée. Considérons les exemples suivants :

(110) Les investisseurs apprécient particulièrement l'annonce par les dirigeants d'une estimation de **marge opérationnelle** proche de 8% sur l'exercice, soit un résultat de l'ordre de 10,4 ME, alors qu'elle atteignait 4,3% sur l'année 2002.

(111) CFF Recycling a ainsi retrouvé au 30 septembre 2003 une **marge d'exploitation** de 3,8% et une marge nette de 1,7%.

(112) Le groupe prévoit pour l'exercice 2003-2004, sur 15 mois, un **chiffre d'affaires net** consolidé de l'ordre de 30 Millions d'Euros et une rentabilité nette de 7% environ.

Les deux séquences mises en évidence au sein des exemples 111 et 112 – *marge d'exploitation* et *chiffre d'affaires net* – génèrent une combinatoire lexicale assez importante, 10 analyses possibles pour la première et 120 pour la seconde (*cf.* Figures 5.1 et 5.2). En considérant ces deux séquences en tant qu'unités, nous élaguons la combinatoire globale de manière conséquente[2].

Par ailleurs, les structures lexicales complexes de type *NA*, *NprepN* et leurs dérivés nous permettent d'évacuer, dès le niveau lexical, une part de l'ambiguïté d'attachement adjectival et

[2]Cette conception rejoint la notion de figement du lexique-grammaire et l'*idiom principle* de Sinclair que nous abordons au point 1.2.2.

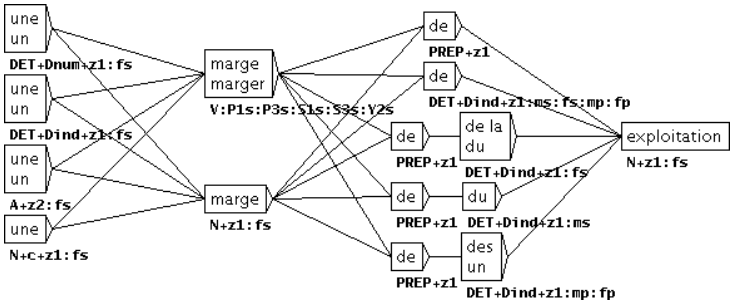


Figure 5.1 : Combinatoire lexicale : marge d'exploitation

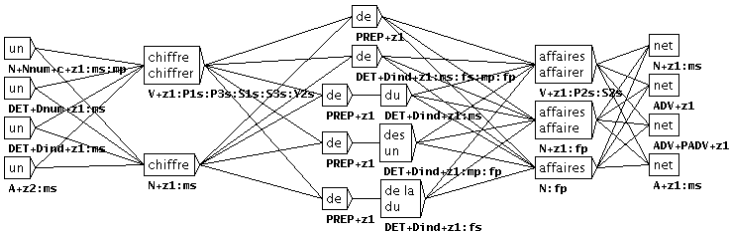


Figure 5.2 : Combinatoire lexicale : chiffre d'affaires net

prépositionnel au nom.

5.2.1 La collocation comme unité distributionnelle

Le *morphème* correspond à l'unité distributionnelle minimale et, tel qu'il est défini par Harris, désigne, dans une certaine mesure, ce que nous appellerions plus volontiers aujourd'hui l'*entrée lexicale*. L'ensemble des morphèmes d'une langue forme donc son lexique et est en ce sens l'unité nécessaire à toute analyse. Dans les paragraphes suivants, nous isolons les propriétés de cette unité pour asseoir notre définition de la collocation.

La notion de morphème, et plus généralement d'unité distributionnelle, se définit à deux niveaux, syntaxique et sémantique :

> In the continuous morphemes, the fundamental criterion which determines that the whole of a sequence of phonemes constitutes one morpheme rather than two, is the fact that the whole sequence occurs together **in a certain class of positions** and **with certain meanings**, and that parts of the sequence do not occur separately with parts of the total meaning of the sequence. [Harris, 1981b : p. 23]

La première étape de toute étude distributionnelle consiste à organiser les unités de la langue en classes lexicales d'équivalences afin de pouvoir traiter un énoncé non plus comme une suite d'éléments discrets mais plutôt comme une suite de classes lexicales, généralisées autour d'une unique étiquette grammaticale. Ainsi, sont jugées équivalentes d'un point de vue syntaxique les unités pouvant se substituer mutuellement dans un contexte particulier (*a certain class of positions*). Si l'on considère le morphème X dans le contexte syntaxique $A - B$, une entrée Y donnée sera considérée comme un morphème de la même classe si l'énoncé $A\ Y\ B$ reste acceptable

grammaticalement. Nous utilisons donc un test de substitution afin, d'une part, d'identifier les frontières de morphèmes et, d'autre part, de les organiser en classe d'équivalences.

Notons que la composante sémantique (*with certain meanings*) est quant à elle plus un constat *a posteriori* qu'une véritable propriété. Il n'est pas possible en effet d'isoler un mot sur base de sa signification sans recourir au jugement humain. Hors, le distributionnalisme repose uniquement sur l'observable. Toutefois, un mot considéré en contexte actualise une sémantique particulière. Nous pouvons donc associer un sens précis à chaque actualisation d'un morphème donné. Nous retrouvons ici le concept d'une désambiguïsation contextuelle que nous observons également pour chaque prédicat dans le contexte de la phrase élémentaire. Par ailleurs, notons que les tables du lexique-grammaire reposent sur le même principe de substitution qui, dans leur cas, est appliqué sur un nombre plus important de propriétés (les propriétés des phrases élémentaires).

Dans un premier temps, Harris a limité le test de substitution aux mots simples. Il a néanmoins dû constater que certaines séquences de morphèmes présentent les mêmes caractéristiques que les morphèmes simples. Ils peuvent se substituer à eux et, ainsi, élargir les classes d'équivalences. La forme complexe devient, dans ce contexte, un cas particulier d'entrée lexicale tout comme les entrées simples. C'est un constat que nous retrouvons également dans [Silberztein, 1993] et [G. Gross, 1996] pour le lexique-grammaire.

D'un point de vue syntaxique, les exemples 113 à 118 définissent la structure élémentaire : N_0 *V* N_1 *sur* N_2, avec N_1 prédicatif et *V* support. En étendant le principe de substitution aux entrées complexes, nous pouvons associer au prédicat *offre* les prédicats *offre publique d'achat*, *offre de rachat*, *offre publique d'échange*, *offre de retrait* et *offre de reprise*.

(113) France Télécom *a lancé* une **offre** sur les titres de sa filiale non détenus qui valorise le FAI 8,86 euros par action contre 7,56 euros vendredi soir à la clôture.

(114) Le groupe de François Pinault *a lancé* son **offre publique d'achat** sur le solde du capital de Gucci à 85,52 Dollars par action après que le couturier italien eut distribué un dividende exceptionnel de 13,50 euros par action.

(115) Tandis que le fonds d'investissement Blackstone Group *a lancé* une **offre de rachat** amicale sur la société ce matin, la direction du chimiste allemand vient d'annoncer qu'elle avait conclu auprès de BASF la vente de la division nylon de sa filiale Ticona.

(116) France Télécom a annoncé les résultats de l'**offre publique d'échange** qu'il *a lancé* sur Orange.

(117) Comme attendu, Crédit Agricole SA *a déposé* une **offre de retrait** suivi d'un retrait obligatoire sur les titres Crédit Lyonnais qu'il ne détient pas encore.

(118) Selon le Wall Street Journal, Edgar Bronfman Junior *prépare* une **offre de reprise** sur les actifs américains de Vivendi Universal Entertainment.

De la même manière, au niveau des arguments, les exemples 119 à 123 nous permettent d'associer au substantif *inflation*, sujet de la structure N_0 *<progresser> de* N_1 à N_2 sur N_3, les entrées *indice des prix à la consommation*, *bénéfice net*, *résultat opérationnel* et *marge opérationnelle*.

(119) Selon les chiffres publiés par le Département du travail américain, et corrigés des variations saisonnières, l'**inflation** *a progressé* de 0,2% sur un mois, en ligne avec le consensus des économistes, et de 2,1% sur un an.

(120) Ainsi, l'**indice des prix à la consommation** *a progressé* de 0,1% sur un mois et de 2% en glissement annuel.

(121) Le **bénéfice net** *a progressé* de 1,3% à 632 millions de dollars, contre 624 millions un an plus tôt.

(122) Le **résultat opérationnel**, plus représentatif, *a progressé* de 12,8%, à 1,39 milliard d'euros, faisant ressortir une marge opérationnelle de 35,6% (contre 33,5% un an auparavant).

(123) La **marge opérationnelle** *a progressé* à 4%, contre 3,6% un an plus tôt, dont 4,9% pour la France.

Vue dans ce cadre syntaxique, la collocation se définit comme une unité grammaticale. Cette notion est couramment acceptée en linguistique. [Sinclair, 1997], par exemple, considère la collocation comme un choix linguistique simple. Ce choix s'exprime au niveau syntaxique mais également, nous l'avons vu, au niveau sémantique (deuxième propriété distributionnelle).

Intuitivement, les structures lexicales des exemples 114 à 118 et 120 à 123 nous semblent former une unité sémantique. Nous pouvons voir une confirmation de cette intuition en envisageant l'exemple 113 qui, au regard des entrées prédicatives complexes, apparaît ambigu. Nous ne pouvons pas déterminer le type d'offre dont il est question. Il nous faut prendre en compte le contexte plus large de la phrase pour envisager une offre (publique ?) d'achat, puisqu'elle porte sur des titres *non détenus*.

Le principe d'unité sémantique est central à la notion de collocation. [Firth, 1969a] est le premier à mettre ce point en évidence de même qu'il semble être à l'origine du terme *collocation*. Ses développements sont particulièrement intéressants et soulignent un point fondamental, sur lequel nous revenons plus tard dans ce chapitre. Si la collocation se présente bien comme une unité sémantique, le sens qui s'en dégage n'est pas contextuel. La notion de contexte intervient nécessairement mais le sens d'une collocation ne peut pas se calculer au départ des relations fonctionnelles qui la lient à son contexte syntaxique, situationnel et culturel ([Firth,

1969a : p. 195]). Cette position peut paraître extrême, la collocation est en effet une notion très fluctuante. Toutefois, elle nous permet d'introduire l'idée du figement. Le sens et la syntaxe d'une collocation sont plus ou moins figés, d'où la nécessité de la considérer en tant qu'unité du point de vue grammatical et sémantique. Le figement se traduit donc doublement : d'une part, il ne nous est pas toujours possible d'envisager le sens global de l'unité au départ du sens de ses composantes (principe de non-compositionnalité) et, d'autre part, la collocation est très peu sujette à la variation syntaxique (insertion, substitution, *etc.*).

En marge des définitions syntaxique et sémantique, nous devons encore étudier un paramètre important : la statistique. Une collocation, qu'elle soit plus ou moins figée, organise un ensemble de collocats autour d'une tête lexicale, *offre* dans les exemples 114 à 118 ([Sinclair, 1997 : p. 115]). Il s'agit donc d'une relation *opérateur – argument*, bien que ce ne soit pas toujours perceptible eu égard au figement de la structure. Dès lors, il nous semble intéressant d'ajouter à la définition la contrainte de vraisemblance énoncée par Harris :

> Each word has a somewhat fuzzy selection of other words that are more likely than average to occur in the position for its argument – that is, more likely than would be expected if the occurrences were random or equal in frequency. [Harris, 1988 : p. 17]

Nous retrouvons là un point fondamental de la définiton de la collocation nous permettant d'entériner définitivement notre hypothèse d'une unité distributionnelle et de prétendre à son utilisation dans un processus d'analyse syntaxique automatique.

5.2.2 Collocation et traitement automatique des langues

Nous citons ici quelques références propres au domaine du traitement automatique des langues et utilisant, de près ou de loin, de manière explicite ou non, la notion de collocation. Nous ne dressons pas un état exhaustif de la question mais présentons brièvement différents modes d'exploitation de la collocation : l'étiquetage, la désambiguïsation et la traduction automatique.

Étiquetage et analyse automatique

[G. Gross, 1996] étudie le figement qu'il présente comme une gageure pour le traitement informatique de texte en ce qu'il peut ne pas respecter les règles communément acceptées par la syntaxe et qu'il est très largement utilisé, particulièrement en langue de spécialité. L'ignorer compromet donc fortement l'analyse.

> Ces suites fonctionnent de façon compacte, en bloc, exactement comme des catégories simples (à des problèmes de graphie près) et c'est à tort ou pour des raisons

> de commodité rédactionnelle qu'on les fait figurer sous l'un ou l'autre des termes dans les dictionnaires. [G. Gross, 1996 : p. 16]

Bien qu'il n'emploie pas le terme de collocation, lui préférant celui de *locution*, de *figement* ou encore d'*unité polylexicale*, le concept d'unité catégorielle qu'il met en avant rejoint ce que nous disions à propos de la collocation. Il associe également au figement une opacité sémantique due à la non-compositionnalité du sens au sein de ces expressions et souligne donc une certaine unité sémantique, également présente dans notre définition de la collocation.

Selon que la structure figée soit prédicative ou non, son étude conduira à la création d'une table de lexique-grammaire ou d'un dictionnaire de formes complexes. L'une comme l'autre sont exploités dans les procédures d'analyse automatique ; le dictionnaire permet l'étiquetage des structures complexes et la ou les tables leur analyse.

La position de Gaston Gross est celle d'un lexicologue et ses orientations théoriques sont celles du lexique-grammaire. Par conséquent, il privilégie une étude systématique des structures figées. Nous retrouvons ce point de vue chez [Gross, 1982] et, également, dans le projet BFQS qui vise le recensement des expressions prédicatives figées francophones (**b**elges, **f**rançaises, **q**uébecoises et **s**uisses, *cf.* [Lamiroy *et al.*, 2003]). De notre côté, nous procédons à une lecture statistique des données textuelles et substituons à la compilation manuelle de dictionnaires, une acquisition automatique de structures complexes.

Cette conception lexicologique ne s'observe pas qu'au sein des disciplines associées au traitement automatique des langues. Il est évident que les tours lexicaux complexes posent problème aux apprenants de la langue. Ainsi, Sinclair est l'instigateur d'un dictionnaire destiné aux apprenants et dont la particularité est d'intégrer la collocation comme entrée lexicale afin de mettre l'utilisateur au contact direct de l'usage linguistique (*cf.* [Hunston et Francis, 2000]).

Désambiguïsation

Selon qu'elle est perçue en tant qu'unité syntaxique ou sémantique, la collocation permet de traiter l'ambiguïté d'attachement prépositionnel ou l'ambiguïté sémantique.

[Hindle et Rooth, 1998] se propose de traiter l'attachement prépositionnel en explorant les préférences lexicales de différentes co-occurrences à complément prépositionnel. Son idée est d'extraire automatiquement une table de bigrammes du type $V + prep$ ou $N + prep$. Cette table servirait ensuite de référence afin de résoudre les cas problématiques.

L'hypothèse qui sous-tend cette expérience est que chaque terme possède une signature distributionnelle. Cette hypothèse est précisément la clé de voûte du lexique-grammaire et des études harrissiennes en général. Les tables du lexique-grammaire établissent cette signature

au niveau prédicatif. Les propriétés qu'elles définissent, de même que leur phrase élémentaire, nous dressent une cartographie exacte et systématique de la distribution des entrées prédicatives d'une langue et de leurs compléments essentiels (prépositionnels ou non). Par conséquent, dans ce chapitre, notre étude de la collocation vise à compléter l'information des tables en étendant l'étude distributionnelle aux éléments non prédicatifs.

[Yarowsky, 1993] exploite, quant à lui, le principe d'unité lexicale des collocations dans un algorithme de désambiguïsation sémantique. Il place son étude dans un contexte de traduction. Un mot de la langue source est considéré sémantiquement ambigu s'il connaît plusieurs traductions possibles en langue cible. Pour chaque mot ambigu, l'algorithme d'entraînement calcule les collocations possibles à droite et à gauche, c'est-à-dire la distribution lexicale du mot, et produit en sortie une table distributionnelle probabilisée dans laquelle chaque sens est associé à un ensemble de collocations. Les résultats sont ensuite utilisés afin de désambiguïser de nouveaux textes. Cette procédure affiche une précision de 92 % qui confirme la propriété d'unité sémantique de la collocation.

Traduction automatique

[Lehrberger, 1982] étudie les possibilités du sous-langage dans un contexte de traduction automatique. Il souligne l'interêt qu'offre l'unité syntactico-sémantique des structures lexicales complexes tant du point de vue de l'analyse syntaxique que du point de vue de la traduction proprement dite.

Au niveau de l'analyse syntaxique, la prise en compte des structures complexes, peu importe qu'elles soient compilées au sein d'un dictionnaire ou formalisées à la manière d'une bibliothèque d'automates, offre un avantage des plus intéressants. En effet, certaines collocations actualisent des structures syntaxiques déviantes. [Lehrberger, 1982] cite pour exemple des cas d'*empilage* nominal issus de manuels d'entretien d'avions :

– external hydraulic power ground test quick-disconnect fittings
– fuselage aft section flight control and utility hydraulic system filter elements
– fan nozzle discharge static pressure water manometer
– landing gear, flight controls, speed brakes, engine air by-pass flaps, and nose steering systems
– stabilizer power control No. 2 system return line check valve failure

Pour définir ces suites nominales d'un point de vue grammatical, nous devrions introduire des règles syntaxiques *ad hoc* qui ne manqueraient pas de générer plus d'ambiguïté encore dans la procédure de rassemblement syntagmatique. Il nous paraît donc nettement plus intéressant d'évacuer les problèmes posés par la collocation dès le niveau lexical.

Au niveau de la traduction, l'intérêt est immédiat. La non-compositionnalité des collocations entraîne bien souvent une violation des règles associant une structure du langage source à sa correspondante en langage cible (*e.g. aspect ratio* – allongement ; *DC power* – courant continu ; *buttock line* – section longitudinale). Par conséquent, la traduction de ces structures doit être envisagée globalement sans considération pour les articulations lexicales qui la composent. [Smadja *et al.*, 1996] rejoint également ce point de vue et décrit un programme, *Champollion*, qui, au départ d'un corpus bilingue aligné à la phrase, permet de mettre en relation des paires de mots ou des syntagmes des deux langues et d'ainsi réaliser un alignement au niveau des structures lexicales complexes.

5.2.3 Collocation et sous-langage

Il serait inexact d'associer, de manière exclusive, le concept de collocation à celui de sous-langage. Il s'agit d'un phénomène linguistique généralisé à l'ensemble de la langue, tout domaine et tout registre confondus.

> Finally, if we take a whole connected discourse as environment, we find that there are certain substitution sets of morphemes which occur regularly (relative to the other sets) throughout the discourse or some portion of it ; these are not the major substitution sets of the language (e.g. nouns) or its grammatical subclasses, but new groupings which are often relevant only to that one discourse. [Harris, 1981b : p. 18]

[Firth, 1969a] développe d'ailleurs le concept de la collocation sur base d'exemples empruntés à la poésie, à certaines œuvres épistolaires, à la langue familière, *etc.* Par contre, nous devons constater que la toute grande majorité d'entre-elles relèvent d'un domaine particulier et constituent le *jargon* du domaine (*cf.* Tableau 5.1[3]).

Nous avons établi au chapitre précédent que la propriété de fermeture d'un sous-langage s'observait aux niveaux lexical et syntaxique. Il ressort de cette situation que nous pouvons représenter l'information propre à un domaine au départ d'un lexique clos, qui s'organise en classes sémantiques, et d'un ensemble contraint de patrons syntaxiques de type *prédicat – arguments*, formalisant les relations entre les différentes classes lexicales. Le tableau 5.1 nous permet également d'apprécier un certain nombre de relations sémantiques à un niveau non prédicatif, celui du groupe nominal. Ces relations représentent des groupes informationnels homogènes et forment ce que nous appelons *collocations* et que [Marsh, 1986] désigne par *frames*.

[3]Cet extrait est tiré d'un corpus regroupant un ensemble d'échanges entre médecins avant et après l'hospitalisation d'un patient. La compilation et l'anonymisation des données ont été réalisées automatiquement par C. Fairon, dans le cadre d'une collaboration entre le Cental et le service informatique des Cliniques universitaires Saint-Luc, dirigé par B. Debande. Ces données nous ont gracieusement été prêtées afin de pouvoir tester notre extracteur terminologique sur un sous-langage différent de la finance et d'ainsi pouvoir estimer sa portabilité de même que son efficacité.

> Sublanguage semantic patterns, which correlate with observed linguistic patterns and encode certain aspects of domain knowledge, are significant for representing the information found in texts. The semantic patterns state specific co-occurrence restrictions active in sublanguage sentences and reflect relations between objects. By identifying these sublanguage patterns, we can represent the knowledge in a domain ; and, because semantic patterns represent stereotypical situations, they can be considered frame representations of the domain. [Marsh, 1986 : p. 117]

Cher Confrère,
[Madame X] , âgée de 78 ans, a séjourné dans le service le 17.12.2002 pour **tentative de cardioversion électrique**.
Antécédents personnels :
– **résection de fibrome utérin**
– **cardiopathie dilatée** avec **altération de la fonction VG - FA chronique**
Affection actuelle :
[Madame X] est hospitalisée pour **tentative de choc électrique**.
Traitement actuel :
– Sintrom mitis 1 mg à adapter selon l'INR
– Zestril 20 mg – 1/jour
– Aldactone 50 mg – 1/jour
– Sotalex 160 mg – 1/2 2/jour
Examen clinique du 17.12.2002 :
– pouls à 90/min
– **auscultation cardiaque** normale
– **examen pulmonaire** normal
– **examen abdominal** normal
– **examen vasculaire périphérique** normal
Examens complémentaires :
– **Biologie courante** du 17.12.2002 : (cfr. en annexe)
– CPK : 35 UI/l (7-160 UI/l)
– INR : 1,63
– urée : 69.
Evolution et conclusion :
[Madame X] , a été hospitalisée pour **tentative de cardioversion électrique**. Choc non réalisé pour cause d'INR mal équilibré. Nous avons majoré les doses de Sintrom à 2-2-1. Une nouvelle hospitalisation est prévue le lundi 23/12.
Traitement médical :
– Sintrom 2 mg, 2 mg, 1 mg.
– traitement par ailleurs inchangé
Nous vous prions de croire , Cher Confrère, en l'expression de nos meilleurs sentiments.

Tableau 5.1 : Extrait du corpus de médecine interne

Tout comme les patrons syntaxiques, ces patrons sémantiques s'organisent autour d'un élément tête fixe (≈ le prédicat) et d'un ensemble de points d'insertion (≈ les arguments). Ainsi, par exemple, le terme *résection* du tableau 5.1 nous permet de définir le patron sémantique : *<résection> <de.PREP> <DET> N A.*

– résection du tablier épiploïque
– résection du dôme vésical

– résection de l'adénome prostatique
– résection du col utérin
– résection d'un polype vésical
– résection du côlon ascendant
– résection du polype rectosigmoïdien
– résection du pontage axillo-poplité
– résection du polype adénovilleux

De la même manière, pour le sous-langage de la finance, les exemples 114 à 118 forment un patron sémantique du type <*offre*> *A* <*de*> *N*, avec *A* facultatif. Nous pouvons donc conclure que chaque sous-langage organise l'information relevant du domaine en un nombre clos de patrons lexicaux et syntaxiques qui lui sont propres. Les patrons lexicaux, plus encore que les patrons syntaxiques, reflètent la sémantique du domaine de spécialité et, en ce sens, offrent également une base solide nous permettant d'identifier, ou non, l'appartenance d'un texte donné à un domaine particulier.

5.2.4 Procédure d'extraction

[Grefenstette, 1993] évalue les performances de deux techniques d'extraction de structures nominales complexes. Ces deux techniques diffèrent par leur définition du contexte lexical. La première consiste à définir un ensemble de patrons syntaxiques caractérisant les différentes structures à extraire (*e.g.* [Grefenstette, 1992; Daille, 1995]). Ces patrons sont appliqués à un texte désambiguïsé et les candidats identifiés sont ensuite évalués statistiquement. La seconde technique développe le même type d'évaluation statistique mais procède à l'identification des candidats à l'aide d'une fenêtre de lecture sans assurer une désambiguïsation préalable. Le contexte d'un nom sera alors calculé par combinaison du nom à l'ensemble des mots identifiés au sein de la fenêtre et pouvant être analysés comme nom, verbe ou adjectif (*e.g.* [Phillips, 1985; Evans *et al.*, 1991]).

Notons que [Smadja, 1993] associe les deux techniques. Il procède, dans un premier temps, à l'identification des candidats par le biais d'une fenêtre de lecture et filtre ensuite les résultats, rejetant les candidats ne ressortissant pas aux structures syntaxiques pertinentes.

[Grefenstette, 1993] conduit son évaluation par comparaison des résultats à des ressources lexicales composées manuellement (dictionnaires, thésaurus, *etc.*). Il ressort de son étude que les solutions dirigées par la syntaxe offrent un meilleur recouvrement des ressources manuelles tandis que la fenêtre de lecture permet une meilleure prise en compte des structures moins fréquentes. D'un autre côté, la fenêtre de lecture entraîne une plus grande combinatoire qui se traduit par un plus long temps de traitement (4 jours pleins sur un DEC-5000 contre 3 heures et

demie pour la solution syntaxique).

Nous ajouterons également à ces résultats le constat suivant : si les techniques syntaxiques nous permettent d'éviter le bruit engendré par l'utilisation d'un *sac de mots*, elles empêchent l'apprentissage de collocations dont la structure syntaxique n'est pas recensée.

Nous privilégions quant à nous la l'acquisition des unités les plus fréquentes. En ce sens, nous croyons préférable de nous engager dans une solution syntaxique. Cette dernière méthode correspond, par ailleurs, à une certaine idée du distributionnalisme et nous semble plus à même de refléter la logique linguistique.

Suivant cette logique, nous avons déterminé les patrons syntaxiques circonscrivant les structures polylexicales nominales (bigrammes et trigrammes) les plus communes (*cf.* 5.2.5). Notre procédure opère en trois temps.

Phase 1 : étiquetage

Nous utilisons le *TreeTagger*[4] afin d'étiqueter le texte traité. De cette manière, nous évacuons l'ambiguïté grammaticale, principale source de bruit, et nous pouvons associer à chaque entrée lexicale un lemme et une étiquette grammaticale uniques (*cf.* Figure 3.2).

Rappelons que le *TreeTagger* affiche une précision de 96,34%, soit une erreur tous les 27 mots. Dans le cadre d'une analyse syntaxique, de tels résultats peuvent facilement conduire à des incohérences. Toutefois, notre procédure d'extraction terminologique repose avant tout sur l'accumulation de données. Dès lors, et puisqu'il ne s'agit pas d'apprécier la phrase dans son ensemble mais bien des segments réduits de celle-ci, les erreurs n'interfèrent pas réellement avec la qualité des résultats. Quand elles se produisent, elles introduisent un bruit à peine perceptible au niveau fréquentiel. Nous abordons ce problème dans l'analyse des résultats (*cf.* Section 5.2.6).

[4]Nous avons dû régler beaucoup de problèmes pour intégrer le *TreeTagger*. Il n'affiche pas une robustesse très convaincante. Ainsi, par exemple, le processus de segmentation en phrases et en tokens devrait être revu en profondeur, principalement pour les titres qui ne sont jamais considérés comme des phrases et sont systématiquement associés à la phrase suivante en raison de l'abscence de ponctuation finale. Les répercussions sur la qualité de l'étiquetage sont par conséquent inévitables. En outre, la lemmatisation des termes français peut être tout simplement aberrante (*e.g.* cours de bourse – *cour* de bourse). Nous ne remettons pas en cause les fondements théoriques soutenant le *TreeTagger*, qui par ailleurs n'a pas été à l'origine développé pour le français, mais nous croyons qu'il serait intéressant, et profitable à notre algorithme, de revoir en profondeur les ressources françaises (principalement lexicales) qui sont pour le moins incomplètes et à certains points de vue erronées.

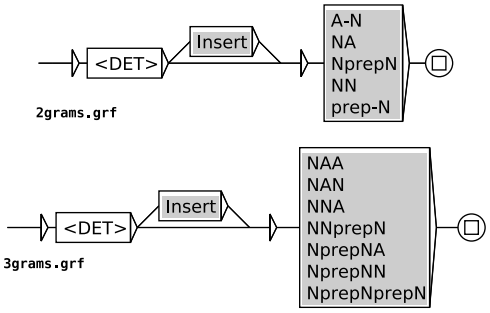


Figure 5.3 : Graphes d'extraction principaux

Phase 2 : extraction

Unitex nous permet d'appliquer nos graphes au texte étiqueté[5] et d'obtenir en sortie un fichier ne comptant que les candidats termes augmentés d'étiquettes XML balisant leurs différentes articulations[6] (*cf.* Figure 5.4). Une dernière étape nous permet de retenir uniquement le lemme, pour chaque étiquette DELA. L'idée ici est, bien entendu, d'optimiser les calculs fréquentiels ultérieurs. Le lemme nous permet en effet d'unifier les différentes flexions d'une même forme.

Au sein de notre formalisme XML, la balise `<ITEM>` délimite les frontières des occurrences. Ce premier élément, racine de chaque unité, possède quatre attributs :

– `size` : désigne la taille du *ngram* extrait,
– `cat` : renseigne la catégorie grammaticale du *ngram*,
– `subcat` : renseigne l'information de sous-catégorisation du *ngram*,
– `struct` : décrit la structure syntaxique interne du *ngram*.

Les informations de taille, de catégorie et de structure interviennent dans le comptage fréquentiel que nous décrivons au point suivant. Le dernier attribut, la sous-catégorisation, est, quant à lui, utilisé au niveau de la formalisation des résultats (*cf.* Section 5.2.7).

Chaque élément `<ITEM>` contient un ou plusieurs éléments `<LEX>` délimitant les frontières des différentes articulations de la structure complexe. Il requiert un unique attribut `pos` qui donne sa position au sein de la structure. Cette information qui peut paraître redondante avec l'attribut `size` de l'élément `<ITEM>` est directement utile dans le cas d'élément coordonnés (*cf. Section 5.2.5*).

Les éléments `<LEX>` pouvant, eux-mêmes, être complexes, la balise `<HEAD>` nous permet de cibler plus précisément les têtes lexicales du *ngram*. Un attribut `cat` est associé à chaque élément `<HEAD>` et indique la catégorie grammaticale de la tête. De plus, dans le cas d'une structure prépositionnelle, l'élément `<LEX>` peut également contenir une balise `<LINK>` dont la fonction

[5]Le texte est transformé en une suite d'entrées de type DELA pour permettre à *Unitex* d'accéder au lemme et à la catégorie grammaticale de chaque mot sans devoir appliquer les dictionnaires, sources d'ambiguïté.

[6]Nous utilisons pour cela la commande *Extract*, en mode *merge* (qui correspond au mode d'annotation d'*Unitex*).

est d'identifier et de catégoriser (attribut `cat`) le lien prépositionnel. Notons toutefois que la structure *prep – N* n'utilise pas la balise `<LINK>` afin d'isoler la préposition. Nous considérons, ici, la préposition comme le premier élément tête de la structure.

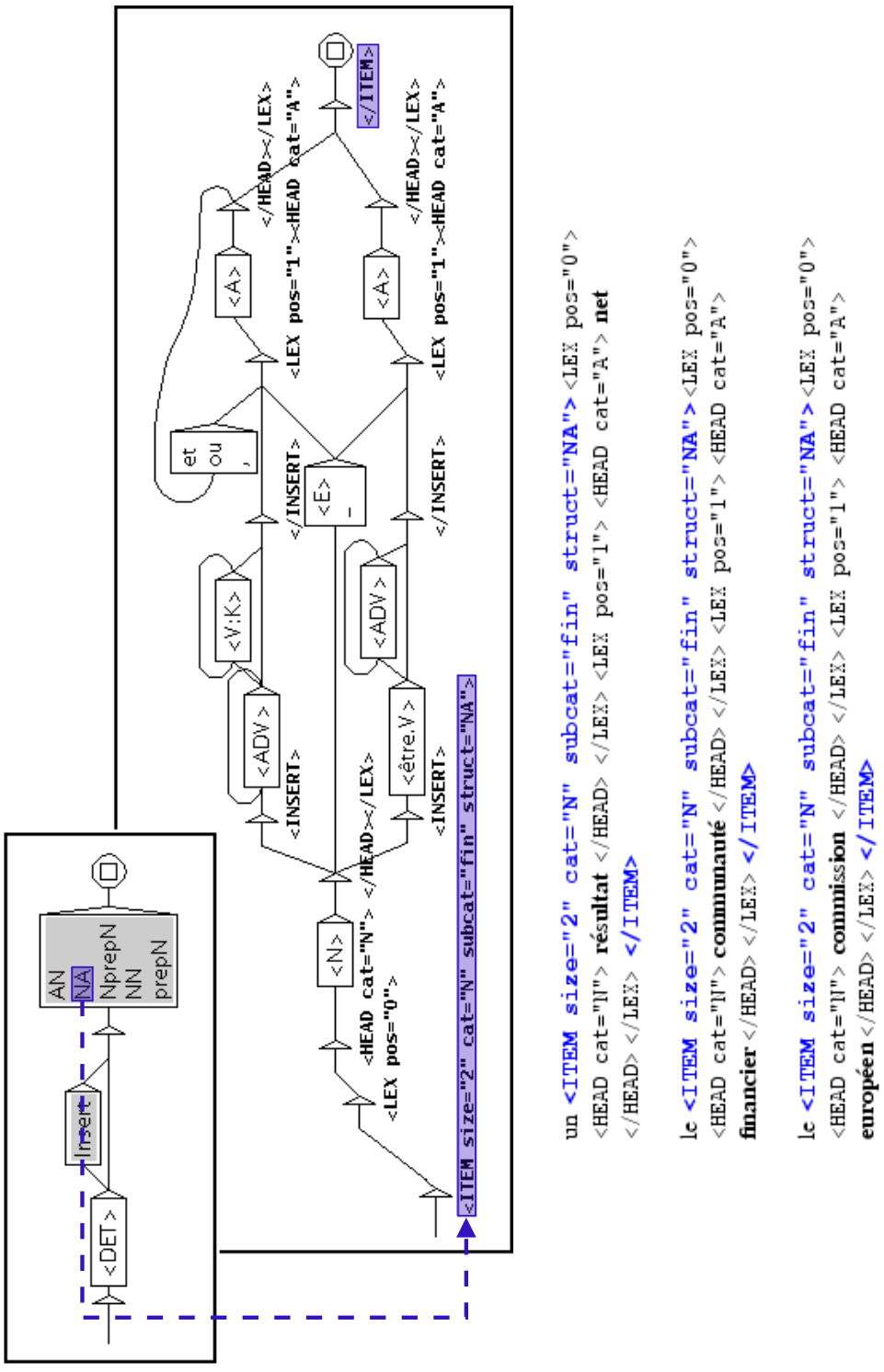


Figure 5.4 : Processus d'extraction

Phase 3 : estimation fréquentielle

> All elements in a language can be grouped into classes whose relative occurrence can be stated exactly. However for the occurrence of a particular member of one class relative to a particular member of another class, it would be necessary to speak in terms of probability based on the frequency of that occurrence in a sample. [Harris, 1981b : p. 4]

Les différentes têtes des structures extraites (balisés `<HEAD>`, dans nos graphes) forment les *ngrams* pour lesquels nous effectuons une analyse fréquentielle. Cette analyse est double, à la fois interne et externe. Pour chaque *ngram*, nous calculons non seulement la force d'association de ses composantes mais également son degré d'attachement au domaine de spécialité. La première analyse nous permet de valider, ou non, la structure extraite et la seconde analyse nous permet de lui associer une information de sous-catégorisation.

Aux deux niveaux d'analyse (interne et externe) correspondent deux variantes du *log-likelihood*. De nombreuses études consacrées aux mesures statistiques ([Dunning, 1993], [Manning et Schütze, 1999], [Daille, 1995]) présentent, en effet, le *log-likelihood* comme la solution la plus adaptée à l'étude des collocations. Nous verrons également que les propriétés caractéristiques de cette mesure offrent une certaine liberté quant à la taille des corpus considérés.

Mesure de co-occurrence

L'étude du *log-likelihood* repose sur la création d'une table de contingence exprimant les différentes fréquences associées à chaque paire (L_i, L_j) :

	L_j	$L_{j'}$ $(j' \neq j)$
L_i	a	b
$L_{i'}$ $(i' \neq i)$	c	d

où,

– a est la fréquence des *ngrams* comprenant à la fois L_i et L_j ;
– b est la fréquence des *ngrams* comprenant à la fois L_i et $L_{j'}$;
– c est la fréquence des *ngrams* comprenant à la fois $L_{i'}$ et L_j ;
– d est la fréquence des *ngrams* comprenant à la fois $L_{i'}$ et $L_{j'}$;
– $a+b+c+d$ est le nombre total (N) d'occurrences d'un motif donné.

Nous créons cette table de contingence pour chaque *ngram* obtenu après extraction des candidats termes. Les valeurs a, b, c et d sont ensuite injectées dans la formule du *log-likelihood* telle qu'elle est proposée par [Daille, 1995] :

$$\begin{aligned} LL_{(L_i,L_j)} &= a\log a + b\log b + c\log c + d\log d \\ &\quad -(a+b)\log(a+b) - (a+c)\log(a+c) \\ &\quad -(b+d)\log(b+d) - (c+d)\log(c+d) \\ &\quad +(a+b+c+d)\log(a+b+c+d) \end{aligned}$$

Le *log-likelihood* est une loi binomiale. Dès lors, son utilisation se limite au test d'hypothèses sur des *ngram* où $n = 2$, c'est-à-dire des bigrammes. Nous ne pouvons donc prétendre analyser de la sorte la force d'association entre les différents membres d'un *ngram* pour $n > 2$.

Nos structures de trigrammes offrent toutefois une perspective intéressante. En découpant un trigramme donné de part et d'autre de l'élément central, nous obtenons une suite de deux bigrammes (*e.g.* $NprepNA \rightarrow NprepN + NA$). Nous pouvons donc envisager l'estimation fréquentielle d'un trigramme et, par transitivité, de tout *ngram*, au départ des bigrammes qu'il contient. Néanmoins, nous ne pouvons simplement additionner les résultats des bigrammes. Un bigramme dont le score d'association est élevé pourrait, en effet, compromettre l'estimation du *ngram* total en lui conférant un score élevé quels que soient les éléments qui suivent. Pour pallier ce problème, [Seretan *et al.*, 2003] propose la moyenne statistique suivante :

$$ILL_{(H_0,H_1,\ldots,H_n)} = \frac{(n-1)\prod_{i=0}^{i<n-1} LL(H_i,H_{i+1})}{\sum_{i=0}^{i<n-1} LL(H_i,H_{i+1})}$$

où $LL(H_i,H_{i+1})$ correspond à la mesure de *log-likelihood* pour deux éléments successifs du *ngram*.

Mesure de récurrence

> The results of such a bigram analysis should highlight collocations common in English as well as collocations peculiar to the financial nature of the analyzed text. As will be seen, the ranking based on likelihood ratio tests does exactly this. Similar comparisons made between a large corpus of general text and a domain-specific text can be used to produce lists consisting only of words and bigrams characteristic of the domain-specific texts. [Dunning, 1993 : p.12]

[Rayson et Garside, 2000] confirme cette expérience au départ d'un corpus composé d'un ensemble de rapports ethnographiques réalisés au sein d'un centre de contrôle aérien. La méthode est à nouveau très simple à mettre en œuvre, c'est ce pourquoi nous l'avons retenue. Au départ d'un corpus de référence et d'un corpus de spécialité, nous créons deux listes de fréquences. Les fréquences de chaque *ngram*, au sein des deux corpus, nous permettent de remplir la table de contingence suivante :

	CORPUS **S**	CORPUS **R**
F_n	a	b
$!F_n$	c	d

où,

- *a* est la fréquence du *ngram* de structure S au sein du corpus de spécialité ;
- *b* est la fréquence du *ngram* de structure S au sein du corpus de référence ;
- *c* est la fréquence totale de *ngram* de structure S dans le corpus de spécialité moins la fréquence du *ngram* étudié ;
- *c* est la fréquence totale de *ngram* de structure S dans le corpus de référence moins la fréquence du *ngram* étudié.

Le *log-likelihood*, tel qu'il est précisé par [Rayson et Garside, 2000], repose d'une part sur les valeurs *observées* (*O*) du *ngram*, notées *a* et *b* au sein de la table de contingence et, d'autre part, sur les valeurs *attendues* (*E*) du *ngram*. Ces dernières se calculent à l'aide des deux formules suivantes et permettent ensuite de résoudre l'équation générale :

$$\begin{aligned} E_S &= \frac{((c+a)*(a+b))}{(c+a+d+b)} \\ E_R &= \frac{((d+b)*(a+b))}{(c+a+d+b)} \end{aligned}$$

$$LL = 2*((a*\log\frac{a}{E_S})+(b*\log\frac{b}{E_R}))$$

Il ne s'agit pas ici de valider la vraisemblance du *ngram* mais plutôt son appartenance à une langue de spécialité donnée. Dès lors, un *ngram* n'est pas considéré comme une suite de *n* unités mais comme une unique entité. Par conséquent, même si la valeur de *n* est supérieure à deux, le *log-likelihood* proposé ci-dessus reste opérationnel.

5.2.5 Typologie des structures

En marge des bigrammes, nous étudions également un ensemble de trigrammes. Au-delà de cette taille, les résultats deviennent moins stables et demandent une vérification humaine accrue compromettant l'automatisation du processus.

Nous visons l'identification de collocations nominales exclusivement, ce qui n'empêche pas l'extension de notre procédure aux structures verbales, par exemple. Nous avons d'ailleurs utilisé cette méthode d'extraction afin d'identifier certains prédicats verbaux complexes (*cf.* Exemples 124 à 126).

(124) La semaine passée, SABMiller **a pris le contrôle** du capital de l'italien Birra Peroni pour un montant estimé à environ 250 millions de dollars.

(125) A en croire le journal allemand, Financial Times Deutschland, Veolia Environnement, n'aurait pas renoncé à **prendre une participation** dans son homologue allemand, la compagnie d'eau, Gelsenwasser.

(126) Axa Private Equity, filiale d'investissement du groupe d'assurances, **monte au capital** de la société Illiad.

Nos filtres syntaxiques reposent sur un ensemble de têtes lexicales (*e.g.* les articulations des bigrammes et trigrammes). Bien qu'essentiellement nominales et adjectivales, nous introduisons également les structures *prep* et *prepN* et définissons de cette manière les différents *modifieurs* de la tête nominale du *ngram* (*cf.* Figure 5.5).

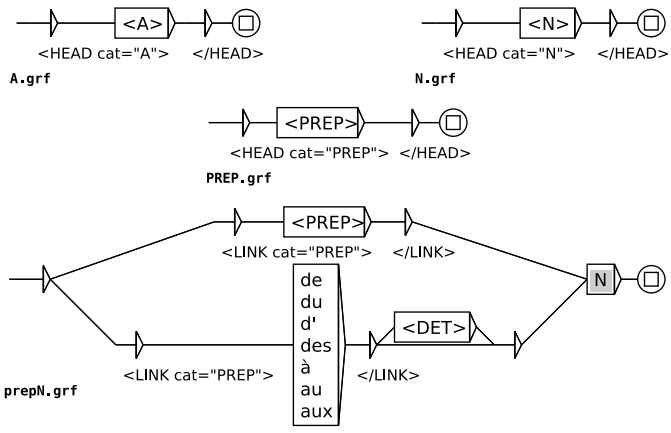


Figure 5.5 : Éléments terminaux

Interférences structurelles

Les collocations sont avant tout des unités sémantiquement homogènes. Par conséquent, leur structure interne n'est pas nécessairement figée.

> An expression which forms an idiom in all contexts in a sublanguage when its components are contiguous may also occur with the same meaning but with its components separated under certain conditions. [Lehrberger, 1982 : p. 101]

[G. Gross, 1996] renseigne deux classes de variation au sein d'un figement : les possibilités de paradigme (*e.g.* vin [*rouge*|*blanc*|*rosé*|*gris*]) et d'insertion (un pré *très* vert, des yeux *très* bleus, une lumière *très* blanche). Dans [Dister *et al.*, 2006], nous avons étudié ces différentes possibilités. Il ressort de notre étude que le figement est effectivement sujet à variation. Nous pouvons, par exemple, imaginer des occurrences comprenant de longues séquences insérées. Ce constat n'est pas nouveau. Toutefois, nous montrons que, dans les faits, ces insertions sont extrêmement limitées, principalement à un ou deux mots.

La collocation n'est pas toujours figée, il s'agit plus d'un usage ou d'une convention lexicale implicite relevant d'un domaine particulier. Des formes telles que *NA* et *NprepN* présentent une structure de groupe nominal, le degré de figement peut donc varier sensiblement ([G. Gross, 1996 : p. 50]). [Daille, 1995] dénombre trois altérations structurelles différentes : la modification (c'est-à-dire l'insertion), la coordination et la juxtaposition. De notre côté, nous ne considérons pas la juxtaposition comme une variation ; nous l'envisageons au niveau des structures de trigrammes.

La modification

Certaines unités peuvent être augmentées d'un ou plusieurs modifieurs (adjectifs ou adverbes). Par conséquent, selon la complexité, et le degré de figement attendu de la structure, nous permettons ou refusons la présence d'insertions (essentiellement adjectivales), balisées `<INSERT>`.

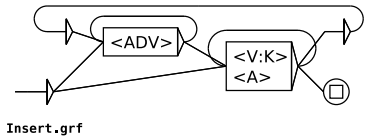


Figure 5.6 : Insertion

(127) plate-forme `<INSERT>`centrale`</INSERT>` de traitement et de production

(128) résultat `<INSERT>`brut`</INSERT>` d'exploitation

(129) rentabilité `<INSERT>`plus`</INSERT>` élevée

(130) demandes `<INSERT>`hebdomadaires`</INSERT>` d'allocation chômage

(131) résultats <INSERT>définitifs</INSERT> au titre de l'exercice

(132) évolution <INSERT>récente</INSERT> des marchés d'actions

(133) amortissement <INSERT>exceptionnel</INSERT> d'écarts d'acquisition

(134) filiale <INSERT>britannique</INSERT> d'acier carboné

L'une ou l'autre insertion adjectivale forme avec le substantif droit une structure de bigramme de type *NA*. Nous pouvons donc, au départ d'un bigramme (*cf.* Exemples 135 et 136) ou d'un trigramme (*cf.* Exemples 137 à 139) sujet à l'insertion, extraire des *ngram* de taille supérieure.

(135) $NprepN \rightarrow NAprepN$: structure <INSERT>financière</INSERT> du groupe

(136) $NprepN \rightarrow NAAprepN$: endettement <INSERT>financier net</INSERT> du groupe

(137) $NprepNprepN \rightarrow NAprepNprepN$: édition <INSERT>spéciale</INSERT> du bulletin interne

(138) $NprepNA \rightarrow NAprepNA$: groupe <INSERT>industriel français</INSERT> du secteur de la mécanique

(139) $NprepNN \rightarrow NAprepNN$: demandes <INSERT>hebdomadaires</INSERT> d'allocation chômage

La coordination

Les structures *NA* et *NprepN*, dont le degré de figement est moindre, présentent une particularité intéressante, la coordination. Si nous ne la prenions pas en compte, la couverture de notre système d'extraction terminologique ne serait, bien entendu, pas optimale. L'information positionnelle que nous annoncions plus avant nous permet de décompresser la coordination et de créer, au départ d'une extraction unique, autant de candidats que nous avons d'éléments coordonnés (*cf.* [Dominguès, 1999]). La procédure est simple : pour chaque lemme de position n, nominal ou adjectival selon le cas, nous créons une entrée en l'associant au lemme de position $n-1$. Ainsi, dans le cas d'une extraction du type des exemples 140 à 146, nous créerons les entrées du tableau 5.2.

(140) <LEX pos="0"> **secteur** </LEX> <LEX pos="1"> **du ciment** </LEX>, <LEX pos="1"> **du béton** </LEX>, <LEX pos="1"> **du granulat** </LEX>, <LEX pos="1"> **du transport** </LEX> et <LEX pos="1"> **du négoce** </LEX>

(141) <LEX pos="0"> **activité** </LEX> <LEX pos="1"> **de crédit** </LEX> <LEX pos="2"> **aux particuliers** </LEX> et <LEX pos="2"> **aux entreprises** </LEX>

(142) <LEX pos="0"> **ralentissement** </LEX> <LEX pos="1"> **des flux** </LEX> <LEX pos="2"> **d'importations** </LEX> et <LEX pos="2"> **d'exportations** </LEX>

(143) <LEX pos="0"> **industrie** </LEX> <LEX pos="1"> **aéronautique** </LEX> <LEX pos="2"> **civile** </LEX> et <LEX pos="2"> **militaire** </LEX>

(144) <LEX pos="0"> **financement** </LEX> <LEX pos="1"> **de croissance** </LEX> <LEX pos="2"> **interne** </LEX> et <LEX pos="2"> **externe** </LEX>

(145) <LEX pos="0"> **activité** </LEX> <LEX pos="1"> **gestion** </LEX> <LEX pos="2"> **d'espace** </LEX> et <LEX pos="2"> **d'événements** </LEX>

(146) <LEX pos="0"> **division** </LEX> <LEX pos="1"> **services** </LEX> <LEX pos="2"> **bancaires** </LEX>, <LEX pos="2"> **financiers** </LEX> et <LEX pos="2"> **technologiques** </LEX>

Exemple 140
secteur du ciment
secteur du béton
secteur du granulat
secteur du transport
secteur du négoce

Exemple 141
activité de crédit aux particuliers
activité de crédit aux entreprises

Exemple 142
ralentissement des flux d'importations
ralentissement des flux d'exportations

Exemple 143
industrie aéronautique civile
industrie aéronautique militaire

Exemple 144
financement de croissance interne
financement de croissance externe

Exemple 145
activité gestion d'espace
activité gestion d'événements

Exemple 146
division Services bancaires
division Services financiers
division Services technologiques

Tableau 5.2 : Décompression des structures coordonnées

Structures de bigrammes

Sur base de la typologie de [Mathieu-Colas, 1996] qui recense plus de 700 structures nominales complexes, [G. Gross, 1996] identifie 17 structures élémentaires (*cf.* Tableau 5.3).

À partir de ces structures élémentaires, nous avons défini cinq filtres syntaxiques : $A-N$, NA, $NprepN$, NN et $prep-N$ (*cf.* Figure 5.7). Ces filtres ne rencontrent pas l'ensemble des structures élémentaires de [G. Gross, 1996]. Nous avons retenu uniquement les structures qui, en

Emprunts	*casus belli, prima donna, pin-up, ilang-ilang*
Onomatopées	*bla-bla, gri-gri, frou-frou, tic-tac*
Lettres, abréviations, sigles	*a b c, b.a.-ba, T.S.F, K-octet*
Composés sur particule	*demi-litre, ci-devant, à-coup, sous-marin, en-avant*
Composés sur cardinaux	*trente et un, quatre-quatre, deux-roues, cinq à sept*
Composés sur thèmes savants	*oligo-élément, oto-rhino, surdi-mutité*
Composés sur verbes	*tire-bouchon, couche-tard, compte-gouttes*
Composés sur adjectifs	*clair-obscur, chaud et froid, fort en thème*
Composés $A+N$	*beau-frère, faux-sens, rouge-gorge, bon vivant*
Composés $N+A$	*table ronde, cordon-bleu, bouc-émissaire*
Composés $N+N$	*poche revolver, café-tabac, thé citron*
Composés $N+de+X$	*pomme de terre, mot de passe, règle de trois*
Composés $N+à+X$	*cuillère à soupe, poste à essence, machine à laver*
Composés $N+en+X$	*mise en scène, arc-en-ciel, fuite en avant*
Composés $N+PREP+X$	*service après-vente, preuve par neuf*
Composés $N+$divers	*traction avant, voix off, gilet pare-balles*
Composés sur phrases	*on-dit, sot-l'y-laisse, qu'en-dira-t-on*

Tableau 5.3 : Structures nominales complexes

première analyse, nous semblaient les plus productives ou qui participaient à la résolution de l'ambiguïté d'attachement adjectival et prépositionnel.

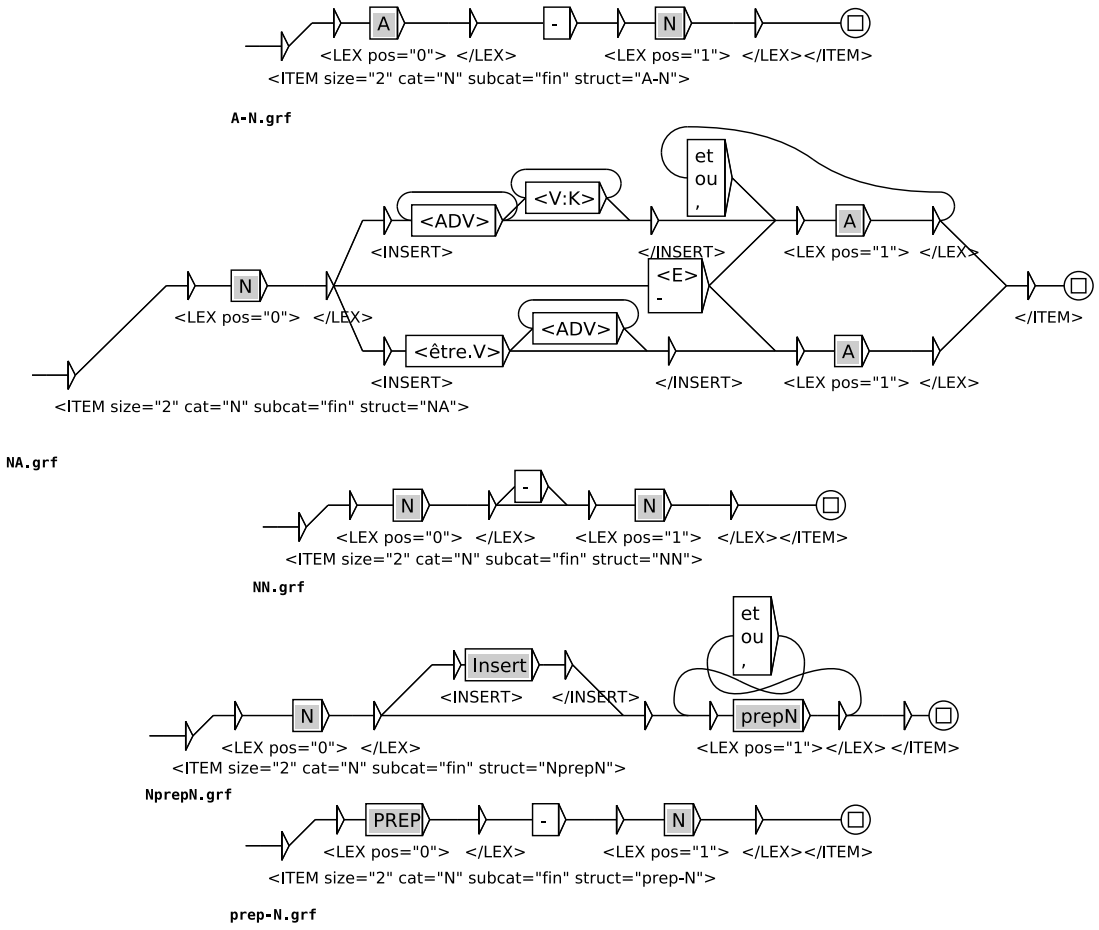


Figure 5.7 : Structures de bigrammes

La productivité des structures $A-N$ et $prep-N$ est très faible. Par ailleurs, leur empreinte fré-

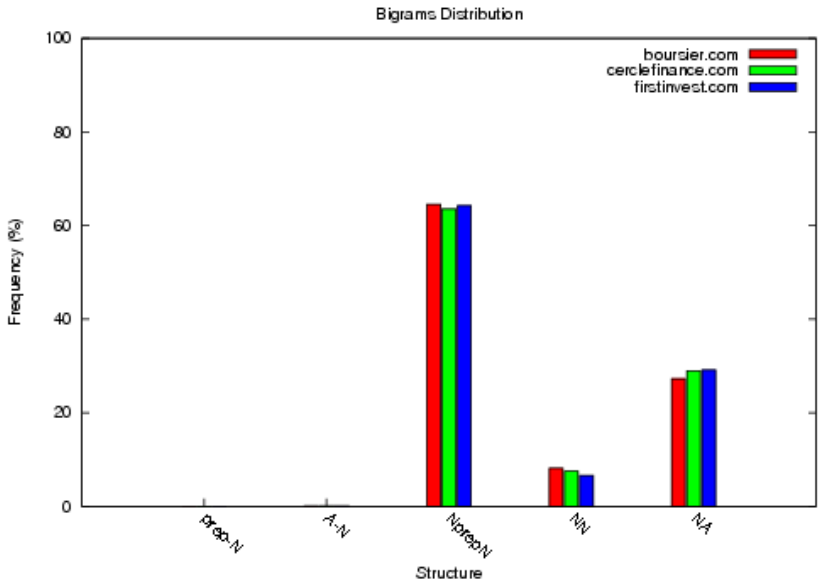


Figure 5.8 : Distribution fréquentielle (bigrammes)

quentielle n'est pas suffisamment significative la plupart du temps pour permettre la validation statistique. La liste que nous proposons au tableau 5.4 regroupe toutes les occurrences des deux structures et sépare les formes de *log-likelihood* égal et supérieur à zéro. Nous voyons là une confirmation des conclusions de l'évaluation de [Grefenstette, 1993] qui reproche aux méthodes « syntaxiques » de négliger les structures plus rares (*cf.* Section 5.2.4).

	$A-N$	$prep-N$
$LL == 0$	haut-fourneau plate-forme belle-famille gros-porteur blanc-seing	sur-performance sur-amortissement
$LL > 0$	pure-player petit-porteur petit-fils	

Tableau 5.4 : Exemples de structures $A-N$ et $prep-N$

NA et *NprepN* sont les structures les plus intéressantes. D'une part, leur productivité dépasse largement celle des autres structures (*cf.* Figure 5.8) et, d'autre part, ce sont elles qui nous permettent d'envisager une résolution d'une partie de l'ambiguïté d'attachement au niveau lexical. Nous reprenons, au tableau 5.6, quelques exemples des deux structures (classés par ordre décroissant de *log-likelihood*). Il nous faut remarquer que certains de ces exemples se retrouvent également dans le DELAC[7]. Toutefois, il s'agit d'un dictionnaire généraliste ; la majorité des structures que nous avons extraites sont trop spécialisées pour y être insérées.

[7]La proportion des structures au sein du DELAC rejoint celle que nous observons pour le langage financier. Les structures *NA* et *NprepN* représentent 71,5% du nombre total d'entrées. Toutefois, *NA* est largement majoritaire (43,5% du nombre total) contrairement aux résultats que nous obtenons au départ de notre procédure automatique

NprepN	*NA*
chiffre d'affaires	bénéfice net
conseil d'administration	perte nette
indice de confiance	communauté financière
transaction hors séance	exercice précédent
bénéfice d'exploitation	marge opérationnelle
groupe de service	résultat opérationnel
effet de change	production industrielle
hausse du prix	marge brute
exercice en cours	réserve fédérale
plan de restructuration	exercice fiscal

Tableau 5.5 : Exemples de structures NA et NprepN

Les composés *NN* sont les plus figés. Ils n'admettent pas l'insertion. Nous avons également pu constater qu'ils s'inscrivaient souvent dans une structure de taille supérieure. Cependant, il n'est pas impossible de les rencontrer sous leur forme binaire qui peut être considérée alors comme une forme abrégée.

NN	
groupe spécialiste	de la réservation de voyage
	de la maintenance électronique
	des solutions d'affranchissement
	des cartes à puce
groupe leader	en matière de télécommunication
	mondial dans la fabrication de microprocesseurs
	des produits de thérapie sanguine
emprunt obligataire	
équipementier télécoms	
activité service	
cash flow	
activité industrie	
activité distribution	en salle
	vidéo
	à valeur ajoutée
profit warning	
maison mère	

Tableau 5.6 : Exemples de structures NN

Plusieurs termes comme *maison mère* et *cash flow* apparaissent indifféremment, dans notre corpus, avec ou sans trait d'union. Il s'agit d'un phénomène de variation orthographique déjà mis en évidence par [Daille, 1995].

Structures de trigrammes

Notre définition des trigrammes repose sur l'addition d'un modifieur à un bigramme donné. En ce sens nous nous alignons sur [G. Gross, 1996] qui envisage huit classes de composés dérivées des structures binaires (*cf.* 5.7).

Nous n'avons pas pris en compte les structures dérivées de *AN*. Leur analyse syntaxique tri-

Expansions de $A + N$	*bon vieux temps, plus petit commun multiple*
Expansions de $N + A$	*produit national brut, résonance magnétique nucléaire*
Expansions de $N + N$	*opération ville morte, jambon assiette beurre*
Expansions de N + *de* + N	*offre publique d'achat, heure de grande écoute*
Expansions de N + *à* + N	*film à grand spectacle, œil au beurre noir*
Expansions de N + *en* + N	*requête en suspicion légitime*
Expansions de N + *Prép* + N	*acte sous seing privé*
Expansions diverses	*non-assistance à personne en danger*

Tableau 5.7 : Dérivés de structures complexes

viale, la forme <*DET*> $A* N$ ne posant pas de réels problèmes d'ambiguïté structurelle. Par conséquent, nous avons étudié les dérivés des structures *NA*, *NN* et $NprepN$. Précisons également que nous n'avons pas envisagé la structure dérivée $NAprepN$, prise en compte au niveau de l'insertion de la structure $NprepN$. D'un point de vue linguistique, ce choix peut paraître discutable mais l'incidence sur les résultats n'est pas significative. Nous pensons donc qu'il est préférable de limiter une potentielle redondance dans le lexique extrait. Ainsi, les exemples *produit national brut* et *offre publique d'achat* du tableau 5.7 seraient considérés comme des collocations de type *NA* et $NprepN$ respectivement.

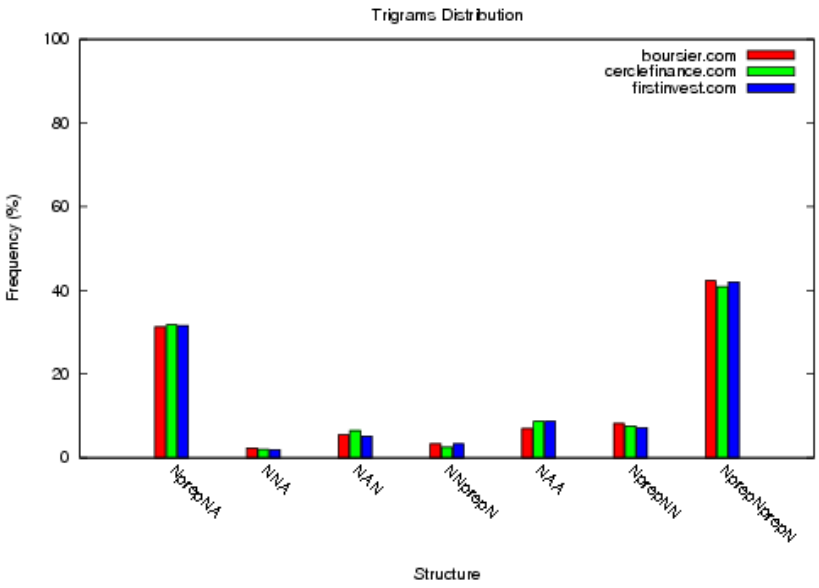


Figure 5.9 : Distribution fréquentielle (trigrammes)

Au vu de la distribution structurelle des bigrammes, nous ne sommes pas étonné de voir les structures $NprepNA$ et $NprepNprepN$ émerger largement au-dessus des autres (*cf.* Figure 5.9). Nous voyons là une nouvelle justification de notre idée d'un attachement au niveau lexical afin de réduire la combinatoire au niveau syntaxique.

(147) *NAA.* Les **marchés financiers européens** clôturent majoritairement dans le rouge mais cette nouvelle baisse est contenue.[8]

[8]Les remarques que nous pourrions faire à propos des trigrammes rejoignent globalement celles que nous avons

(148) *NAA.* Le Neurotin réalise un **chiffre d'affaires annuel** d'environ 1,6 milliard de dollars.

(149) *NAA.* Le **produit net bancaire** se contracte de 2,8% à 2,886 milliards d'euros.

(150) *NAN.* Le bénéfice net ressort à 192 millions, contre une **perte nette record** de 9,26 milliards d'euros, un an plutôt, consécutive en particulier à la dépréciation de ses investissements dans la téléphonie mobile.

(151) *NAN.* Plusieurs compagnies aériennes chinoises pourraient commander une trentaine d'avions au **constructeur américain Boeing**[9] pour un montant d'environ 2 milliards de dollars.

(152) *NAN.* Le commissaire européen à la concurrence, Mario Monti a annoncé l'échec des discussions engagées avec le **géant américain Microsoft** sur l'abus de position dominante reproché par Bruxelles au numéro un mondial du logiciel, et son intention de proposer mercredi prochain à la Commission d'infliger une amende au groupe américain.

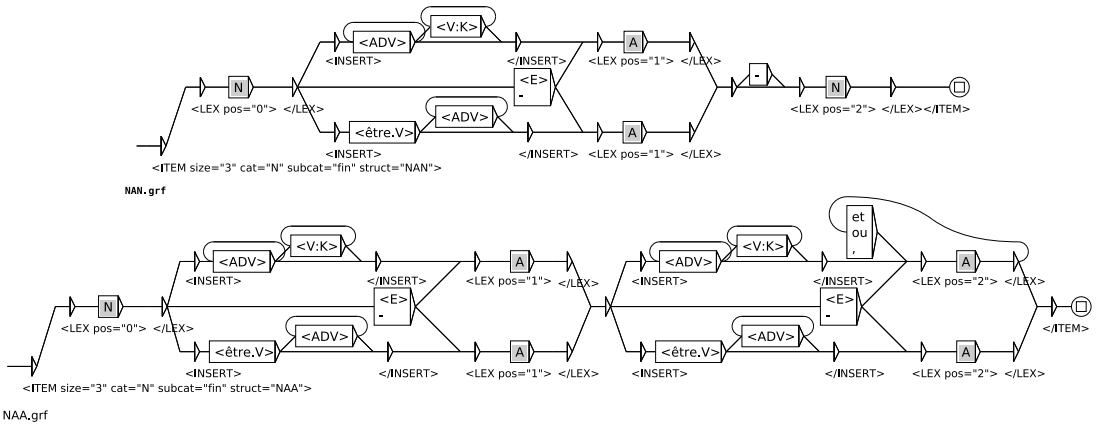


Figure 5.10 : Structures de trigrammes NAX (cf. Exemples 147 à 152)

(153) *NNA.* L'**opérateur télécoms allemand** a réalisé un bénéfice net de 850 millions d'euros, contre une perte de 1,8 milliard d'euros au premier trimestre 2002.

(154) *NNA.* A cette date, Stephan Newhouse sera le nouveau numéro deux de Morgan Stanley, assistant de près Philip Purcell, **président directeur général** de l'établissement.

proposées pour les bigrammes. Nous limitons donc notre présentation de ces structures à quelques exemples d'extraction afin de mettre en évidence l'avantage que confère l'étude des collocations. Ces exemples nous permettent en effet d'apprécier la simplification que les regroupements lexicaux apportent à la définition du syntagme nominal.

[9]Cette occurrence, de même que la suivante, ne sont pas du même ordre que des exemples tels que *offre publique d'achat* ou *chiffre d'affaires*. L'élément mis en avant ici est l'aspect fréquentiel prédominant, selon nous, dans l'établissement d'une collocation.

(155) *NNA***.** Cette performance résulte des marges opérationnelles toujours importantes dans la **division téléphonie mobile** et comprend une charge de restructuration de 350 à 400 millions d'euros dans le pôle des infrastructures de réseaux.

(156) *NNprepN***.** Les **banques conseils de l'opération** sont Morgan Stanley et Credit Suisse First Boston.

(157) *NNprepN***.** Le **sous-indice de l'emploi** s'est inscrit à 50,4 contre 48,5 un mois plus tôt, soit la première amélioration depuis le mois de juillet 2002.

(158) *NNprepN***.** Précisons que c'est la Saur, le **pôle services aux collectivités** de Bouygues, qui est le propriétaire de South East Water et qui a mandaté NM Rothschild pour examiner les options de vente totale ou partielle de cette entité.

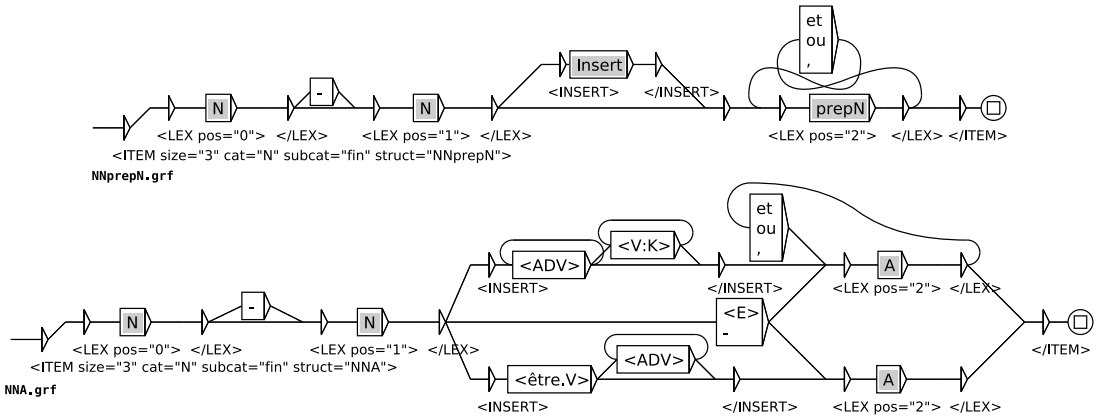


Figure 5.11 : Structures de trigrammes NNX (cf. Exemples 153 à 158)

(159) *NprepNA***.** Unilever (-7%) : le groupe néerlandais de grande distribution a révisé ses prévisions de **chiffre d'affaires annuel**.

(160) *NprepNA***.** Ce matin, les transactions hors séance préfigurent d'une **ouverture en territoire positif**.

(161) *NprepNA***.** Par ailleurs, les **commandes de biens durables** ont augmenté de 1,1% alors que celles de biens non durables se sont légèrement repliées de 0,2%.

(162) *NprepNN***.** Les résultats du deuxième trimestre 2003 de Boursorama reflètent la **hausse des marchés actions** depuis la crise irakienne.

(163) *NprepNN***.** Le rythme de **croissance de l'activité commerciale** est identique en Amérique latine.

(164) *NprepNN***.** Les **demandes d'allocations chômage** ont reculé de 4.000, à 384.000 contre 380.000 la semaine précédente (chiffre révisé).

(165) *NprepNprepN***.** Par ailleurs, le **chiffre d'affaires de la société** a reculé de 28,5% sur un an, à 209 millions de dollars.

(166) *NprepNprepN*. Le distributeur français de prêt à porter féminin, Etam Developpement, annonce pour le 3ème trimestre 2003, un recul de 4,4% de son **chiffre d'affaires hors taxe**, à 755,56 millions d'euros contre 790,66 millions d'euros un an plus tôt.

(167) *NprepNprepN*. La **baisse d'activité du groupe** s'explique donc en grande partie par un effet de périmètre mais on retrouve également un impact négatif de change (notamment chez Colas).

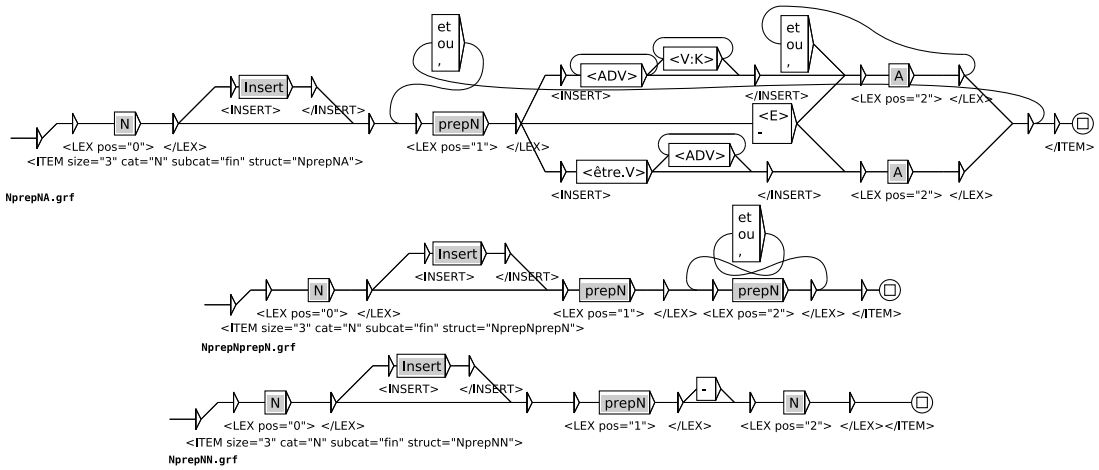


Figure 5.12 : Structures de trigrammes NprepNX (cf. Exemples 159 à 167)

5.2.6 Quelques résultats

Rappelons qu'il ne s'agit pas d'extraire des mots composés mais bien un ensemble de collocations propres à un domaine donné afin d'associer au final une, et une seule, catégorie grammaticale à un groupe de mots, potentiellement ambigus si nous les analysons de manière compositionnelle.

Quel corpus pour l'analyse ?

La pertinence d'un test statistique est en étroite relation avec la taille du corpus. Toutefois, les propriétés inhérentes au *log-likelihood* nous permettent d'utiliser des corpus de taille réduite.

> For text analysis and similar problems, the use of likelihood ratios leads to very much improved statistical results. The practical effect of this improvement is that statistical textual analysis can be done effectively with very much smaller volumes of text than is necessary for conventional tests based on assumed normal distributions, and it allows comparisons to be made between the significance of the occurrences of both rare and common phenomenon. [Dunning, 1993 : p. 6]

Nous pouvons donc envisager deux cas de figures différents. Travailler sur un gros corpus rend possible l'élaboration d'une base lexicale importante, un dictionnaire de collocations couvrant une grande partie du sous-langage traité (*cf.* Section 5.2.7). D'un autre côté, nous pourrions compléter cette ressource en soumettant à l'analyse fréquentielle chaque texte traité, pour autant qu'il ne s'agisse pas d'une simple dépêche ([Dunning, 1993] utilise un corpus de 31 777 mots).

De notre côté, nous avons défini deux types de corpus : un corpus de spécialité, cible de l'analyse, et un corpus de référence, pour estimer l'appartenance au domaine des termes extraits. Notre corpus de spécialité rassemble 10 000 dépêches du corpus `firstinvest.com` (*cf.* Section 4.4.1). Nous avons également appliqué notre procédure d'extraction à l'intégralité des corpus `boursier.com` et `cerclefinance.com` afin d'estimer plus facilement la couverture du dictionnaire créé au départ de `firstinvest.com` (*cf.* Tableau 5.2.6). Le corpus de référence rassemble le premier tiers de l'année 2 000 du journal *Le Soir*, l'un des principaux quotidiens francophones belges (*cf.* Tableau 5.8).

	MOTS
Corpus de spécialité	**1 939 424**
Corpus de référence	**6 308 625**

Tableau 5.8 : Corpus d'évaluation (extraction terminologique)

Quelle précision ?

Pour évaluer la précision de notre extracteur, nous avons passé en revue les 8 000 premières entrées de bigrammes et de trigrammes ce qui, par rapport à la totalité des structures identifiées, équivaut à 22,5% des bigrammes différents (c'est-à-dire les *types*[10]) et à 34% des trigrammes différents (*cf.* 5.9).

	Tokens	*Types*	*Ratio*
Bigrammes	115 424	35 574	**30,8%**
Trigrammes	38 842	23 647	**60,9%**

Tableau 5.9 : Tokens et types

La précision de la procédure est directement liée à l'introduction d'un seuil fréquentiel. Au plus nous abaissons ce seuil (c'est-à-dire la valeur de *log-likelihood*), au plus large sera la couverture. Il est évident que l'augmentation de la couverture s'accompagne également d'une diminution de la pertinence globale des résultats.

[10]Pour un corpus donné, les *tokens* désignent l'ensemble des formes identifiées (ici, les bigrammes et trigrammes) et, les *types*, l'ensemble des formes différentes. Le *ratio* des deux nous donne quant à lui une bonne idée de la diversité lexicale d'un corpus.

La figure 5.13 nous montre la fréquence relative cumulée des éléments extraits. Les éléments sont classés par ordre décroissant de *log-likelihood*. Nous constatons donc que les extractions les plus pertinentes sont également les plus fréquentes. Il n'y a rien de réellement surprenant dans ce constat mais nous voyons que les 8 000 premières entrées couvrent déjà 65,5% des co-occurrences totales de bigrammes (c'est-à-dire les *tokens*) et 46,56% des co-occurrences de trigrammes. Appréciées en termes de valeur de *log-likelihood*, ces entrées correspondent à un seuil de 5,32 et 1,6 respectivement. Cette différence s'explique assez facilement si l'on considère le tableau des fréquences cumulées. Nous constatons, en effet, une plus grande diversité au niveau des trigrammes, diversité qui se traduit par une plus faible stabilité statistique. Cette diversité s'observe plus facilement au niveau du ratio des *types* et *tokens* présenté dans le tableau 5.9.

	Précision
Bigrammes	91,63%
Trigrammes	84,93%

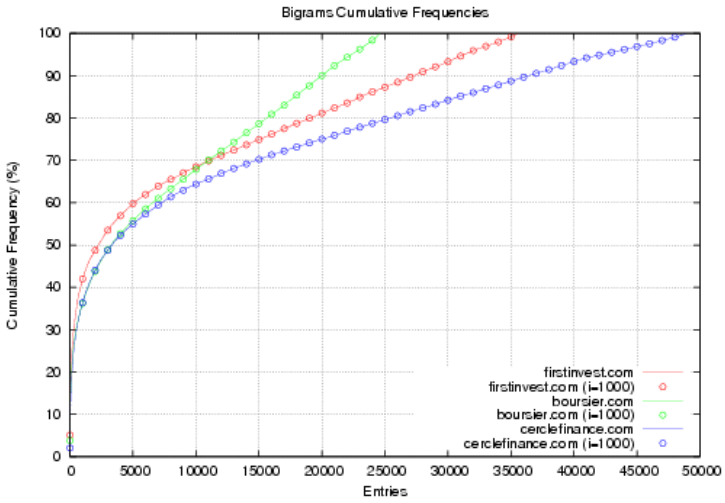
Tableau 5.10 : Précision

En termes de précision, les 8 000 entrées considérées offrent 91,63% de pertinence pour les bigrammes et 84,93% pour les trigrammes. Ces résultats, particulièrement intéressants, s'expliquent, selon nous, parce que nous travaillons dans le cadre d'un sous-langage dont une des propriétés est la fermeture aux niveaux lexical et syntaxique (*cf.* Chapitre 4). Par conséquent, en nous positionnant dans un sous-langage particulier, nous contraignons davantage le score statistique (mesure de récurrence).

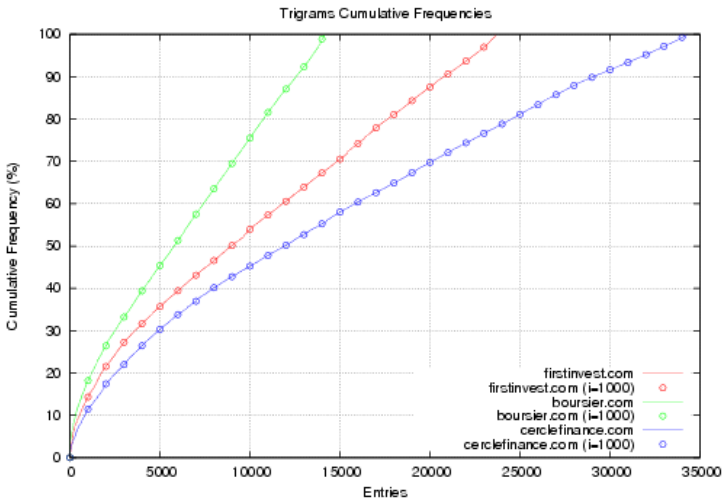
		BIGRAMMES		TRIGRAMMES	
Erreurs internes	Coordination	9,375%	47,76%	11,54%	46,02%
	Attachement	43,75%		51,92%	
	Incomplet	37,5%		34,62%	
	Insertion	9,375%		1,92%	
Erreurs externes	Etiquetage	20%	52,24%	8,2%	53,98%
	Corpus	80%		91,8%	

Tableau 5.11 : Répartition des erreurs internes et externes

Les erreurs que nous avons pu observer sont majoritairement dues à des phénomènes extérieurs à l'estimation fréquentielle qui représentent 52,24% du nombre total d'erreurs pour les bigrammes et 53,98% pour les trigrammes. Ces erreurs se répartissent en deux catégories selon qu'elles ressortissent à la qualité graphique du corpus (*cf.* Exemples 168 et 169) ou à l'étiquetage préalable à l'extraction (*cf.* Exemples 170 et 171). Les erreurs dues au *TreeTagger* sont très nombreuses et représentent 80% des erreurs extérieures à l'estimation fréquentielle pour les bigrammes et 91,8% pour les trigrammes.



(a) Bigrammes



(b) Trigrammes

Figure 5.13 : Fréquences cumulées

(168) *NN*. Celle-ci est le résultat du ralentissement macro-économique observé en Europe et qui n'est que partiellement compensé par le **maintient** de l'activité aux Etats-Unis (53% du chiffre d'affaires) et dans la **zone zone** Pacifique-Asie.

(169) *NprepNprepN*. Notre **avisFace** au succès de l'opération et au prix satisfaisant du nominal, nous maintenons notre position conserver.

(170) *NN*. Dans le même temps, les Activités diversifiées voient leurs ventes {reculer, reculer. N} de 13,6%, à 1,9 million d'euros.

(171) *NAN*. De plus, la marge brute {ressort, ressort. N} à son plus haut niveau historique, avec 50,4%.

Les erreurs imputables à la procédure d'extraction proprement dite relèvent majoritairement d'un mauvais attachement et représentent 43,75% des erreurs *internes* pour les bigrammes et 51,92% pour les trigrammes. Cette observation vient donc modérer notre hypothèse de départ, même si les performances restent réellement intéressantes et justifient pleinement, selon nous, l'idée d'un attachement au niveau lexical (*cf.* Exemples 172 et 173). Les erreurs restantes mettent en cause soit des structures de bigrammes ou trigrammes incomplètes (*cf.* Exemples 174 et 175), soit une méprise quant à une potentielle coordination (*cf.* Exemples 176 et 177), soit, finalement, certains modifieurs (*cf.* 178 et 179).

(172) *NprepN***.** MSN Premium, qui est encore testé, sera lancé officiellement avant la fin de l'**année au prix** de 9,95 dollars par mois, soit environ 9,2 euros.

(173) *NprepNN***.** Le chiffre d'affaires a été principalement soutenu par la progression de 14% sur un **an de l'activité édition** de logiciels.

(174) *NprepN***.** Saint-Gobain (-1,2%) : le groupe annonce la création d'une société commune détenue à parité avec l'américain Owens Corning pour construire une **usine de fils** de verre pour le renforcement.

(175) *NprepNA***.** Thomson (-4%) : le **groupe d'électronique grand** public a publié un chiffre d'affaires de 8,46 milliards d'euros en 2003, en baisse de 6,5% à taux de change constant, et un résultat net de 26 millions d'euros, contre 373 millions en 2002.

(176) *NprepN***.** Le groupe table sur une croissance de l'ordre de 10% de **son bénéfice** net par action et **de son bénéfice** d'exploitation pour l'ensemble de l'exercice 2004 et sur une hausse de 4% environ de ses ventes, hors effets de change.

(177) *NprepNprepN***.** Cette déprime des ménages reflète leurs craintes quant à l'**évolution du marché** du travail et **de la situation** économique en général.

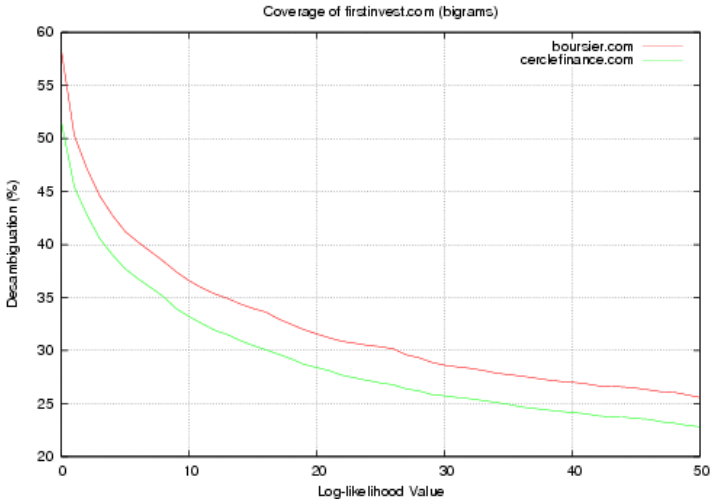
(178) *NprepN***.** La **manie <INSERT> aiguë associée </INSERT> au trouble** bipolaire est une maladie mentale qui touche 3 à 4% de la population adulte.

(179) *NprepNprepN***.** Une cour de justice américaine a confirmé la décision de premier appel de juger invalides les brevets de l'Augmentin, l'**antibiotique <INSERT> mis </INSERT> au point par GlaxoSmithKline** (GSK).

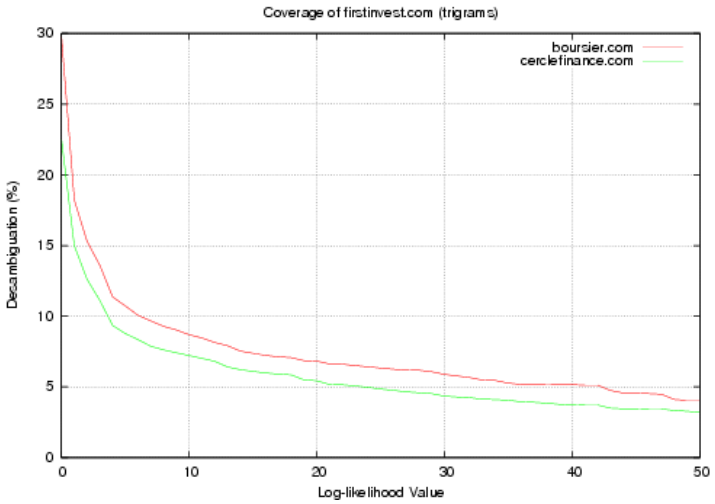
Quelle couverture ?

Pour juger la couverture des entrées extraites automatiquement, nous avons calculé l'intersection des listes extraites à partir de `firstinvest.com`, d'une part, et de `boursier.com` et `cerclefinance.com`, d'autre part. Notons que ces deux dernières listes comprennent l'en-

semble des structures décrites par nos filtres linguistiques et non pas uniquement les collocations pertinentes. Cela n'est pas innocent. Nous voulons avant tout estimer ce qu'implique notre extraction en termes d'attachements (prépositionnel et adjectival) au niveau lexical.



(a) Bigrammes



(b) Trigrammes

Figure 5.14 : Couverture lexicale

À l'inverse de la précision, la couverture lexicale diminue au fur et à mesure que le seuil statistique augmente. Toutes structures confondues, les seuils qui délimitent nos 8 000 entrées – 5,32 pour les bigrammes et 1,6 pour les trigrammes – correspondent respectivement à une couverture de 41,2% et 18,22% pour le corpus `boursier.com` et de 37,69% et 14,98% pour le corpus `cerclefinance.com`, nettement plus important en taille (*cf.* Figure 5.14).

Notons, pour conclure, que le *log-likelihood* permet, dans une certaine mesure, de circonstancier la limite évoquée par [Grefenstette, 1993] pour les méthodes linguistiques (*cf.* 5.2.4). S'il est évident que les structures les plus rares sont généralement associées à un score statistique moindre, nous constatons néanmoins que les structures dont la fréquence absolue (f) est inférieure à deux unités représentent 44,85% des 8 000 bigrammes considérés et 82,76% des 8 000 trigrammes (*cf.* Tableau 5.12).

	$f = 1$	$f = 2$	$f > 2$
Bigrammes	2175	1414	4411
Trigrammes	5578	1040	1382

Tableau 5.12 : Répartition fréquentielle

5.2.7 Un dictionnaire de collocations

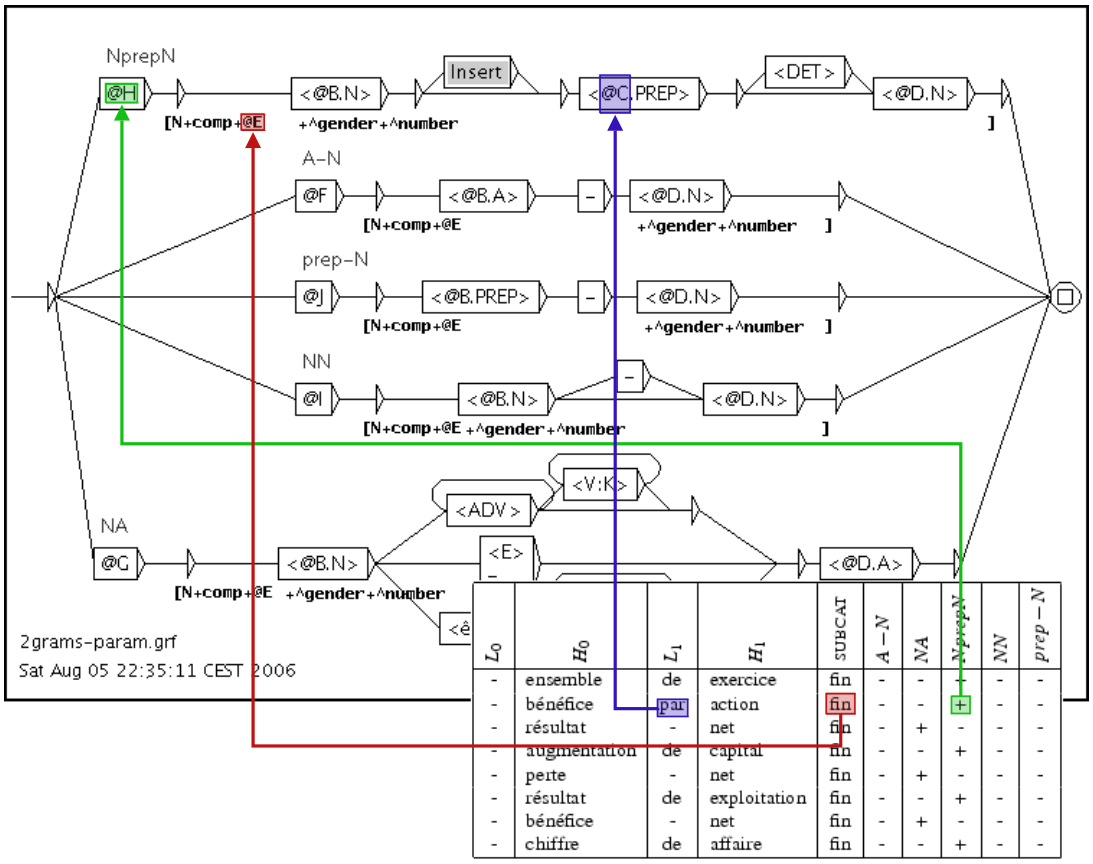
L'extraction terminologique que nous venons de détailler n'est pas une fin en soi. Nous visons avant tout l'amélioration de l'analyse lexicale. Dès lors, se pose la question de la réutilisation des ressources. Pour garantir cette réutilisation, nous devons impérativement valider deux conditions :

- **généralisation**, afin de prendre en compte d'éventuelles variations (modifieurs) ;
- **automatisation**, pour permettre l'inclusion du processus au sein d'une chaîne de traitement complexe et non supervisée.

En sortie du processus d'extraction, nous obtenons une suite de *ngrams* formalisés au sein d'une matrice lexico-syntaxique (*cf.* Figure 5.15). D'un point de vue formel, cette matrice s'apparente à une table du lexique-grammaire : chaque ligne correspond à une entrée lexicale complexe et chaque colonne représente une propriété lexico-syntaxique (*e.g.* la structure syntaxique du collocat et son trait de sous-catégorisation). Aux intersections ligne-colonne, un signe $+$, respectivement $-$, signifie que l'entrée actualise, respectivement n'actualise pas, la structure correspondante.

Nous exploitons le contenu des tables au moyen de *graphes paramétrés* (*cf.* Section 3.3). Rappelons que ces graphes formalisent les constructions linguistiques pertinentes pour une table donnée. Chaque construction, c'est-à-dire les structures des différents collocats, correspond à un parcours du graphe (*cf.* Figure 5.15). Tout comme pour les filtres que nous utilisons lors de l'extraction, nous conservons les points d'insertion.

Dans cette partie de notre travail, nous n'utilisons plus *Unitex* afin de procéder aux analyses lexicale et syntaxique mais *Outilex* (*cf.* [Blanc *et al.*, 2006]). Cet analyseur repose, nous l'avons précisé au chapitre 1, sur un formalisme similaire au premier mais offre un plus grand contrôle



L_0	H_0	L_1	H_1	SUBCAT	$A-N$	NA	$NprepN$	NN	$prep-N$
-	ensemble	de	exercice	fin	-	-	+	-	-
-	bénéfice	par	action	fin	-	-	+	-	-
-	résultat	-	net	fin	-	+	-	-	-
-	augmentation	de	capital	fin	-	-	+	-	-
-	perte	-	net	fin	-	+	-	-	-
-	résultat	de	exploitation	fin	-	-	+	-	-
-	bénéfice	-	net	fin	-	+	-	-	-
-	chiffre	de	affaire	fin	-	-	+	-	-

Figure 5.15 : Graphe paramétré

des données linguistiques. En effet, il introduit des outils comme l'unification et la pondération qui nous permettent de contraindre davantage nos automates de reconnaissance et d'éviter ainsi une partie de l'explosion combinatoire que nous évoquions au chapitre 3. De plus, il nous permet de réaliser des regroupements au niveau lexical (commande *decore-fsa*). C'est précisément ce dernier outil que nous utilisons afin d'appliquer nos collocations au texte. Quand une structure est reconnue, l'automate du texte, qui, rappelons-le, décrit l'ensemble des analyses possibles pour un énoncé donné (c'est-à-dire toute l'ambiguïté de cet énoncé), est augmenté d'un chemin qui rassemble la séquence en un unique élément.

Afin de ne pas perdre toute l'information associée aux entrées qui composent la collocation, nous faisons remonter, au niveau de la structure complexe, le genre et le nombre de la tête lexicale en plus de la catégorie grammaticale. Par ailleurs, chaque structure est augmentée de deux propriétés distributionnelles. La première, $+comp$ (pour complexe), caractérise la structure de l'entrée. La seconde relève, quant à elle, du domaine de spécialité : trait $+fin$ pour la finance, $+med$ pour la médecine, *etc.*

La remontée de l'information flexionnelle (symbolisée par le caractère ˆ au sein du graphe paramétré) est une opération assez triviale. Nous définissons, pour chaque structure de collocation, une tête lexicale à laquelle nous associons la notation $+\hat{g}ender+\hat{n}umber$. Cette notation est in-

terprétée par la commande *decore-fsa* d'*Outilex* qui, en sortie, adjoindra à la structure reconnue la valeur des variables *gender* et *number* (c'est-à-dire le genre et le nombre de la tête lexicale, *cf.* Figure 5.15). Dans le cas de la structure *NN*, ce procédé relève plus d'une approximation que d'une procédure canonique. [Daille, 1995] recense deux types d'accord en nombre, selon qu'il agisse sur le premier élément (*e.g* génératrice diesel – génératrices diesel) ou sur les deux (*e.g* mémoire tampon – mémoires tampons). Par ailleurs, le genre des composés *NN* peut, dans de très rares cas, ne pas correspondre au genre du premier nom (*e.g.* auto-stoppeur, [Silberztein, 1993 : p. 75]). Toutefois, bien qu'il soit linguistiquement hasardeux, ce procédé ne pose pas de réels problèmes dans la pratique. Par contre, le cas de $prep - N$ est plus problématique. Son genre comme son nombre ne sont pas toujours prédictibles au départ de la tête nominale. Nous n'avons pas vraiment approfondi ce sujet. L'extraction de ces structures est très rare et leur validation statistique nulle.

La propriété $+comp$ est intégrée au graphe paramétré et peut être utilisée afin de contraindre les regroupements réalisés au niveau syntaxique. En effet, le formalisme d'*Outilex* nous permet de donner plus de poids à un chemin particulier. Ainsi, en pondérant les chemins actualisant une entrée marquée $+comp$, nous reléguons au second plan l'analyse compositionnelle sujette à ambiguïté.

La sous-catégorisation lexicale (*e.g.* $+fin$, $+med$, *etc.*) intervient dans notre grammaire d'extraction afin de contraindre le type des compléments de chaque constituant de la phrase élémentaire. De cette manière, nous nous assurons de l'homogénéité et de la pertinence des syntagmes. Cette technique nous permet de résoudre une part des problèmes d'attachement et d'éviter certaines analyses extérieures au domaine.

Notons que nous avons également envisagé la notion de sous-catégorisation au niveau des termes simples. La phrase élémentaire, considérée dans le contexte d'un sous-langage, nous permet de préciser l'information associée à chaque constituant essentiel. Par conséquent, en appliquant la *mesure de récurrence* aux substantifs des syntagmes isolés par l'automate d'extraction de la figure 3.6, nous sommes à même de préciser leurs traits sémantiques. La table 36DT, par exemple, nous permet, après estimation fréquentielle, d'associer le couple de traits $+fin+trans$ (pour *transaction financière*) aux termes *capital*, *groupe* et *actions* des exemples 180 et 181.

(180) <ESENT pred="acquérir"> Artemis a acquis 30,2% du <HEAD cat="N"> **{capital, capital.N}** </HEAD> du <HEAD cat="N"> **{groupe, groupe.N}** </HEAD> américain de bagagerie Samsonite. </ESENT>

(181) <ESENT pred="racheter"> Alcatel a racheté 9,45% des <HEAD cat="N"> **{actions, action.N}** </HEAD> de Thomson-CSF à l'Etat. </ESENT>

Cette solution rappelle le principe soutenant les classes d'objets que nous évoquions aux chapitres 1 et 4. Les premiers résultats que nous avons obtenus sont encourageants mais pas encore optimaux. Certains termes extraits relèvent, par exemple, d'une structure lexicale complexe du type de celles que nous venons d'étudier (*cf.* Exemples 180 : *groupe – groupe américain de bagagerie*). Or, pour l'instant, nous avons uniquement considéré les formes simples. Cette procédure est donc proposée comme perspective de développement et devra être complétée.

5.3 Entités nommées

La question des entités nommées en extraction d'information est un problème à part entière (dont la résolution est traitée séparément dans les conférences MUC et ACE) et a déjà donné lieu à de nombreux travaux (*e.g.* [Poibeau, 2003], [Fairon et Watrin, 2003], [Kevers, 2006]). Nous ne prétendons pas ici résoudre ce problème. Toutefois, le domaine de la finance regorge de noms propres, principalement des noms de société ou d'organisation, sous des formes diverses (abréviations, acronymes, *etc.*). Par conséquent, nous ne pouvions pas totalement évacuer le problème sans que cela ne se répercute sur les résultats de notre extracteur.

Nous avons envisagé la question à deux niveaux : la compilation d'un dictionnaire d'entreprises et la définition d'une structure lexicale générique.

5.3.1 Un dictionnaire d'entreprises

Notre dictionnaire des entreprises intègre deux sources différentes : les données de *Forbes*[11] – un site internet listant, entre autres, les 2000 entreprises les plus importantes au monde – et celles de *CercleFinance*[12] qui propose une rubrique au sein de laquelle les informations sont classées par société. Notre méthode d'acquisition est identique à celle utilisée pour la compilation de nos corpus ; des marqueurs récurrents au sein du code de la page nous permettent d'identifier la séquence textuelle à extraire[13] (*cf.* Figure 4.2). Nous obtenons, après rassemblement des sources et évacuation des doublons, un dictionnaire de 3445 entrées (*cf.* 5.13).

	Forbes	*CercleFinance*
ENTRÉES	2000	1810
TOTAL	**3445**	

Tableau 5.13 : Taille du dictionnaire entreprises.dic

Le format adopté est naturellement celui du DELA (*cf.* Tableau 5.13). Nous signalons la *sé-*

[11] http://www.forbes.com/lists/2006/18/06f2k_The-Forbes-2000_Rank.html
[12] http://www.cerclefinance.com/default.asp?pub=societes&rub=SECURITE
[13] L'extraction des données de Forbes a été réalisée par Michel Thomas, informaticien au Cental.

mantique du nom propre par la combinaison de traits $+PR+soc$, pour *nom propre* et *société*. Par conséquent, nous conservons la catégorie grammaticale nominale et nous pouvons accéder aux entrées du dictionnaire de façon générique, en spécifiant uniquement l'étiquette *N*, ou particulière, en associant à *N* les traits $+PR$ ou $+soc$ ou les deux. Il nous semble préférable, de manière générale, d'éviter la multiplication des catégories grammaticales génériques. En effet, nous aurions pu imaginer une catégorie du type *SOC* relevant du même niveau d'abstraction que *N* (nom), *A* (adjectif) ou *PRO* (pronom) mais cette méthode compliquerait nos automates d'extraction et empêcherait toute généralisation du nom d'entreprise à la catégorie nominale. Nous préférons considérer les informations *nom propre* et *société* comme une contrainte que nous activons si nécessaire afin, par exemple, de limiter les analyses associées à une suite syntaxique.

```
Unilog,.N+PR+soc
Union Bank of India,.N+PR+soc
Union Financière de France,.N+PR+soc
Union Pacific,.N+PR+soc
Unipol,.N+PR+soc
Uniqa,.N+PR+soc
Unisn Fenosa,.N+PR+soc
Unisys,.N+PR+soc
United Auto Group,.N+PR+soc
United Microelectronics,.N+PR+soc
United Mizrahi Bank,.N+PR+soc
United Overseas Bank,.N+PR+soc
United Parcel Service,.N+PR+soc
United Technologies,.N+PR+soc
United Utilities,.N+PR+soc
UnitedHealth Group,.N+PR+soc
Unitrin,.N+PR+soc
```

Tableau 5.14 : Dictionnaire des entreprises (extrait)

5.3.2 Une formalisation générique

Même très complet, un dictionnaire peut toujours contenir des lacunes. Il va sans dire que dans un monde aussi fluctuant que la finance, où les sociétés se créent, changent de nom ou de mains et disparaissent, nous ne pouvons reposer nos ressources sur un dictionnaire qui rend compte d'une situation inévitablement figée dans le temps. Par conséquent, nous utilisons, en parallèle à notre dictionnaire, une description approximative de la notion de nom propre, formalisée sous la forme d'un automate (au format *Outilex*[14]).

[14]Dans la figure 5.16, `<lex>` désigne une entrée lexicale quelconque (`<MOT>` dans *Unitex*). Associés à `lex`, les attributs `+cap` et `+maj` signifient respectivement une entrée *tout en majuscules* et *commençant par une majuscule*. La balise `<prep>` représente la préposition (`<PREP>`, dans *Unitex*) ; `<noun+pr>`, un nom propre (`<N+PR>`, dans *Unitex*) ; `<punc>`, un signe de ponctuation (`<PUNC>`, dans *Unitex*) ; et `<unknow>`, un mot inconnu des dictionnaires.

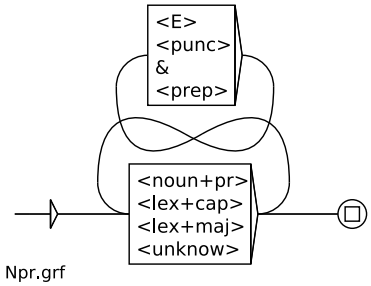


Figure 5.16 : Formalisation générique des noms propres

La figure 5.16 repose sur l'hypothèse suivante : dans un système à large couverture lexicale, les mots inconnus commençant par une majuscule sont des noms propres (*cf.* Tableau 5.15). [Maurel, 2004] circonstancie toutefois cette hypothèse et précise que «la seule approximation acceptable est de les reconnaître comme partie de nom propre» (p. 783).

```
<BNP,.unknow> <Paribas,.unknow>
<France,France.noun+top+f+s+pr-comp> <Télécom,.unknow>
<Serge,Serge.noun+hum+m+s+pr-comp> <Kampf,.unknow>
<Deutsche,.unknow> <Telekom,.unknow>
<AOL,.unknow> <Europe,Europe.noun+f+s+pr-comp>
<Blue,.unknow> <Circle,.unknow>
<Framatome,.unknow>
<Thomson,.unknow> - <CSF,.unknow>
```

Tableau 5.15 : Exemples de mots inconnus

Sur base de ces deux solutions, la formalisation générique d'un automate des noms propres et la création d'un dictionnaire d'entreprises, nous atteignons une couverture tout à fait acceptable[15]. Toutefois, il reste que la notion de nom propre est très complexe et dépasse ce que nous en disons. Un nom propre n'est pas uniquement un mot inconnu de même qu'il ne commence pas nécessairement par une majuscule. Certains noms propres se construisent à partir de noms communs (*e.g. le pain quotidien*, les *Cahiers du Cental*, l'*Organisation des Nations Unies*). Ces constructions, mises en évidence par [Maurel, 2004], présentent souvent une syntaxe régulière et peuvent donc être attachées au niveau syntaxique. Par ailleurs, notre extracteur de collocations nous permet d'en appréhender certaines au niveau lexical (*cf.* Tableau 5.16).

NA	Union européenne
NN	agence Reuters
NN	Galeries Lafayette
NAA	Agence spatiale européenne
NAN	compagnie néerlandaise KLM
NprepNN	indice du Conference Board

Tableau 5.16 : Noms propres complexes

[15] Nos tests n'ont pas montré d'erreurs imputables à un nom propre non reconnu mais nous n'avons pas étudié la couverture de manière spécifique.

5.4 Conclusion

Nous avions suggéré au chapitre 4 la nécessaire spécialisation au domaine traité des connaissances linguistiques. Au niveau lexical, nous avons envisagé cette spécialisation de deux manières différentes :

– l'augmentation des traits sémantiques de sous-catégorisation et
– l'apprentissage de structures nominales complexes.

Il résulte de notre étude la mise au point d'une procédure opérationnelle d'apprentissage lexical couvrant ces deux articulations et nous permettant d'envisager la création d'un lexique de spécialité. Cette solution obtient de bons résultats principalement du point de vue de la résolution des ambiguïtés d'attachements prépositionnel et adjectival particulièrement problématiques. Afin d'exploiter les lexiques extraits, nous avons mis au point un formalisme s'inspirant largement des graphes paramétrés. Les résultats peuvent donc être très facilement injectés au processus d'analyse syntaxique. Nous présentons, ci-dessous, un exemple de texte dans lequel nous avons mis en gras les noms propres et souligné les structures lexicales complexes. Il s'agit des entrées effectivement reconnues. Par ailleurs, nous avons limité notre soulignement aux collocations appartenant aux 8 000 entrées que nous avons revues et corrigées. D'autres auraient sans doute pu être théoriquement prises en compte mais nous ne les avons pas identifiées (automatiquement).

> **Microsoft** : son activité internet bénéficiaire pour la première fois
>
> Le leader mondial des logiciels, le groupe américain Microsoft, a indiqué que sa division internet **MSN** a été bénéficiaire au troisième trimestre 2003, et ce pour la première fois depuis sa création il y a de cela huit ans. Le bénéfice net de **MSN** s'élève à 58 millions de dollars, contre une perte de 147 millions il y a un an. L'activité de la division a été dopée par le bond des recettes publicitaires (+50%), qui ont entraîné une hausse du chiffre d'affaires à 128 millions de dollars. D'autre part, **Microsoft** a indiqué que dans l'ensemble de ses sept divisions, quatre ont dégagé un résultat d'exploitation positif. La division Windows a réalisé un bénéfice d'exploitation de 2,26 milliards pour un chiffre d'affaires de 2,81 milliards. La division Office a dégagé un bénéfice de 1,59 milliard pour 2,29 milliards de ventes et la division serveurs, 370 millions pour 1,87 milliard de chiffre d'affaires. Les divisions **Business** solutions, des logiciels pour téléphonie mobile et "**Home Entertainment**" ont toutes subi des pertes. De ce fait, en incluant d'autre perte d'exploitation de 751 millions de dollars, **Microsoft** a dégagé un bénéfice d'exploitation total de 3,15 milliards de dollars.

Par ailleurs, notre étude d'un sous-langage tel que celui de la médecine interne, très différent de la finance, montre que notre procédure n'est pas *ad hoc* et est applicable à tout domaine.

Finalement, nous avons confirmé qu'une solution hybride mêlant estimation fréquentielle et définitions syntaxiques apporte une réelle plus-value au processus d'analyse lexicale, du moins dans le cas d'un sous-langage. Nous ne croyons pas qu'il serait possible d'établir un lexique financier ou médical sans recourir à une batterie de spécialistes du domaine. En ce sens, nous

rencontrons les impératifs d'adaptabilité d'un moteur d'extraction.

CHAPITRE 6

ANALYSE SYNTAXIQUE DE SURFACE

6.1 Introduction

Dans un premier temps, nous aborderons le principe de segmentation des énoncés en *chunks* (*i.e.* en tronçons syntaxiques non récursifs). Nous empruntons cette notion à [Abney, 1991]. L'idée que nous poursuivons est, rappelons-le, la réduction de la complexité syntagmatique. Il s'agit de restreindre le nombre d'éléments grammaticaux en effectuant certains regroupements au départ de règles purement syntaxiques[1]. De cette manière, nous travaillons au niveau de la phrase élémentaire avec un ensemble d'éléments plus facilement manipulable et moins ambigu.

Parallèlement, cette étape nous permet de séparer les informations syntaxiques générale et spécialisée. Les chunks forment en effet des unités indépendantes du contexte d'énonciation. Il s'agit de structures syntaxiques minimales, généralisables à la langue dans son ensemble et, par conséquent, réutilisables. Au contraire, la procédure d'attachement gagne à être envisagée dans un contexte particulier (le sous-langage). C'est ce que nous suggérions en conclusion du chapitre 3 et c'est en ce sens que nous l'aborderons ici.

Nous avons déjà souligné les implications de la spécialisation du sous-langage au niveau lexical, en introduisant la notion de collocation qui correspond souvent à une convention lexicale (implicite ou explicite) propre à un domaine. Dans ce chapitre, nous développerons davantage les conséquences syntaxiques de ce phénomène. Nous envisagerons les différentes caractéristiques inhérentes à la spécialisation linguistique des sous-langages et les perspectives que leur prise en compte offre au niveau de l'attachement syntagmatique. Nous aborderons certains éléments transversaux, déjà mis en évidence par [Kittredge, 1982] et [Lehrberger, 1982], mais nous limiterons plus généralement nos considérations à la table 36DT. Cette table, bien que

[1]Une grammaire de chunks est exclusivement syntaxique. Nous entendons par là qu'aucun élément n'est lexicalisé. La lexicalisation intervient lors de l'attachement en syntagmes (prédicats, prépositions, *etc.*). Nous verrons, cependant, que nous nous écartons partiellement de cette règle afin de prendre en compte certains adverbes qui présentent une structure de chunk nominale, prépositionnelle ou verbale et peuvent compliquer inutilement l'attachement.

très productive dans le domaine des transactions financières, ne suffit pas à formaliser le sous-langage. Ceci étant, notre objectif, dans ce chapitre, n'est pas la présentation d'un sous-langage mais la mise en évidence des perspectives qu'offre la spécialisation linguistique par rapport à la définition standard. En ce sens, la table 36DT donne un excellent point de comparaison.

6.2 Segmentation en syntagmes non récursifs

La procédure de segmentation en chunks repose sur une **cascade de transducteurs**, c'est-à-dire un ensemble d'automates à états finis augmentés d'une **sortie** et organisés en différents **niveaux successifs** de raffinement syntaxique. Cette idée a déjà fait son chemin dans le traitement automatique des langues. De très nombreux articles sont consacrés à cette procédure dont les principaux sont [Abney, 1991], [Abney, 1996], [Federici *et al.*, 1996] et [Ait-Mokhtar et Chanod, 1997]. Notons également que [Appelt *et al.*, 1993] exploite, avec succès, ce principe dans le cadre de l'extraction d'information, afin d'effectuer une analyse de surface des textes (*cf.* Chapitre 2).

NIVEAU	EXEMPLE
Énoncé	Le groupe de télécommunications néerlandais KPN a annoncé avoir acquis une participation de 77,5% dans le troisième opérateur allemand de téléphonie mobile E-Plus.
Lexical	Le **[**$_N$ groupe de télécommunications **]** néerlandais KPN a annoncé avoir acquis une participation de 77,5% dans le troisième **[**$_N$ opérateur allemand de téléphonie mobile **]** E-Plus.
Chunk	Le **[**$_N$ groupe de télécommunications **] [**$_{XA}$ néerlandais **]** KPN a annoncé **[**$_{XVI}$ avoir acquis **]** une participation de 77,5% dans le **[**$_{XA}$ troisième **] [**$_N$ opérateur allemand de téléphonie mobile **]** E-Plus.
	[$_{XN}$ Le groupe de télécommunications **] [**$_{XA}$ néerlandais **] [**$_{XN}$ KPN **]** a annoncé **[**$_{XVI}$ avoir acquis **] [**$_{XN}$ une participation **]** de **[**$_{XN}$ 77,5% **]** dans **[**$_{XN}$ le troisième opérateur allemand de téléphonie mobile E-Plus **]**.
	[$_{XN}$ Le groupe de télécommunications **] [**$_{XA}$ néerlandais **] [**$_{XN}$ KPN **] [**$_{XV}$ a annoncé avoir acquis **] [**$_{XN}$ une participation **] [**$_{XP}$ de 77,5% **] [**$_{XP}$ dans le troisième opérateur allemand de téléphonie mobile E-Plus **]**.
Phrase	**[**$_{N0}$ Le groupe de télécommunications néerlandais KPN **] [**$_V$ a annoncé avoir acquis **] [**$_{N1}$ une participation de 77,5% dans le troisième opérateur allemand de téléphonie mobile E-Plus **]**.

Tableau 6.1 : Processus d'analyse

Nous illustrons, dans le tableau 6.1, le processus que ce concept suggère. Pratiquement, nous

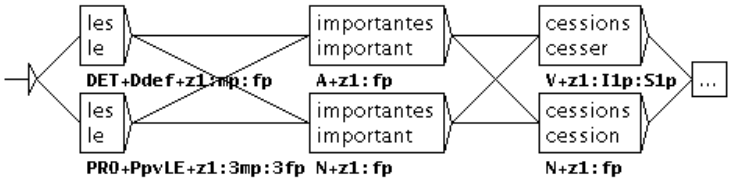
appliquons notre grammaire sur l'automate du texte (*cf.* niveau `Chunk` du tableau 6.1) après avoir effectué l'attachement des structures lexicales complexes propres au domaine (*i.e.* les collocations, *cf.* niveau `Lexical` du tableau 6.1). Nous utilisons pour cela la fonction *decorefsa* d'*Outilex* à laquelle nous associons les trois options suivantes :

1. L'heuristique du plus long motif. Nous retenons, comme pertinentes, uniquement les séquences les plus longues (option *-longest-match*).

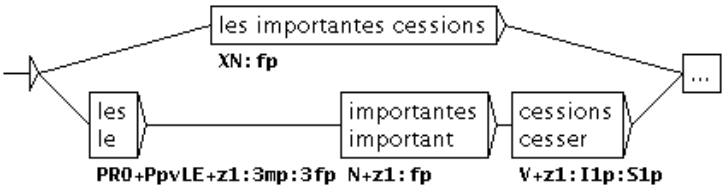
2. L'évacuation de la redondance. Si, pour un chunk donné, nous obtenons différentes reconnaissances qui présentent les mêmes frontières initiale et finale, mais qui diffèrent par leur chemin (*i.e.* les transitions empruntées pour aller de l'état initial à l'état final), nous en conservons une seule (option *-ipath*).

3. Le mode «remplacement». Chaque transition ayant mené à l'identification d'un chunk est supprimée de l'automate du texte (option *-r*).

Nous obtenons, en sortie, un automate du texte partiellement désambiguïsé et augmenté de nouvelles transitions, les chunks. L'étape suivante, l'analyse syntaxique proprement dite, utilise ces chunks afin de reconstruire les structures élémentaires, axiomes de notre grammaire d'extraction (*cf.* niveau `Phrase` du tableau 6.1).



(a) Avant segmentation



(b) Après segmentation

Figure 6.1 : Segmentation en chunks

Chunks et ambiguïté

Le chunk définit une unité syntaxique minimale introduite par certains mots grammaticaux et terminée par la tête de la structure (*cf.* [Ait-Mokhtar et Chanod, 1997]). L'ambiguïté d'attache-

ment (prépositionnel ou adjectival) est donc évacuée et traitée ultérieurement.

Considérons, par exemple, la définition prototypique du chunk nominal :

$$XN \quad \rightarrow \quad DET? \quad A* \quad N$$

Considérons ensuite la séquence : *les importantes cessions*, dont la complexité est illustrée par la figure 6.1(a). En appliquant notre grammaire à cette séquence selon l'heuristique du plus long motif, nous pouvons remplacer la suite `{les, le. DET+Ddef} {importantes, important. A} {cessions, cession. N}` par XN. De cette manière, nous évacuons, en amont de l'analyse syntaxique, une part non négligeable de la complexité lexicale en réduisant les six analyses de départ à deux analyses effectives[2] (*cf.* Figure 6.1(b)). Précisons encore que nous travaillons, au niveau syntagmatique, au départ des chunks uniquement. Dès lors, les unités lexicales qui n'ont pu être rassemblées en chunks sont ignorées (à l'exception des conjonctions de coordination et de subordination et des pronoms relatifs qui ne sont pas intégrés au formalisme).

Chunks et étiquetage

Chaque chunk est associé à une étiquette qui se substitue à la séquence reconnue au sein de l'automate du texte et nous permet de travailler, au niveau syntagmatique, au départ d'un ensemble plus restreint et donc moins complexe de catégories grammaticales. La nouvelle unité hérite également des propriétés flexionnelles et lexicales de sa tête. C'est en ce sens que nous inscrivons le processus de segmentation en chunks dans la continuité de l'étiquetage au même titre que la reconnaissance des collocations. Nous reprenons dans le tableau 6.2, les différents chunks que nous avons définis de même que les traits qui leur sont associés.

TYPE	CHUNK	TRAITS
`Adverbial`	$XADV$	–
`Adjectival`	XA	genre et nombre
`Nominal`	XN	genre, nombre et sous-catégorisation
`Prépositionnel`	XP	genre, nombre et sous-catégorisation
`Verbal`	$XV(P)$	genre, nombre, mode et personne
	$XVI(P)$	mode (genre et nombre)
	$XVK(P)$	genre, nombre et mode
	$XVG(P)$	mode (genre et nombre)

Tableau 6.2 : Types de chunks

[2]Ces regroupements peuvent, bien sûr, être incorrects (de la même manière qu'ils le seraient dans un contexte d'analyse profonde). Toutefois, les chunks, parce que non récursifs, sont nettement moins ambigus. Par ailleurs, nous conservons l'ambiguïté lexicale. Ce qui signifie en pratique qu'une même séquence graphique peut conduire à plusieurs regroupements différents.

Chunks et performance

La principale plus-value en termes de performance est la rapidité d'exécution. Les structures d'un niveau donné de la cascade se définissent uniquement sur base des structures des niveaux précédents. En d'autres mots, aucun chunk ne repose sur la définition d'un chunk du même niveau ou d'un niveau supérieur. La récursion est donc totalement évacuée au profit d'une procédure itérative, plus rapide (*cf.* [Abney, 1991], pour une évaluation chiffrée).

Nous avons, en outre, évoqué une potentielle réduction de l'ambiguïté lexicale. Les résultats que nous obtenons viennent confirmer cette propriété. L'application des dictionnaires DELA et DELACF sur notre corpus d'évaluation de 1 761 phrases (*cf.* Section 3.4) engendre une complexité lexicale qui, considérée du point de vue de l'automate du texte, représente 74 538 états pour 182 203 transitions, c'est-à-dire une moyenne de 36,9 états et 90,2 transitions par phrase. Une fois la segmentation effectuée, nous obtenons un nouvel automate du texte qui compte cette fois 54 709 états et 113 056 transitions (27,1 états et 56 transitions par phrase) soit une réduction de 26,6% du nombre d'états et de 37,9% du nombre de transitions[3].

6.2.1 Typologie des structures

Un chunk s'articule autour d'un élément tête. Cet élément tête est le dernier élément de sa définition formelle et forme donc sa frontière *droite*. Ainsi, un chunk adverbial s'arrêtera à l'adverbe, un chunk adjectival à l'adjectif, *etc.*

Notre cascade définit douze nouvelles catégories réparties en quatre niveaux de raffinement successifs. Nous résumons ces informations dans le tableau 6.3. Nous abordons, dans les points suivants, la définition formelle de ces unités que nous regroupons en trois catégories : les **modifieurs**, les **structures nominales** et les **structures verbales**.

Niveau	Chunk
1	$XADV$
2	XA $XVK(P)$ $XVI(P)$
3	XN $XVG(P)$
4	XP $XV(P)$

Tableau 6.3 : Étapes de la segmentation en chunks

[3]Notons que nous n'éliminons pas, de ce résultat, les entrées qui, parce qu'elles ne s'inscrivent dans aucun chunk, ne seront pas prises en compte au niveau de l'analyse syntaxique.

Les modifieurs

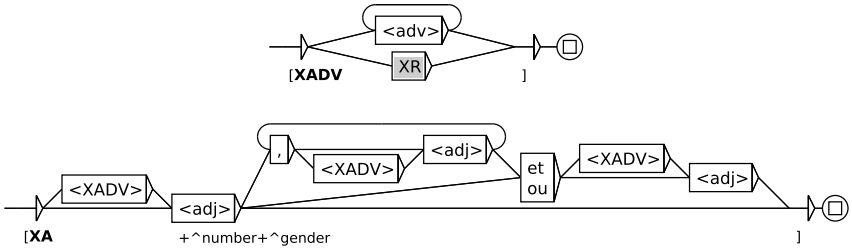


Figure 6.2 : Modifieurs (XADV et XA)

Le **chunk adverbial** (*XADV*, *cf.* Figure 6.2) est intuitivement le plus évident à définir. La plupart du temps, nos énoncés présentent simplement un empilement d'adverbes théoriquement infini[4] et modifiant telle ou telle partie de l'énoncé (*cf.* Exemples 182 à 184). Toutefois, la notion d'adverbe ne se limite pas à cette seule considération. Certaines structures complexes (*XR*) peuvent et doivent être assimilées à cette catégorie (*cf.* Exemples 185 à 188). C'est un phénomène bien connu de la communauté du lexique-grammaire que nous ne pouvions pas ignorer (*cf.* [Gross, 1986a; Maurel, 1989; G. Gross, 1996; Senellart, 1999]). D'un point de vue syntaxique, ces structures relèvent tantôt du chunk nominal (*cf.* Exemple 187), tantôt du chunk prépositionnel (*cf.* Exemple 185 et 186) et tantôt du chunk verbal (*cf.* 188). Mais, parce qu'elles présentent un degré de figement assez important, nous n'avons pas intérêt à les aborder de manière compositionnelle et nous préférons les considérer comme des unités simples (tout comme les noms composés et les collocations).

(182) Il a [$_{XADV}$ en effet discrètement] annoncé [$_{XADV}$ hier] qu'il avait acquis (via la CLT-UFA) 1,99% du capital de M6.

(183) Selon nos informations, sur cette participation de 5,4%, quelque 2,5% seront cédés [$_{XADV}$ très rapidement].

(184) Il faudra [$_{XADV}$ toutefois probablement] attendre de savoir si le groupe allemand passe sous le contrôle de Vodafone.

(185) Rhodia s'est également engagé à céder pour 700 millions d'euros d'actifs [$_{XADV}$ d'ici la fin 2004].

(186) De la sorte, Sepi n'envisage pas de vendre ses parts [$_{XADV}$ à court terme] alors qu'elle détient une option de vente [$_{XADV}$ à partir de juillet prochain].

(187) De son coté , Havas Advertising s'est propulsé au 4^{e} rang mondial de la publicité en rachetant [$_{XADV}$ le 21 février] Snyder pour 2,1 milliards d'euros.

[4]Il s'agit d'un point de vue théorique comme pour la récursivité du *GN*. Dans les faits, l'empilement dépasse très rarement deux unités.

(188) L'euro qui perd depuis son lancement, [$_{XADV}$ il y a un an], 17%, s'effrite encore à 0,9650 dollar.

Outre les adverbes de temps, nous considérons encore les adverbes de lieu (*cf.* Exemple 189 et 190). Quand ils sont introduits par une préposition, ils perturbent l'attachement. En les analysant dès la segmentation en chunks, nous évitons de traiter, au niveau syntagmatique, l'ambiguïté qu'ils génèrent. Notons toutefois que le sens de certains noms propres (*i.e.* noms de ville ou de pays) n'est pas toujours évident à interpréter en recourant à la syntaxe uniquement. Dans la phrase 190, par exemple, il n'est pas possible de déterminer avec certitude si la séquence *aux Etats-Unis* est un adverbe ou le N_2 de la phrase dative élémentaire. Un pays ou une ville peut, en effet, acheter ou vendre des sociétés.

(189) En 1999 , Publicis a racheté Burrell Communications, [$_{XADV}$ aux Etats-Unis], Publi-Graphics, [$_{XADV}$ au Moyen-Orient], Publicis Ad Link [$_{XADV}$ à Hong Kong et en Chine], AMA [$_{XADV}$ aux Philippines], Publicis Zen [$_{XADV}$ en Inde] et Publicis Welcomm [$_{XADV}$ en Corée du sud].

(190) SG vend sa gestion d'actifs [$_{XADV}$ aux Etats-Unis].

Nous avons limité notre étude aux structures adverbiales complexes indicatrices de temps et de lieu. Les autres type d'indicateurs (intensité, aspect, *etc.*) sont nettement moins fréquents ou simplement absents des prédicats relatifs aux transactions financières[5]. Nous présentons un extrait des structures couvertes par notre grammaire dans le tableau 6.4[6].

à l'ouverture des marchés
à la mi-séance
à très court terme
après la clôture de Wall Street
au plus tard au mois d'octobre 2003
au plus tôt l'année prochaine
ce matin
d'ici le 15 avril 2004
d'ici au 31 décembre
d'ici 2007
le 24 février dernier
aux États-Unis
à New York

Tableau 6.4 : Structures adverbiales de temps (extrait)

Le **chunk adjectival** (*XA*, *cf.* Figure 6.2) s'arrête à l'adjectif tête et peut contenir un modifieur

[5] Au sein de la structure N_0 *V* N_1 *Advp* de table 36DT, le complément adverbial, *Advp*, spécifiant le montant, est considéré comme un constituant essentiel (*cf.* Section 6.3).

[6] Notons que le dictionnaire DELAC (mots composés) contient déjà lui-même un nombre très important de structures adverbiales complexes (5661 entrées, tous types confondus) : *au plus (tôt+tard), la semaine (dernière+passée), au cours de l'année, etc.*

adverbial gauche (*cf.* Exemples 191 à 193). Nous prenons également en compte d'éventuelles coordinations (*cf.* Exemples 192 et 193). Précisons que les adjectifs (et les adverbes) ne forment pas toujours un chunk *maximal*[7] et sont intégrés au chunk nominal en cours de processus (*cf.* Exemples 191 et 193). Il reste, au final, les adjectifs post-nominaux et ceux séparés de l'élément qu'ils modifient.

(191) La BNP prend une participation dans EADS, [$_{XN}$ le [$_{XA}$ futur] groupe] d'aéronautique et de défense.

(192) La Commission Bancaire Fédérale Suisse a annoncé aujourd'hui que les autorités de la concurrence [$_{XA}$ suisses, allemandes, irlandaises et américaines] avaient autorisé LVMH à reprendre TAG Heuer (dans le cadre de son OPA sur l'horloger).

(193) Chase Manhattan Bank, l'une [$_{XP}$ des [$_{XA}$ [$_{XADV}$ plus] importantes et prestigieuses] banques] [$_{XA}$ américaines], va étendre sa gamme de produits d'investissements en rachetant la banque Hambrecht & Quist, pour 1,35 Md $.

Structures nominales et prépositionnelles

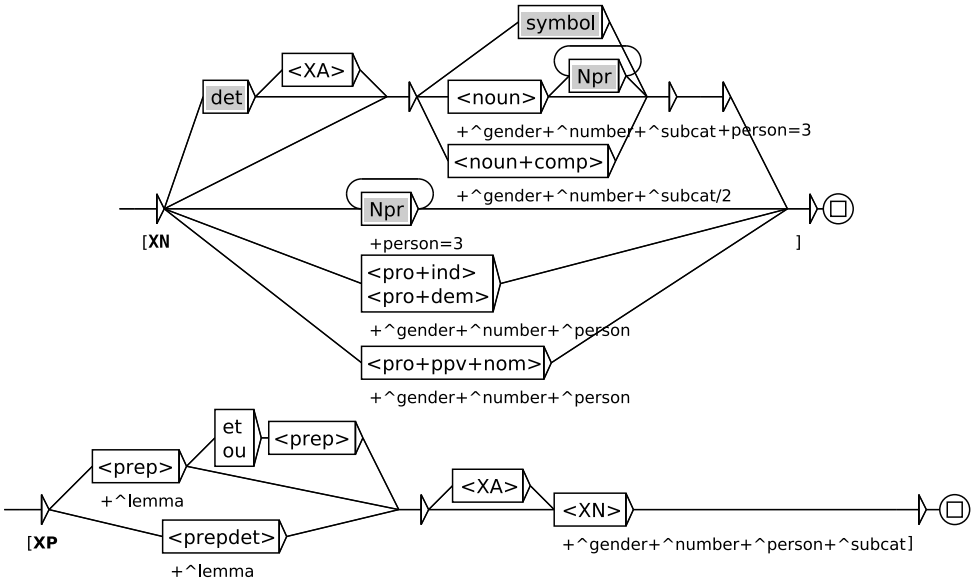


Figure 6.3 : Chunks nominal et prépositionnel (XN et XP)

Le **chunk nominal**, tout comme l'ensemble des chunks, se termine à l'élément tête de la structure, c'est-à-dire le nom (commun ou propre, simple ou composé). Dès lors, les éventuelles appositions de même que les modifieurs droits et les compléments prépositionnels ne sont pas

[7] *A maximal chunk is a chunk that is contained in no other chunk.* [Abney, 1996 : p. 2]

pris en compte. Toutefois, le nom tête peut être une unité complexe ; c'est-à-dire un nom composé ou une collocation (notée <*noun+comp*>, dans la figure 6.3 et [$_N$ ...], dans les exemples). Nous privilégions d'ailleurs cette analyse en la pondérant au sein du graphe de reconnaissance[8]. De cette manière, nous obtenons des structures qui dépassent la signature syntaxique de XN et qui peuvent déjà couvrir l'entièreté du syntagme nominal sans qu'il ait été nécessaire d'en formaliser toute la complexité (*cf.* Exemples 194 à 195).

(194) L'information a été donnée aujourd'hui par Gucci qui a racheté Yves Saint Laurent (et [$_{XN}$ l' [$_N$ ensemble des actifs]] de Sanofi Beauté à Artémis , [$_{XN}$ le [$_N$ holding personnel]] du [$_{XN}$ [$_N$ milliardaire français] François Pinault]).[9]

(195) Dans l'ensemble, [$_{XN}$ le [$_N$ chiffre d'affaires]] dévoilé est supérieur aux anticipations initiales de la direction qui visait [$_{XN}$ un [$_N$ chiffre d'affaires annuel]] de [$_{XN}$ 280 [$_N$ millions d'euros]].

(196) [$_{XN}$ Le [$_N$ résultat net]] apparaît en baisse en raison de [$_{XN}$ la forte plus-value] enregistrée l'année dernière par [$_{XN}$ la [$_N$ vente de l'activité services]].

Le **chunk prépositionnel** combine simplement une préposition ou une combinaison de prépositions et un chunk nominal qui peut être amputé du déterminant initial (*cf.* Exemples 197 et 198).

(197) L'information a été donnée aujourd'hui par Gucci qui a racheté Yves Saint Laurent (et l' ensemble des actifs de Sanofi Beauté à Artémis, le holding personnel [$_{XP}$ du [$_N$ milliardaire français] François Pinault]).

(198) Carrefour confirme son objectif d'une progression à deux chiffres de son bénéfice net par action, [$_{XP}$ avant et après [$_N$ amortissement des survaleurs]], de 10% sur l'ensemble de l'année 2003.

Structures verbales

Nous distinguons trois structures verbales différentes selon le mode du verbe tête : **simple** (*i.e.* indicative, subjonctive et conditionnelle ; *cf.* Exemples 199 à 201), **infinitive** (*cf.* Exemples 202 à 204) et **participiale** (*cf.* Exemples 205 à 206). Nous définissons comme lemme du chunk le lemme du verbe tête afin de permettre son utilisation au sein d'un graphe paramétré.

Nous distinguons les formes verbales **actives** (XV, XVG, XVI et XVK) des formes **passives**

[8] La notation $/2$ de la figure 6.3 signifie qu'un poids de 2 est associé au chemin <*noun+comp*>.

[9] Nous présentons, dans ces exemples, les groupements qui offrent un intérêt particulier. Nous ne mettons pas en évidence l'ensemble des groupements afin de ne pas rendre la lecture trop ardue.

(*XVP*, *XVGP*, *XVIP* et *XVKP*). Nous étendons également la notion de chunk verbal aux structures verbales composées d'un verbe *plein* et d'un verbe *opérateur* pour autant que celles-ci ne soient pas séparées par un segment dépassant le chunk adverbial (*cf.* Exemples 199 et 201).

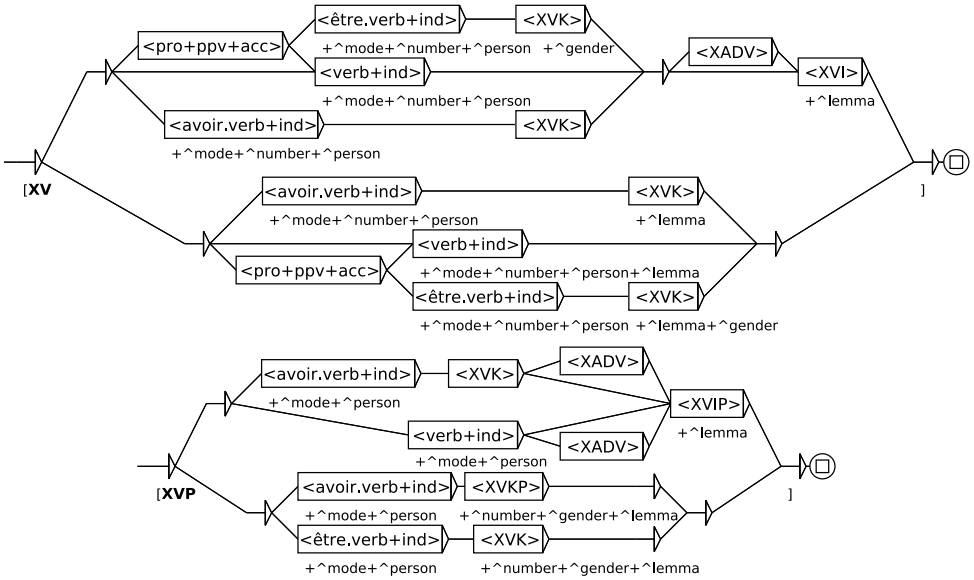


Figure 6.4 : Structures verbales simples (XV et XVP)

(199) Il [$_{XV}$ a ainsi confirmé lundi avoir acquis], via son holding CNP (Compagnie Nationale à Protefeuille), 54% du capital du groupe Joseph (maison de mode).

(200) Air France et Lufthansa [$_{XVP}$ étaient cités] comme les partenaires les plus probables mais il [$_{XV}$ semble] que les allemands [$_{XV}$ aient pris] une longueur d'avance.

(201) Le groupe hôtelier SHB [$_{XVP}$ va en effet être intégralement repris] par Accor pour un montant de 54,8 millions d'euros.

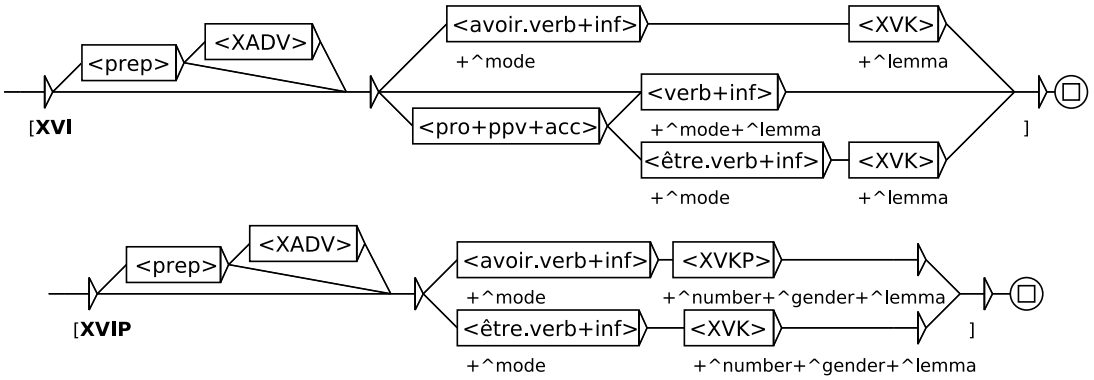


Figure 6.5 : Structures verbales infinitives (XVI et XVIP)

(202) Dexia serait en effet assez bien placé [$_{XVI}$ pour reprendre] la banque grâce à son association avec l'homme d'affaires israélien Avi Roami.

(203) Après ETPM , c'est au tour d'une société de BTP jugée non stratégique d'[$_{XVIP}$ être cédée].

(204) Rappelons que Dexia n'a jamais caché son intention [$_{XVI}$ de se rapprocher] d'un réseau de distribution en France pour vendre ses produits financiers.

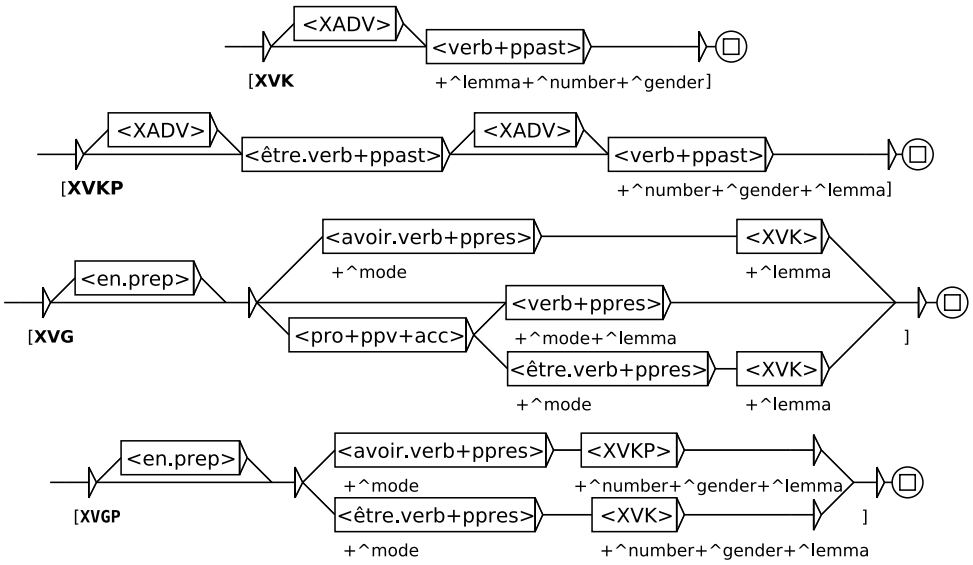


Figure 6.6 : Structures verbales participiales (XVG et XVGP)

(205) L'opérateur de téléphonie mobile [$_{XVGP}$ ayant été racheté] à plus de 9000 euros par abonné, Vivendi et Bouygues ont soudainement vu la valeur de leurs activités de téléphonie mobile augmenter à vitesse grand V.

(206) Dans le détail, BellSouth va dans un premier temps racheter 100% d'E-Plus puis céder 77,5% à KPN [$_{XVG}$ en maintenant] sa participation à 22,5%.

(207) [$_{XVG}$ En acquérant] 35% du capital de Devimo, Ségécé est désormais présent en Belgique.

Exemple de segmentation en chunks

Nous reprenons l'exemple que nous proposions en fin de chapitre 5 pour illustrer l'attachement en collocations (segments soulignés) et la reconnaissance des noms propres (segments mis en gras). Les séquences de chunks ont été délimitées par les crochets ouvrant et fermant. Leur étiquette est précisée au niveau du crochet ouvrant.

[$_{XN}$ **Microsoft**] : [$_{XN}$ son activité internet] [$_{XA}$ bénéficiaire] [$_{XP}$ pour la première fois]

[$_{XN}$ Le leader mondial des logiciels], [$_{XN}$ le groupe américain Microsoft], [$_{XV}$ a indiqué] que [$_{XN}$ sa division internet **MSN**] [$_{XVP}$ a été] [$_{XA}$ bénéficiaire] [$_{XADV}$ au troisième trimestre 2003], et [$_{XN}$ ce] [$_{XP}$ pour la première fois] [$_{XP}$ depuis sa création] [$_{XADV}$ il y a de cela huit ans]. [$_{XN}$ Le bénéfice net] [$_{XP}$ de **MSN**] [$_{XV}$ s'élève] [$_{XP}$ à 58 millions de dollars], [$_{XP}$ contre une perte] [$_{XP}$ de 147 millions] [$_{XADV}$ il y a un an]. [$_{XN}$ L'activité de la division] [$_{XVP}$ a été dopée] [$_{XP}$ par le bond des recettes publicitaires] ([$_{XN}$ +50%]), qui [$_{XV}$ ont entraîné] [$_{XN}$ une hausse du chiffre d'affaires] à [$_{XN}$ 128 millions de dollars]. [$_{XP}$ D'autre part], [$_{XN}$ **Microsoft**] [$_{XV}$ a indiqué] que [$_{XP}$ dans l'ensemble] [$_{XN}$ de ses sept divisions], [$_{XN}$ quatre] [$_{XV}$ ont dégagé] [$_{XN}$ un résultat d'exploitation positif]. [$_{XN}$ La division Windows] [$_{XV}$ a réalisé] [$_{XN}$ un bénéfice d'exploitation] [$_{XP}$ de 2,26 milliards] [$_{XP}$ pour un chiffre d'affaires] [$_{XP}$ de 2,81 milliards]. [$_{XN}$ La division Office [$_{XV}$ a dégagé] [$_{XN}$ un bénéfice] [$_{XP}$ de 1,59 milliard] [$_{XP}$ pour 2,29 milliards] [$_{XP}$ de ventes] et [$_{XP}$ la division serveurs], [$_{XN}$ 370 millions] [$_{XP}$ pour 1,87 milliard] [$_{XP}$ de chiffre d'affaires]. [$_{XN}$ Les divisions **Business**] [$_{XN}$ solutions], [$_{XP}$ des logiciels] [$_{XP}$ pour téléphonie mobile] et "[$_{XN}$ **Home Entertainment**]" [$_{XV}$ ont] [$_{XN}$ toutes] [$_{XVK}$ subi] [$_{XN}$ des pertes]. [$_{XP}$ De ce fait], [$_{XVG}$ en incluant] [$_{XP}$ d'autres pertes d'exploitation] [$_{XP}$ de 751 millions de dollars], [$_{XN}$ **Microsoft** [$_{XV}$ a dégagé] [$_{XN}$ un bénéfice d'exploitation total] [$_{XP}$ de 3,15 milliards de dollars].

Il s'agit bien sûr d'un exemple théorique. Nous n'avons retenu que l'enchaînement correct. Contrairement à [Abney, 1991] et [Ait-Mokhtar et Chanod, 1997] qui travaillent au départ d'un texte désambiguïsé, nous conservons toute la combinatoire induite par l'application des dictionnaires. Par conséquent, notre procédure réduit la complexité sans l'évacuer complètement (*cf.* Tableau 6.5). L'ambiguïté restante est de deux types :

– certaines entrées lexicales n'entrent pas dans une structure de chunk (*i.e. Le*, nom ; *a*, nom ; *la*, nom et pronom ; *etc.*). Dans ce cas, l'ambiguïté est totalement artificielle puisque nous ne prenons plus en compte ces éléments au niveau de l'attachement en syntagmes[10].
– certains chunks entrent en concurrence l'un avec l'autre. Ces conflits constituent l'ambiguïté réelle qu'il nous faudra gérer au niveau de l'attachement :
  – *de 281 milliards de dollars* : *XP vs XN*
  – *dans l'ensemble* : *XADV vs XP*
  – *s'élève* : *XV vs XN XN*
  – *etc.*

Niveau	États		Transitions	
	Total	x/Phrase	Total	x/Phrase
Énoncé	326	36,22	809	89,89
Lexical	280	31,11	640	71,11
Syntaxique	189	21	413	45,89

Tableau 6.5 : Évolution successive de l'ambiguïté

[10]Nous n'incluons pas dans ces éléments les conjonctions de coordination et de subordination qui sont expressément écartées de la segmentation

6.3 Attachement des constituants et sous-langage

Dans le paradigme d'une grammaire lexicalisée, comme celle définie par le lexique-grammaire, chaque prédicat, ou famille de prédicats, forme une ancre lexicale à laquelle sont associées différentes propriétés : le nombre et la nature des arguments (*i.e.* la fenêtre de sous-catégorisation), les prépositions introductrices des arguments et les transformations que le prédicat actualise. Ces trois propriétés définissent, pour un prédicat donné, un ensemble de *patrons d'attachement* qui correspondent à la phrase élémentaire et à ses dérivés transformationnels. Ainsi, par exemple, [Guillet et Leclère, 1992] associe, aux prédicats de la table 36DT, les patrons suivants :

– N_0 *V* N_1 *à* N_2
– N_1 *est Vpp par* N_0 *à* N_2
– N_0 *Vsup le V-n de* N_1
– N_1 *est Vpp*
– N_0 *lui V* N_1
– N_0 *V Dnum N*
– *etc.*

Par ailleurs, les propriétés distributionnelles caractérisant les différents constituants (*i.e.* la nature des arguments qui détermine la *sélection lexicale* du prédicat) nous permettent de contraindre davantage la procédure d'attachement en imposant le trait *humain* au sujet, le trait *concret* à l'objet direct, *etc.*

Théoriquement, sur base des trois propriétés que nous avons définies ici, nous devrions être à même de répondre aux problèmes posés par l'attachement[11]. Toutefois, au chapitre 3, nous avons dû constater qu'elles ne permettent pas toujours de discriminer attachements au verbe et au nom (principalement dans le cas des prépositions *de* et *à*). Pour expliquer ce phénomène, nous avons avancé deux causes possibles[12] : la récursivité du syntagme nominal et l'objet d'étude du lexique-grammaire (*i.e.* la langue standard). Pour évacuer le premier problème, nous avons envisagé une étape préalable à l'attachement, la segmentation en syntagmes non-récursifs. Il est évident que cette solution ne résoud rien mais elle offre le double avantage d'accélérer le processus d'analyse et de limiter la combinatoire lexicale. À présent, il nous reste à considérer la question de la **spécialisation linguistique**.

[Lehrberger, 1982] isole différents facteurs qui permettent de caractériser cette spécialisation propre à un sous-langage donné au regard d'un autre sous-langage ou de la langue standard :

[11] L'attachement peut être une question d'interprétation qui demande alors la prise en compte du contexte. Dans ce cas, le cadre phrastique est trop limité ; la simple analyse syntaxique ne suffit donc pas.

[12] Nous pourrions également ajouter une troisième cause : nos grammaires ne formalisent pas nécessairement des phrases complètes pour identifier l'ensemble des constituants essentiels. Par conséquent, nous ne sommes pas forcément en mesure de profiter pleinement de la désambiguïsation contextuelle que nous évoquions au point 3.1.

les **restrictions lexicales**, **syntaxiques** et **sémantiques**, la **déviance grammaticale** et les **symboles**. Nous avons exploré ces facteurs afin de mesurer leur impact sur l'analyse syntaxique en général et sur l'attachement en particulier.

6.3.1 Restrictions lexicales

La limitation du focus sémantique induit nécessairement une restriction du vocabulaire aux concepts du domaine traité. Si nous comparons le *ratio types* / *tokens*[13] d'un corpus de spécialité et d'un corpus généraliste de tailles équivalentes, nous constatons une nette différence qui vient confirmer cette intuition (*cf.* Tableau 6.6).

	firstinvest.com	Le Soir
TOKENS	1 679 169	1 728 281
TYPES	33 407	82 698
RATIO	**1,99%**	**4,78%**

Tableau 6.6 : Ratio type/token spécialisé vs général

Considérée du point de vue de la phrase élémentaire, la restriction lexicale confère une plus grande cohérence aux classes d'équivalences que forment les constituants essentiels de la phrase. Cette cohérence ne s'exprime pas uniquement de manière quantitative par une réduction du nombre d'entrées. Dans son étude des bulletins boursiers, [Nakamura, 2004] constate en effet que les groupes nominaux sont réductibles à quelques schémas prototypiques suggérant une limitation de la cooccurrence au niveau lexical[14].

(208) ABC gagnait 2,84% vendredi 24 août dans les premières transactions, à 18,10 euros.[15]

(209) L'action ABC s'apprécait de 0,94% vendredi dans les premiers échanges, à 37,48 euros.

(210) Le titre d'ABC a bondi de 3,29% à 16,78 euros, mercredi matin.

(211) L'action du spécialiste de la carte à puce ABC restait stable dans les premières transactions, lundi 20 août, à 3,09 euros.

À partir des exemples 208 à 209 [Nakamura, 2004] définit la structure élémentaire $N_0\ V\ N_1\ N_2$

[13] Le *ratio* entre *types*, le nombre d'unités lexicales total, et *tokens*, le nombre d'unités lexicales différentes, détermine la variété lexicale d'un corpus.

[14] Au chapitre 5, nous avons montré que la restriction de cooccurrence est généralisée à l'ensemble du sous-langage et peut être mesurée hors du contexte de la phrase élémentaire. Nous avons mis au point une procédure d'acquisition de ces structures lexicales qui apporte également une solution à l'ambiguïté d'attachement (*cf.* Section 5.2.6).

[15] Ces exemples, de même que les structures nominales qui leur sont associées, sont empruntés à [Nakamura, 2004].

AdvT AdvL, où N_0 représente le sujet ; N_1, le complément numérique relatif ; N_2, le complément numérique absolu ; *AdvT* et *AdvL*, les compléments adverbiaux de temps et de lieu. Il formalise ces compléments de la manière suivante :

- N_0 : *(l'action+le titre+la valeur) ((de+E) Npr + de Ns Npr + E)*[16]
- N_1 : *(de+E) Dnum (% + pour cent)*
- N_2 : *Dnum Nunité*
- *AdvT* : *(dans <le> <premier> + au début de <le> + au cours de <le>) (transaction+échange+cotation)*

Ces structures définissent avec précision l'attachement au niveau du groupe nominal et viennent augmenter les contraintes pesant sur la phrase élémentaire. Dès lors, il n'y a plus à proprement parler ambiguïté. Notons, cependant, que les structures étudiées par [Nakamura, 2004] sont très figées. Il n'est pas toujours possible de résumer les classes d'équivalences à un ensemble d'expressions rationnelles. Les constituants de la table 36DT présentent une variation plus importante. Certaines structures émergent (*cf.* Exemples 212 à 217) mais elles sont mêlées à une variété d'autres moins évidentes à généraliser.

(212) En début de semaine, Ahold avait déjà vendu ses enseignes en Indonésie [$_{N3}$ pour un montant de 12 millions d'euros].

(213) Suez a vendu sur le marché 3,7 millions d'actions de Total [$_{N3}$ au prix unitaire de 134,10 euros] par l'intermédiaire de Citigroup, avancent certaines sources du marché.

(214) Précisons que préalablement à l'offre, SFL avait acquis 28% du capital [$_{N3}$ au prix de 38,5 euros par action], dont 14,5% auprès des AGF et 13,5% auprès de la Société Générale.

(215) Le gouvernement néerlandais a annoncé avoir vendu sur le marché [$_{N1}$ une participation de 12% dans l'opérateur télécoms national KPN] pour près de deux milliards d'euros.

(216) On savait déjà que Lagardère allait reprendre [$_{N1}$ les 10% de TimeWarner et les 20% de Vivendi].

(217) Concrètement , Thomson-CSF va acquérir [$_{N1}$ 50% des activités de défense de Samsung Electronics].

La restriction lexicale nous permet, également, de limiter la combinatoire engendrée par l'application des dictionnaires. Les DELA et DELACF offrent une très large couverture parce qu'ils visent l'analyse de la langue générale, plus variée au niveau lexical. Par conséquent, une grande

[16] *Ns désigne un ensemble de groupes nominaux comportant une sous-classification des activités. Ex : société de production audiovisuelle, entreprise de fournitures de bureau, etc. E représente l'élément vide. "+" représente un choix dans le paradigme.* [Nakamura, 2004 : p. 841]

partie de l'ambiguïté est imputable à des entrées qui ne ressortissent pas au domaine de la finance et est donc totalement artificielle. Considérons un simple exemple : le pronom de la deuxième personne du singulier, *tu*, est absent de nos corpus. Dès lors, nous pouvons limiter l'interprétation de formes telles que *chiffres* ou *titres*, très fréquentes, à la catégorie nominale et évacuer l'ambiguïté $N - V$[17].

6.3.2 Restrictions syntaxiques

La spécialisation syntaxique s'observe doublement : par un déviance grammaticale que nous avancions au chapitre 4, et sur laquelle nous revenons plus loin dans ce chapitre, et par une **réduction du nombre de structures et de transformations**, et donc du nombre de patrons d'attachement. Ainsi, par exemple, les phrases impératives sont totalement absentes de nos corpus. L'emploi des interrogatives n'est pas, quant à lui, absent mais est limité à quelques structures fragmentaires, principalement des titres de dépêches, au sein desquelles la seule marque d'interrogation visible est, bien souvent, la ponctuation finale (*cf.* Exemples 218 à 221).

(218) Vivarte : les trois principaux actionnaires prêts à céder leurs titres ?

(219) Danone en lice pour acquérir Dann Foods ?

(220) Sportfive cédé à Advent International ?

(221) Suez prêt à vendre Ondeo Nalco ?

L'analyse de nos corpus financiers nous a permis d'isoler trois catégories structurelles selon que le prédicat étudié est principal ou subordonné, ou qu'il apparaisse coordonné à un autre prédicat.

Structures principales

En position principale, le prédicat actualise le schéma de la phrase élémentaire (N_0 *V* N_1 *à* N_2), augmenté de certaines appositions, insertions et/ou relatives (*cf.* Exemples 222 et 225).

(222) Selon les informations publiées aujourd'hui par Le Figaro, [$_{N0}$ Bull] a acquis [$_{N1}$ deux entreprises], [$_{N3}$ pour un montant inférieur à 100 millions d'euros], afin de se renforcer dans les activités de services et celles liées à l'Internet.

(223) [$_{N0}$ La Fnac (groupe Pinault-Printemps-Redoute)] a décidé de vendre [$_{N1}$ son réseau de petites boutiques télécoms] à [$_{N2}$ The Phone House].

[17] Nous utilisons pour cela un *anti-dictionnaire* ou *dictionnaire filtre* qui force l'assignement d'une (ou plusieurs) catégorie(s) à un mot donné.

(224) Selon les informations de l'intersyndicale CGT-FO-CFTC-CFDT-UAI-CGC, [$_{N0}$ la CNP] serait sur le point de revendre à [$_{N2}$ la Caisse des Dépôts et Consignations (CDC)] [$_{N1}$ 50% d'Agifrance].

(225) Rappelons que [$_{N0}$ Vivendi (qui détient 49% de Canal+)] doit céder [$_{N1}$ 9% du capital de sa filiale] [$_{N2}$ à un tiers].

Structures subordonnées

En position subordonnée, le prédicat s'inscrit dans une structure complexe qu'il nous faudrait décrire si nous visions l'analyse syntaxique complète de la phrase. Or, nous cherchons uniquement l'identification des syntagmes pertinents pour l'extraction, c'est-à-dire les constituants essentiels du prédicat spécialisé. Par conséquent, si nous ne pouvons pas totalement faire abstraction de la complexité syntaxique, nous pouvons limiter nos grammaires à la description de l'organisation de ces constituants autour de la tête prédicative. Nous rejoignons en cela les formalismes de [Grishman, 2001] et [Riloff et Jones, 1999] qui limitent leurs patrons d'extraction à des fragments de phrases.

Nous avons rencontré trois types de subordination – **infinitive**, **relative** et **participiale** (participes passé et présent). Nous détaillons, ci-dessous, les structures subordonnées les plus fréquentes.

– N_0 (*P*) <*PREP*> <*XVI*> N_1 *à* N_2[18]

  [$_{N0}$ France Telecom] a en effet annoncé avoir conclu un accord avec Vodafone AirTouch plc pour **racheter** [$_{N1}$ sa participation de 17,24% dans la société de téléphonie mobile allemande E-Plus Mobilfunk GmbH (E-Plus)] [$_{N3}$ au prix de 3,4 milliards de marks (soit 1,7 Md euros)]

– <*PREP*> <XVI> N_1 *à* N_2, N_0

  Après **avoir cédé** en juin [$_{N1}$ 100% du capital de Radio Classique] [$_{N2}$ au groupe Desfossés (La Tribune - propriété de LVMH)], [$_{N0}$ Sagem] a annoncé au quotidien Le Monde son désir de vendre sa participation de 5%.

– N_0 *qui* <*XV*> N_1 *à* N_2

  A force de faire affaire avec d'autres investisseurs, il va finir par ne plus en avoir besoin et M. Murdoch pourrait bien devoir se mettre à négocier avec [$_{N0}$ le tandem Canal+/Vivendi] (qui **a pris** [$_{N1}$ 25% de BSkyB - fleuron de l'empire de Murdoch]).

– N_2 *à qui* N_0 <*XV*> N_1

  Ce consortium est composé de BNP-Paribas, Cinven et Deutsche Bank Capital

[18] Dans cette structure (et la suivante), <*PREP*> abstrait deux classes d'équivalences indiquant soit une intention (*pour*, *dans le but de*, *etc.*) soit une temporalité (*après*, *sur le point de*, *etc.*).

Partners et est associé à [$_{N2}$ Danone] à qui [$_{N1}$ des actifs] **devraient être cédés**.

– N_1 *que* N_0 <*XV*> *à* N_2

Les salles de marché continuent de bruisser de rumeurs au sujet de l'avenir des [$_{N1}$ 60% d'actifs d'Editis] que [$_{N0}$ Lagardère] **va céder**.

– N_1, <*XVK*> *par* N_0 *à* N_2[19]

Egg France est né du rachat par Egg de la [$_{N1}$ Zebank], **acquis** en 2002 [$_{N2}$ auprès de Bernard Arnault et de la banque franco-belge Dexia], pour 5,5 millions d'euros[20].

– <*XVK*> *par* N_0 *à* N_2, N_1

Racheté par [$_{N0}$ l'américain Carnival] en 1997, [$_{N1}$ Costa] s'intègre dans un ensemble de 66 navires, soit le premier groupement mondial.

– (*en*+<E>) <*XVG*> N_1 *à* N_2, N_0

En **cédant** [$_{N1}$ ses titres ABN Amro], [$_{N0}$ ING Groep] vise à réduire son exposition aux marchés actions et renforcer la structure financière de son bilan.

– N_0 (*P*+<E>), *en* <*XVG*> N_1 *à* N_2

[$_{N0}$ Le groupe Comcast], en **rachetant** [$_{N1}$ AT&T systems] en novembre dernier [$_{N3}$ pour 30 milliards de dollars], a fait main basse sur 13 millions d'abonnés.

Structures coordonnées

[Appelt et Israel, 1999] présente la coordination[21] comme étant le problème le plus complexe à traiter en extraction d'information.

> Coordination presents what is probably the most difficult problem in syntactic analysis for any natural-language processing system. There is no single approach to handling coordination in shallow parsing – the particular approach taken depends on just what is being coordinated, as well as domain-relevant properties of the sentence.

Nous ne traitons pas les cas de coordination internes au *GN* qui, au sein de nos corpus financiers, ne présentent pas de réelles difficultés de formalisation. Nous ne traitons pas non plus les cas de coordination de phrases qui, dans le cadre d'une analyse syntaxique de surface, peuvent tout simplement être ignorés (*cf.* Exemples 226 et 228).

[19] Les marques de ponctuation, dans nos structures, sont optionnelles. Nos corpus présentent un emploi généralisé de la ponctuation mais nous avons rencontré certaines occurrences (rares) qui contredisent ce phénomène.

[20] Ce dernier complément est ambigu. Il n'est pas possible de décider s'il relève du prédicat verbal *acquérir* ou nominal *rachat*. Par défaut, nous arrêtons l'analyse à la première marque de ponctuation (dans ce cas-ci la virgule nous empêche de reconnaître le potentiel N_3).

[21] Nous n'abordons pas ici la *disjonction* qui pose des problèmes dépassant le seul cadre syntaxique. Nous limitons donc notre propos à l'étude de la *conjonction*.

(226) Unibail s'engage à acquérir, à un prix de 117 Euros par action, les actions Paris Expo détenues par la Sagi **et** la Sagi s'engage à céder à Unibail ses actions Paris Expo.

(227) De même, Vodafone devra céder Orange (téléphonie mobile en Grande-Bretagne) **et** Vivendi sera sur les rangs.

(228) En l'espace de quelques semaines , Deutsche Telekom a racheté One2One, le quatrième opérateur de téléphonie mobile en Grande-Bretagne**,** France Telecom a repris E-Plus (Allemagne) **et** Mannesmann a annoncé hier soir être arrivé à un accord pour la reprise d'Orange, troisième opérateur de téléphonie mobile en Grande-Bretagne.

La réelle difficulté consiste en définitive à gérer la coordination verbale (*cf.* Exemple 229 à 230). C'est un point que nous avons étudié au chapitre 3 (*cf.* Section 3.5.2). Nous avions identifié la structure suivante :

$$N_0\ V_1\ (W_1)\ et\ V_2\ (W_2)\ W$$

(229) Rappelons que France Telecom semble envoyer des signaux en ce sens puisque [$_{N0}$ l'opérateur] a déclaré vouloir lever [$_{W1}$ 40 milliards d'euros en bourse], céder [$_{W2}$ pour 15 milliards d'actifs] et réaliser [$_{W3}$ des opérations de croissance externe dans la téléphonie mobile en payant en cash].

(230) Plusieurs opérateurs avaient en effet rapporté qu'[$_{N0}$ AXA] préparait [$_{W1}$ une OPA sur Generali] et avait d'ailleurs racheté [$_{W2}$ les actions Generali que Lazard détenait].

(231) [$_{N0}$ Bouygues] rachèterait [$_{W1}$ les minoritaires de TF1] et fusionnerait [$_{W2}$ avec sa filiale de télévision].

Les transformations

En marge des structures de phrase, la restriction syntaxique se reflète également au niveau des transformations. Les prédicats de transaction actualisent trois transformations unaires[22] différentes : la passivation (*cf.* Exemples 232 et 233), la nominalisation (*cf.* Exemples 234 et 235) et l'extraction (*cf.* Exemples 238 et 239)

(232) Revlon [$_{XVP}$ a été repris] dans les années 80 par le milliardaire américain Ronald Perelman (réputé pour ses luxueuses soirées new-yorkaises).

[22]Les transformations binaires on été traitées au niveau des structures syntaxiques subordonnées et coordonnées.

(233) Le site de micro-ondes [$_{XVP}$ pourrait être cédé] au coréen Golstar (N°4 mondial de ce marché).

(234) Le spécialiste français des gaz industriels, Air Liquide, [$_{XV}$ a annoncé] [$_{XN}$ l'achat] des activités métrologie (hors Royaume-Uni) au groupe Brammer, spécialiste européen des instruments de test et mesure.

(235) La banque helvétique [$_{XV}$ négocierait] [$_{XN}$ la reprise] auprès du Crédit Lyonnais (Crédit Agricole) du courtier Laing&Cruickshank.

(236) **C'est** Zodiac **qui** va reprendre la participation du breton ainsi que celle de Dassault Industries (le holding de participations de la famille Dassault).

(237) Dans un premier temps, **ce sont** quelque 15,5% du laboratoire pharmaceutique (soit 10 Mds euros) **qui** seront cédés, comme l'y autorise une clause du pacte d'actionnaire avec L'Oréal.

La nominalisation est la caractéristique syntaxique la plus saillante du domaine de la finance. Elle ne s'observe pas seulement pour les prédicats de transaction ; elle est également mise en évidence par les études des prédicats de variations boursières de [Kittredge, 1982] et [Nakamura, 2004]. Nous avons recensé deux constructions de surface selon que la nominalisation s'intègre à une phrase à verbe support (*cf.* 234 et 235) ou qu'elle apparaisse simplement en tant que groupe nominal (*cf.* 238 et 239) :

$$((N_0\ Vsup) + <E>)\ <DET>\ <V\text{-}n>\ de\ N_1\ à\ N_2$$

(238) Fimalac (stable) : [$_{XN}$ la cession] de la filiale LBC, pour 200 à 250 millions d'euros, est imminente selon la presse.

(239) En décembre, Sammy Corp, le premier constructeur japonais de machines à sous, est devenu le premier actionnaire de Sega par [$_{XN}$ l'acquisition] de 22% de son capital, au prix de 45 milliards de yens, une opération destinée à renforcer sa position sur les marché des jeux d'arcade.

Finalement, on constate un nombre important de structures verbales complexes (*i.e.* un couple *verbe opérateur – prédicat*). Ces structures posent problèmes aux niveaux sémantique (introduction de la modalité[23], *cf.* Exemples 240 à 241) et syntaxique (possibilité d'insertion, *cf.* Exemple 240).

[23]Dans ce travail, nous nous contentons de marquer la modalité. Une prise en compte réelle du problème demanderait un traitement au niveau de la segmentation en chunks. Nous réduisons, en effet, les structures verbales complexes au lemme du verbe plein.

(240) Selon la presse allemande (le quotidien économique Handelsblatt), Deutsche Telekom [$_{XV}$ aurait proposé] à Lagardère de lui [$_{XV}$ racheter] 51% de Club-Internet pour 450 millions d'euros.

(241) Le Point croit savoir que Jean-Marie Messier [$_{XV}$ pourrait céder] Cegetel pour un prix élevé.

(242) LVMH [$_{XV}$ veut racheter] Chaumet.

Si le champ des possibles syntaxiques reste assez varié dans le langage des transactions, d'autres sous-langages, comme celui de la médecine interne, offrent une variété beaucoup plus réduite. [Kittredge, 1982] fait, à ce sujet, une remarque particulièrement intéressante. Il constate que l'on trouve plus de stabilité syntaxique entre deux langues différentes pour un sous-langage donné qu'entre deux sous-langages pour une langue donnée. Son étude des bulletins boursiers présage d'ailleurs celle proposée par [Nakamura, 2004]. Les structures élémentaires identifiées par l'un comme par l'autre sont strictement identiques (*cf.* Exemples 208 à 211, pour le français, et 243 à 246, pour l'anglais). Au regard de l'extraction d'information, cette dernière remarque nous semble particulièrement pertinente et nous permet d'envisager la création d'un fonds de propriétés transversal au domaine[24].

(243) Energy company Perez Companc SA, the most-traded stock on the Buenos Aires exchange, fell 0.74 percent to 6.75 pesos ($6.75) while oil producer YPF SA, Argentina's biggest company, dropped 0.32 percent to 31.5 pesos.

(244) Bangkok Metropolitan Bank Pcl fell 4.5 percent to 2.1 baht.

(245) Shares of Digital Theater Systems Inc. (DTSI) jumped 9.1% to $17.35 after the maker of multichannel audio systems for movie theaters said it had entered an exclusive licensing agreement with digital-cinema operator Avica Technology Corp.

(246) Shares of The9 Ltd. (NCTY) rose 9% to $25 after the Chinese online gaming service company said it earned $10.5 million, or 43 cents an American depositary receipt.

6.3.3 Restrictions sémantiques

La restriction du sujet à un domaine précis entraîne, nous l'avons vu, une limitation du lexique et des structures syntaxiques. Elle s'accompagne également d'une réduction de la polysémie[25].

[24] Nous devrions toutefois poursuivre la comparaison au-delà du couple *français – anglais* trop stable pour être véritablement significatif.

[25] Cependant, bien que cela soit rarement étudié ou mis en avant, les langues de spécialité restent des ensembles dans lesquels la polysémie existe. [Bertels *et al.*, 2006] étudie ce phénomène en essayant de la quantifier.

Tout comme la limitation du lexique, cette caractéristique nous permet de limiter l'ambiguïté des dictionnaires DELA. En supprimant les sens extérieurs au contexte financier, un mot actualisant différentes catégories grammaticales dans la langue standard peut, dans certains cas, être réduit à une unique catégorie dans un sous-langage :

– **cours :** le {cours,.N :ms} de la bourse *vs* les {cours,cour.N :fp} intérieures
– **devise :** la {devise,.N :fs} allemande *vs* devise,deviser.V
– **embarras :** l'embarras,.N :ms financier *vs* embarras,embarrer.V

Cette réduction du sens se reflète également au niveau de la sous-catégorisation des prédicats du domaine. Nous pouvons, en effet, préciser les traits sémantiques associés aux constituants essentiels.

Dans un contexte général, les prédicats associés à la table 36DT sont très variés. Il n'est pas possible de leur associer une sémantique unique. La seule caractéristique commune est le procès que formalise la phrase élémentaire, un échange entre un N_0 *humain* et un N_2 bénéficiaire. Si nous limitons le focus linguistique à un échange financier, nous pouvons réduire et préciser les propriétés et, par conséquent, contraindre davantage l'attachement. Pour les quatre compléments essentiels – N_0, N_1, N_2 et N_3[26] – nous avons défini les propriétés suivantes :

– $N_0 = : N_{hum}$
– $N_1 = : (N_{hum} + Dnum\%\ N + (Dnum{+}{<}E{>})\ N_{titre})$
– $N_2 = : N_{hum}$
– $N_3 = : Dnum\ N_{montant}$

Dnum N caractérise des compléments *numériques*. Pour le complément N_1, il désigne plus particulièrement un pourcentage d'une société tandis que pour N_3, il représente le montant de la transaction (*cf.* Exemple 247). N_{titre} regroupe les compléments de *valeurs* (au sens financier) ; il est comparable au N_0 défini par [Nakamura, 2004] pour les compléments de variations (*cf.* Section 6.3.1).

(247) Altadis rachète [$_{N1}$ 80% du capital de la régie] [$_{N3}$ pour 1,29 milliard d'euros].

Le trait *humain* des compléments N_0, N_1 et N_2 nous permet d'isoler une caractéristique intéressante du sous-langage de la finance : l'emploi des noms propres. Une entreprise est, *a priori*, un objet concret qui peut toutefois désigner un objet humain par référence au conseil d'administration (*cf.* Exemples 248) ou à son porte-parole (*cf.* Exemple 249). Ce point est également mis en avant, par [Nakamura, 2004] et [Kittredge, 1982] au niveau des prédicats de variations boursières (*cf.* Exemple 208).

[26] Le complément adverbial N_3 n'est pas, pour la langue générale, définitoire de la classe. Nous verrons cependant au point suivant qu'il est essentiel dans le sous-langage de la finance.

(248) **Elf cède** une partie de ses actifs anglais.

(249) Nouvelle donne dans le secteur bancaire : **la Société Générale annonce** ce matin avoir pris 3,8% du Crédit Lyonnais.

6.3.4 Déviance grammaticale

Il existe deux types de déviance grammaticale. D'une part, certaines phrases, jugées correctes dans un sous-langage, peuvent être agrammaticales du point de vue de la langue générale. D'autre part, le sous-langage peut présenter des propriétés qui n'existent pas en langue générale.

Agrammaticalité

Pour le sous-langage des transactions financières, le phénomène d'agrammaticalité s'observe principalement au niveau des titres de dépêches[27]. On y rencontre deux sources d'agrammaticalité différentes : l'effacement des déterminants et l'effacement partiel ou complet de la structure verbale (*cf.* Exemples 250 à 252).

(250) Axa : emprunt obligataire en vue

(251) US : demandes d'allocations chômage en baisse

(252) Eurazeo : résultats affectés par les dépréciations

(253) Bayer prêt à conserver seul sa pharmacie[28]

L'agrammaticalité est donc assez localisée dans ce domaine et son impact sur le processus de reconnaissance très limité. D'autres sous-langages, comme celui de la médecine, sont plus concernés.

Typiquement, un rapport médical[29] présente trois articulations : une introduction présentant le patient et le problème, un ensemble de faits (résultats d'examens, proposition de traitement, prescriptions, *etc.*) et une conclusion. La première et la troisième articulations relèvent d'une syntaxe proche de la langue générale. Les fait exposés, par contre, présentent systématiquement une structure syntaxique fragmentaire dont les caractéristiques les plus communes sont l'effacement des déterminants et du verbe (*cf.* Exemples 254 à 256).

[27] Nous pourrions objecter que le titre n'est pas une marque du sous-langage de la finance mais une caractéristique propre au registre journalistique. Cette objection n'est pas pertinente. Un sous-langage détermine un système grammatical clos qui se caractérise par une sémantique ET une syntaxe particulières. Les propriétés syntaxiques du registre journalistique ressortissent donc directement à la définition du sous-langage des transactions financières.

[28] L'agrammaticalité de cette phrase est discutable. Il s'agit, en effet, d'une phrase averbale typique des titres de la presse écrite. Par conséquent, selon la définition que l'on donne à la notion de phrase, elle peut être considérée comme pleinement grammaticale (*cf.* [Lefeuvre, 1999]).

[29] *Patient discharge summary* ou lettre de sortie (après hospitalisation) ou lettre de consultation.

(254) Cliniquement : patiente en état général moyen.

(255) ECG : rythme sinusal régulier à 99 bpm.

(256) Conclusion :

1. Anémie hémolytique à caractère immun avec discrète splénomégalie, d'étiologie exacte non précisée.
2. Valve aortique prothétique : flux diastolique périphérique décelé au doppler transoesophagien, sans signification pathologique certaine.
3. Aucun argument formel pour une endocardite malgré l'accélération importante de la VS et la discrète splénomégalie.

Propriétés spécifiques

La spécialisation syntaxique s'observe non seulement au niveau de la nature des prédicats (*cf.* Section 6.3.3) mais aussi au niveau du nombre des constituants essentiels.

La phrase élémentaire de la table 36DT est composée, en langue standard, de trois constituants essentiels : N_0, sujet de la phrase élémentaire ; N_1, objet et N_2, bénéficiaire. Les propriétés renseignent également un complément optionnel, N_3, précisant le montant de l'échange. Ce complément est généralisé à l'ensemble des prédicats de transaction et peut donc être considéré comme essentiel[30] dans le sous-langage de la finance.

En marge du constituant N_3, nous avons recensé un cinquième constituant, également adverbial, précisant un potentiel intermédiaire de la transaction (*cf.* Exemples 257 à 259). Ce complément (noté N_4) vérifie le trait *humain* (au sens entendu pour N_0 et N_2).

(257) Pernod Ricard vient d'acquérir [$_{N4}$ par l'intermédiaire de sa filiale Pernod Ricard Americas] la totalité du capital de la société mexicaine Tequila Viuda de Romero SA de CV.

(258) Havas Advertising vient de prendre une participation majoritaire, [$_{N4}$ via sa filiale à 100% Euro RSCG Worldwide], au capital de la société américaine The Tyee Group.

(259) Il a en effet discrètement annoncé hier qu'il avait acquis ([$_{N4}$ via la CLT-UFA]) 1,99% du capital de M6.

Un dernier point remarquable est la potentielle agglutination des compléments N_3 et N_1 en un complément N_5 (*cf.* Exemples 260 à 262). Ce complément est très figé (le complément

[30] Il n'est pas à proprement parler essentiel d'un point de vue syntaxique. Nous pouvons en effet l'effacer sans contredire la grammaticalité de la phrase. Mais il apporte une plus-value informationnelle qui le rend essentiel (d'un point de vue informationnel donc) à l'extraction.

N_4 également mais dans une moindre mesure). Sa structure peut être résumée à une simple expression rationnelle :

$$N_5 = : \textit{pour Dnum } N_{\textit{devise}} \textit{ <de.}\text{PREP}\textit{> (<}\text{DET}\textit{>) (actions+titres+actifs)}$$

(260) SCH, qui a acquis [$_{N5}$ pour 17 millions de dollars d'actifs] en Amérique latine au cours des dix dernières années, va désormais axer ses efforts sur son marché local, l'Espagne.

(261) L'équipementier télécoms suédois a vendu [$_{N5}$ pour 456 millions d'euros de ses titres] à un investisseur financier dont l'identité n'a pas été révélée.

(262) Ainsi le chimiste de spécialités doit avoir cédé [$_{N5}$ pour 400 millions d'euros d'actifs] au 30 juin 2004.

En augmentant le nombre de compléments, nous précisons la fenêtre de sous-catégorisation du prédicat et, par conséquent, nous facilitons le processus d'attachement.

6.3.5 Symboles et unités

De nombreux sous-langages présentent des symboles ou des unités qui sont propres au domaine auquel ils appartiennent. Ainsi, la médecine présente des symboles tels que :

– *mg/dl* : miligramme(s) par décilitre,
– *co/jour* : comprimé(s) par jour
– *mg/kg/semaine* : miligramme(s) par kilo et par semaine
– *etc.*

Ces unités confèrent une sémantique particulière aux constituants dans lesquels elles interviennent. Les symboles médicaux que nous avons présentés nous permettent, par exemple, d'associer un trait N_{dose} (pour dosage) aux compléments suivants :

– Emuthiol 10 mg/kg/semaine
– Créatinine 0.9 mg/dl
– Bactrim forte : 2 X 1 co/jour

La finance présente également certains symboles intéressants qui déterminent tantôt un pourcentage (*%*), tantôt un montant (*$*, *MdF*, *MEuros*, *etc.*). Ces deux catégories de symboles caractérisent respectivement les compléments N_1 et N_2 de la table 36DT (*cf.* Exemples 263 et 264). Par conséquent, leur reconnaissance offre un bon indice d'attachement, nous permettant d'éviter des erreurs du type de l'exemple 265.

(263) Concrètement, la Société Générale va céder [$_{N1}$ 20**%** du Crédit du Nord] à Dexia pour 368,9 millions d'euros (soit 2,42 milliards de francs).

(264) Gucci (PPR) rachète Yves Saint Laurent [$_{N1}$ pour 1 **Md $**].

(265) A cet égard, le CCF a cédé son activité de conservation de titres à Natexis Banques Populaires [$_{N3}$ pour les mêmes raisons].

Nous avons intégré ces informations dans une grammaire de symboles (*cf.* Figure 6.7[31]) liée au chunk nominal. Elle nous permet d'augmenter les informations de sous-catégorisation du constituant XN lors de la segmentation en chunks et de contraindre davantage l'attachement en syntagmes.

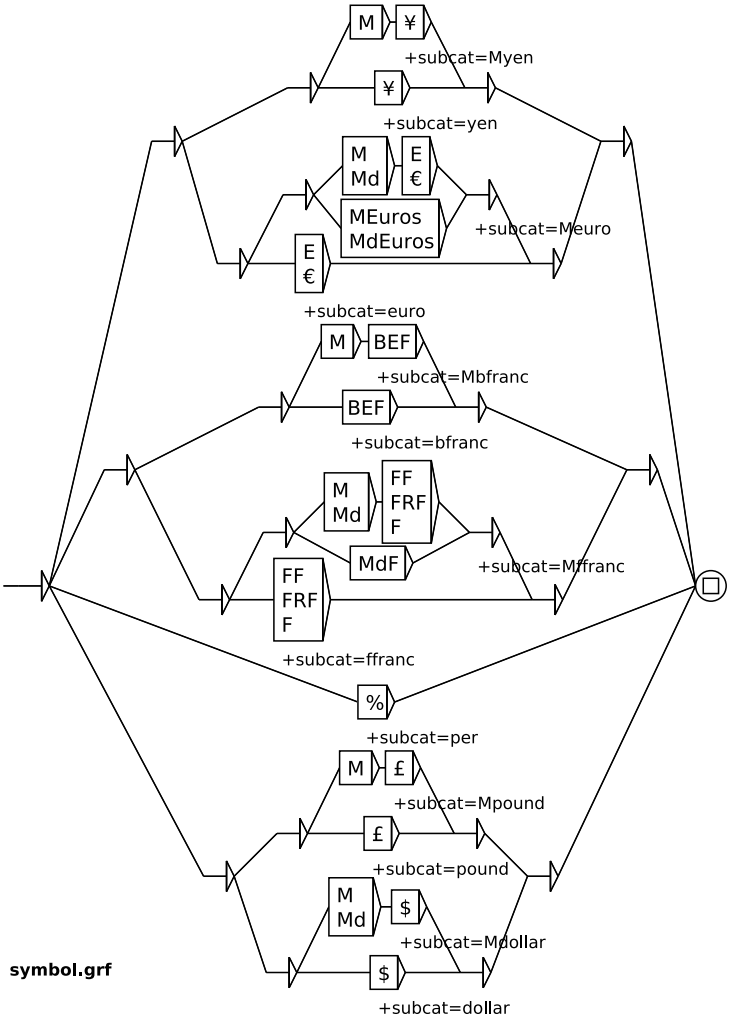


Figure 6.7 : Graphe des symboles (extrait)

[31]Cette grammaire est un extrait d'une grammaire beaucoup plus vaste décrivant l'ensemble des devises et des modifieurs du type *M* ou *Md*. Nous développons actuellement une application de démonstration analysant les fluctuations boursières au sein de textes financiers et calculant automatiquement leur rendu graphique. Dès lors, nous avons besoin d'informations plus précises qu'un simple trait *+devise* ne pourrait nous apporter.

6.4 Conclusion

Nous poursuivions un double objectif dans ce chapitre :

– envisager les perspectives offertes par la segmentation en chunks et
– étudier les implications de la spécialisation linguistique au niveau de l'attachement en syntagmes.

La segmentation en chunks évacue une part non négligeable de l'ambiguïté lexicale tout en augmentant l'efficacité globale en terme de temps de traitement. Au niveau du constituant adverbial, elle nous permet également d'effectuer certains regroupements plus complexes qui dépassent la simple notion de chunk. Nous avons envisagé uniquement les adverbes de temps et de lieu mais la procédure peut s'appliquer à l'ensemble des structures adverbiales figées.

Notre étude de la spécialisation linguistique a, quant à elle, conduit à l'identification de différents facteurs contraignant davantage le processus d'attachement. La réduction à un sous-langage nous permet, d'une part, de restreindre la complexité lexicale des dictionnaires et, d'autre part, de préciser la nature et le nombre de constituants associés aux patrons d'extraction, c'est-à-dire la fenêtre de sous-catégorisation des prédicats.

L'introduction des symboles et des unités spécialisées au niveau de la segmentation en chunks nous a amené à revoir partiellement notre hypothèse de départ d'une séparation des ressources linguistiques générales (les chunks) et spécialisées (les structures élémentaires). Cependant, les implications au niveau de la modularité du système d'extraction sont nulles. Nous conservons une base syntaxique générale qui peut, le cas échéant, être augmentée de certaines structures spécialisées.

Il ressort de notre étude la mise au point d'une procédure s'inscrivant dans la continuité de [Harris, 1964] et [Abney, 1991] et obéissant au principe d'un raffinement linguistique incrémental du même type que celui envisagé au niveau lexical. Nous avons ainsi développé un cadre «réaliste» pour l'utilisation du lexique-grammaire dans des applications réelles.

Conclusion

Dans ce travail, nous souhaitions évaluer la pertinence des bases de données syntaxiques du lexique-grammaire dans le cadre applicatif de l'extraction d'information. Nous avons donc naturellement commencé notre discussion par l'étude du formalisme lui-même. Cette étude, plus théorique, nous a mis au contact de deux outils linguistiques particulièrement pertinents – la phrase élémentaire, unité minimale de sens, et le transformationnalisme harrissien. Ces deux outils nous ont, ensuite, amené à porter un regard nouveau sur l'extraction d'information. En effet, chaque phrase élémentaire définit un patron d'extraction qui, associé aux transformations, offre deux atouts majeurs.

– D'un point de vue syntaxique, le couple *structure élémentaire – transformations* synthétise la variété syntaxique des prédicats et nous permet de réduire considérablement le travail de formalisation.
– D'un point de vue informationnel, la limitation du contexte d'analyse (au sous-langage) nous permet de préciser davantage la sémantique des constituants et d'établir une correspondance directe entre les champs du formulaire et les syntagmes de l'énoncé. Dès lors, il n'y a plus de réelle différence entre analyse syntaxique et extraction proprement dite.

A partir de cette hypothèse initiale, nous avons développé un premier prototype qui repose sur une analyse syntaxique à base de graphes paramétrés présentée dans [Roche, 1993], [Senellart, 1999] et [Paumier, 2003]. Nous avons pu ainsi identifier un problème majeur : le lexique-grammaire, parce qu'il formalise la langue standard, ne décrit pas les spécificités lexicales et syntaxiques de domaines très spécialisés. Dans les faits, ce problème se traduit par une nette insuffisance au niveau de la couverture lexicale et par une augmentation considérable de l'ambiguïté d'attachement.

Pour résoudre cette situation, nous avons envisagé la définition d'un lexique-grammaire de spécialité et la mise au point d'une méthode d'analyse plus à même de répondre aux besoins de l'extraction.

Notre idée d'un lexique-grammaire de spécialité s'inscrit dans la lignée de [Harris, 1971] et du concept de sous-langage. L'application de ce concept nous a permis de préciser les proprié-

tés distributionnelles associées aux tables du lexique-grammaire (structures spécialisées) et de prendre en compte les structures propres au domaine (structures de spécialité). De cette manière, nous contraignons davantage l'attachement des unités lexicales en syntagmes et nous comblons les lacunes de la définition générale.

La méthode d'analyse que nous avons proposée s'inspire également des recherches menées par Zellig Harris qui nous sont apparues, tout au long de ce travail, comme étonnamment actuelles et extrêmement pertinentes. Elle participe de l'idée d'un attachement incrémental qui, pratiquement, opère en quatre temps :

1. Application des dictionnaires des mots simples et composés.
2. Etiquetage des collocations propres au domaine.
3. Segmentation en chunks (ou tronçons syntaxiques non récursifs).
4. Reconnaissance des patrons d'extraction.

L'étiquetage des collocations nous permet d'envisager un attachement au niveau lexical et vient, en cela, compléter la notion de noms composés telle qu'elle est proposée par [Silberztein, 1993] et [G. Gross, 1996]. Tout comme la segmentation en chunks, cette procédure réduit également la combinatoire syntaxique résultant de l'application des dictionnaires. Par conséquent, nous travaillons au niveau de la phrase élémentaire au départ d'un ensemble restreint (et donc moins ambigu) d'unités.

Un corollaire intéressant de cette procédure est la séparation effective des connaissances linguistiques générales et spécialisées. Au contraire des entrées lexicales de nos dictionnaires et des chunks syntaxiques (tous deux généralisables à la langue dans son ensemble), les collocations, de même que les patrons d'extractions, relèvent directement du sous-langage étudié. La modularité qui en découle rejoint l'impératif d'adaptabilité propre à l'extraction d'information.

Parallèlement à ces développements, nous avons également apporté une réponse aux deux questions emblématiques que soulève l'extraction d'information :

Solution linguistique ou statistique ?

Cette première question ne ressortit pas directement à l'extraction d'information mais plutôt au traitement automatique dans son ensemble. Si elle trouve une pertinence particulière au sein de notre discipline, c'est avant tout dans le but de répondre à l'impératif d'adaptabilité que nous évoquions plus avant. À cette fin, nous avons développé une procédure automatique d'acquisition de structures lexicales complexes propres au domaine étudié – les collocations.

Analyse profonde ou partielle ?

L'extraction vise uniquement l'identification des frontières de syntagmes. Par conséquent, se pose la question de la profondeur d'analyse. Notre premier prototype d'extraction reposait sur une analyse syntaxique profonde (bien que limitée aux prédicats du domaine). Cette solution entraîne, nous l'avons vu, une complexité qui n'ajoute rien au résultat final sinon des erreurs. C'est pourquoi nous lui avons substitué une analyse à base de chunks (lexicaux et syntaxiques) qui nous permet de simplifier le processus global.

Au terme de ce travail, nous avons implémenté un extracteur fonctionnel qui valide l'utilisation du lexique-grammaire et confirme, par conséquent, notre hypothèse initiale. Cet extracteur que nous détaillons en annexe, intègre l'ensemble des développements proposés dans ce travail. Il offre donc une solution viable et «prête à l'emploi» rencontrant tous les niveaux d'analyse, de l'extraction d'information jusqu'au remplissage de formulaire. Précisons toutefois que nous nous sommes exclusivement occupé de l'aspect syntaxique de l'extraction et avons dû négliger un certain nombre de problèmes annexes dont le plus intéressant, selon nous, concerne la gestion de la coréférence. C'est un point que nous ne manquerons pas d'étudier dans nos développements futurs.

ANNEXE A

NOTATIONS *Unitex* ET *Outilex*

Nous reprenons ici les notations principales utilisées au sein des dictionnaires et des automates[1].

Code *Unitex*	Code *Outilex*	Signification
A	adj	adjectif
ADV	adverb	adverbe
CONJC	conjc	conjonction de coordination
CONJS	conjs	conjonction de subordination
DET	det	déterminant
INTJ	intj	interjection
N	noun	nom
PREP	prep	préposition
PRO	pronoun	pronom
V	verb	verbe

Tableau A.1 : Principaux codes grammaticaux

Code *Unitex*	Code *Outilex*	Signification
z1	-	langage courant
z2	-	langage spécialisé
z3	-	langage très spécialisé
Abst	abstract	abstrait
Anl	animal	animal
AnlColl	animal+collective	animal collectif
Conc	concret	concret
ConcColl	concret+collective	concret collectif
Hum	human	humain
HumColl	human+collective	humain collectif
t	+transitive	verbe transitif
i	-transitive	verbe intransitif
en	en	particule pré-verbale (PPV) obligatoire
se	se	verbe pronominal
ne	ne	verbe à négation obligatoire

Tableau A.2 : Quelques codes sémantiques

[1] Les informations propres à *Unitex* sont empruntées à [Paumier, 2006].

Code *Unitex*	**Code *Outilex***	**Signification**
m	masculine	masculin
f	feminine	féminin
n	-	neutre
s	singular	singulier
p	plural	pluriel
1,2,3	1,2,3	1^{re}, 2^{e}, 3^{e} *personne*
P	present	présent de l'indicatif
I	imperfect	imparfait de l'indicatif
S	subjpresent	présent du subjonctif
T	subjpast	imparfait du subjonctif
Y	imperative	présent de l'impératif
C	condpresent	présent du conditionnel
J	simplepast	passé simple
W	infinitive	infinitif
G	gerondif	participe présent
K	pastparticipe	participe passé
F	futur	futur

Tableau A.3 : Principaux codes flexionnels

Symbole	**Signification**
<E>, ε	expression vide
W	suite de constituants
<MOT>,<lex>[2]	unité alphabétique
<MIN>,+min	suite alphabétique minuscule
<MAJ>,+maj	suite alphabétique majuscule
<PRE>,+cap	suite alphabétique commençant par une majuscule
<NB>,<num>	suite numérique contigue
#	interdit la présence de l'espace (non supporté par *Outilex*)
=:	marque d'explicitation d'une structure ou d'un constituant
*	signale l'agrammaticalité d'une phrase
?	signale une propriété ou une phrase non vérifiée
+	opérateur de disjonction (*e.g.* (<E>+<DET>) <N>)
	introducteur de traits (*e.g.* <N+Hum>)
:	introducteur de flexion (*e.g.* <N :ms>)

Tableau A.4 : Liste des symboles et unités

ANNEXE B

LA TABLE 36DT : PROPRIÉTÉS STANDARD

Nous reprenons ici l'ensemble des propriétés standard associées à la table 36DT présentée au chapitre 1 (*cf.* Table 1.1).

- $N_0 =: N_{hum}$: propriété transversale à la table et imposant le trait *humain* au sujet du prédicat
- $N_0 =: V\text{-}n$: sujet morphologiquement lié au verbe (*e.g.* Max a escroqué 100 F à Ida[1])
- N_2 *bénéficiaire* : propriété distinguant les datifs bénéfactifs des datifs privatifs (*e.g.* Max offre un diamant à Ida)
- $N_1 =: N_{hum}$: la norme impose un objet *concret* sauf pour quelques verbes (*cf.* déléguer) qui actualisent un objet *humain* (*e.g.* Max délègue son adjoint au ministre)
- $N_1 =: N_{-hum}$: s'ajoute à la propriété précédente et permet de distinguer les verbes dont l'objet est strictement *humain* (éléments notés $-$ dans la table ; *e.g.* Max a kidnappé mon dossier)
- $N_0\ V\ N^0pc$: l'objet est une partie du corps (*e.g.* Max donne le bras à Ida) et appartient au sujet (il s'agit du bras de Max)
- $N_1 =: V\text{-}n$: l'objet est morphologiquement lié au verbe
- $N_1 =: Dnum\ N$: propriété associée aux verbes actualisant un échange d'argent (*e.g.* Max donne une grosse somme à Ida)
- $N_1 =:$ *coup* : l'objet est un coup (*e.g.* Max donne une gifle à Ida)
- *Prép* $=:$ *à* : complément N_2 introduit par la préposition *à*
- *Prép* $=:$ *de* : complément N_2 introduit par la préposition *de* (très rare ; *e.g.* Ida hérite cette pendule de son oncle)
- $N_2 =: N_{-hum}$: bien que la propriété définitoire de la classe impose un N_2 *humain*, il est possible de rencontrer certaines structures dérivées du type N de N_hum (*e.g.* Le jury a attribué le prix au livre de Max)
- *Ppv* $=:$ *lui* : la propriété $N_2 =: N_hum$ impose une pronominalisation du N_2 en *lui* (*e.g.* Max lui attribue le prix)
- *Ppv* $=: y$: les quelques cas vérifiant la propriété $N_2 =: N_{-hum}$ pronominalisent le N_2 en y (*cf.*

[1] Les exemples sont empruntés à [Guillet et Leclère, 1992].

Max y adresse sa lettre)

- $N_2 =: V\text{-}n$: N_2 morphologiquement lié au verbe (*e.g.* Max a affermé sa terre à un fermier)
- $N_2 =: Npl\ obl$: pluriel obligatoire (*e.g.* Max distribue du lait aux enfants)
- $N_1\ est\ Vpp$: participe passé à valeur adjectivale caractérisant le N_1 (*e.g.* Ces biens sont confisqués)
- $N_1 =: N_{abst}$: construction métaphorique (*e.g.* Ce canapé rend tout son charme au salon)
- $N_0\ V\ N_1\ Dnum\ N$: le complément *Dnum N* donne le montant de l'échange (uniquement pour les prédicats actualisant un échange d'argent ; *e.g.* Max a vendu ce livre 10 francs à Ida)
- $N_0\ V\ N_2$ (*de* N_1) : construction à N_1 prépositionnel (*e.g.* Livrer du vin à Paul → Livrer Paul de vin)
- $N_0\ V\ N_1$ *à* $N_2\ Advp$: structure à complément adverbial de prix (*e.g.* Luc a cédé ce livre à Rhoshandiatellyneshiaunneveshenk Koyaanfsquatsiuty[2] pour 60 francs

[2]Prénom de Mlle Williams née à Beaumont – Texas – le 12 septembre 1984. Ceci explique que Maurice Gross ait suggéré l'utilisation de prénoms de trois lettres.

ANNEXE C

MANUEL DE PATEX

C.1 Introduction

Le but de ce manuel est d'exposer de manière claire et concise, les procédures relatives à l'installation, la configuration et l'utilisation de *PatEx*, application résultant des recherches menées dans ce travail. Les aspects plus théoriques liés à cette application sont abordés aux chapitres 3, 4, 5 et 6. Nous indiquerons autant que possible les renvois théoriques au fur et à mesure de nos explications[1].

C.2 Installation

PatEx a été développé de manière à rendre son fonctionnement indépendant de la plate-forme utilisée. Dans ce tutoriel, nous limitons les illustrations à *Linux* mais la procédure reste la même quel que soit le système d'exploitation (**nix* ou *Windows*).

C.2.1 Pré-requis

PatEx repose sur un ensemble de logiciels et de librairies que nous reprenons ci-dessous. Ces logiciels sont disponibles gratuitement et sont fournis avec un manuel détaillant leur installation.

- *TreeTagger* : `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`
- *Unitex* : `http://www-igm.univ-mlv.fr/~unitex/`
- *Outilex*[2]
- *Python (>=2.3)* : `http://www.python.org/download/`
- *libxml2-python* : `ftp://xmlsoft.org/libxml2/python/`
- *libxml2 & libxslt* : `ftp://xmlsoft.org/libxml2/`

[1] Si, après lecture du manuel, certains points restent obscurs ou si certaines erreurs se produisent lors de l'installation et/ou de l'utilisation, n'hésitez pas à contacter l'auteur (Patrick Watrin `watrin@tedm.ucl.ac.be`).

[2] *Outilex* est encore en phase de développement et n'est donc pas proposé en téléchargement pour l'instant. Pour toute information à ce sujet, contactez le professeur Eric Laporte, responsable du projet (`eric.laporte@univ-mlv.fr`).

C.2.2 Installation de *PatEx*

L'installation de *PatEx* est extrêmement triviale. Une procédure automatique effectue les différentes opérations nécessaires (compilation et copie des modules dans le système de fichier). Il suffit en définitive d'ouvrir un terminal et d'y entrer les commandes suivantes :

```
wget http://patlab.org/python/patex-1.0.tar.gz
tar -zxvf patex-1.0.tar.gz
cd patex-1.0/
sudo python setup.py install #l' installation requiert les droits " administrateur "
```

Listing C.1: Commandes relatives à l'installation

C.2.3 Configuration

Nous envisageons ici la configuration de *PatEx* uniquement. Cependant, certains outils utilisés par notre application peuvent nécessiter une configuration particulière. C'est principalement le cas d'*Outilex* si l'utilisateur cherche à définir un ensemble de structures ou de traits propres à un sous-langage donné (*cf.* Section C.6.1 et C.6.2).

Variables d'environnement

Nous l'avons dit, *PatEx* repose sur différentes applications. Plutôt que d'imposer un répertoire d'installation, nous proposons la définition de six variables d'environnement. Ces variables peuvent être définies au niveau du système (fichier `/etc/profile`) ou pour un utilisateur particulier (fichiers `.bashrc`, `.cshrc`, `.kshrc`, *etc.* du répertoire de l'utilisateur).

```
export TREETAGGER_CMD="/usr/local/treetagger/cmd"
export TREETAGGER_CFG="/usr/local/treetagger/lib"
export UNITEX_BIN="/usr/local/unitex/App"
export UNITEX_CFG="$HOME/unitex"
export OUTILEX_BIN="/usr/local/outilex/bin"
export OUTILEX_CFG="/usr/lib/python2.4/site-packages/patex/DOCS/conf"
```

Listing C.2: Export des variables d'environnement

L'exemple proposé ci-dessus varie bien entendu selon le système utilisé et est proposé à titre indicatif. Nous distinguons les répertoires de commandes des répertoires de configurations pour permettre, si besoin, de spécifier un répertoire de configuration alternatif. De cette manière, il est possible à l'utilisateur de modifier les éléments de configuration par défaut sans altérer la configuration initiale. L'intérêt n'est pas immédiat pour le *TreeTagger* qui n'offre pas

de réelle flexibilité à ce niveau. *Unitex* et *Outilex*, par contre, reposent sur un ensemble de fichiers (ressources linguistiques, *etc.*) qui peuvent nécessiter certains changements (au niveau de la segmentation des textes en phrases et en mots, par exemple) afin de répondre aux besoins spécifiques d'une application particulière.

Redirection

L'architecture de *PatEx* a été pensée afin de répondre aux exigences d'une application *Internet*. En ce sens, une attention particulière a été consacrée aux redirections des sorties (standard et erreur) qui peuvent perturber le fonctionnement d'un serveur *Web* comme *Apache*. Différents niveaux de sorties sont définissables par le biais des trois variables d'environnement suivantes[3] :

```
1 export PATEX_VERBOSE="1"
2 export PATEX_DEBUG="0"
3 export PATEX_LOG="/var/log/patex.log"
```

Listing C.3: Export des variables de redirection

La variable `$PATEX_VERBOSE` définit trois niveaux de sorties (notés de `0` à `2`) correspondant à trois degrés de verbosité. Parallèlement la variable `$PATEX_DEBUG` permet d'activer (`1`) ou désactiver (`0`) la sortie d'erreur, essentiellement utilisée lors du débogage des programmes et des ressources linguistiques.

Accessoirement, il est possible de préciser un fichier (en lieu et place de la sortie standard) afin de stocker ces informations (variable `$PATEX_LOG`).

C.3 Utilisation

PatEx est avant tout un ensemble de modules (développés en *Python*) pouvant être réutilisés afin de développer de nouvelles applications ou servir de support à un cours introductif à la programmation liée au traitement automatique de la langue. Nous ne nous étendrons pas sur ce point qui nécessiterait de très longs développements ne ressortissant pas directement à notre propos. Nous proposons toutefois une brève introduction de l'architecture et de l'utilisation de *PatEx* en tant que module.

Parallèlement, une application – *lgx* (*Lexicon-Grammar eXtractor*) – est fournie avec la distribution et permettra au lecteur de réitérer l'ensemble des expériences effectuées dans ce travail. Nous fournissons d'ailleurs (dans le répertoire `data/` de la distribution de *PatEx*) la majorité

[3] Sous *Windows*, il n'est pas possible de supprimer les redirections. Pour éviter toute interférence avec le serveur *Web*, nous suggérons l'utilisation d'un fichier défini par la variable `$PATEX_LOG`.

des ressources linguistiques ayant servi de supports à ce travail (grammaire de chunks et d'extraction, structures lexicales complexes, dictionnaire d'entreprises, *etc*).

C.3.1 Le module *patex*

Le module *patex* est composé de plusieurs sous-modules couvrant l'ensemble des points abordés dans ce travail (analyse et apprentissage lexical, segmentation en chunks, analyse syntaxique, *etc.*). Nous présentons un bref aperçu de ces différentes articulations au sein du tableau C.1.

patex	
`patex.extract`	classes relatives à l'extraction et à l'organisation au sein d'un formulaire des syntagmes identifiés lors de l'analyse syntaxique
`patex.grammar`	classes relatives à l'analyse syntaxique (*chunking* et *parsing*)
`patex.lexicon`	classes relatives à l'apprentissage lexical

patex.stats	
`patex.stats.mwu`	classes relatives au traitement statistique des structures lexicales complexes
`patex.stats.ngrams`	classes génériques ressortissant à la définition des *ngrams*

patex.syntax	
`patex.syntax.chart`	classes gérant les différentes sorties d'*Outilex* (automate du texte, *chart*, *etc.*)

patex.utils	
`patex.utils.graph`	structure de donnée `Graph`
`patex.utils.tree`	structure de donnée `Tree`

patex.wrapper	
`patex.wrapper.outilex`	classes gérant l'ensemble des fonctionnalités d'*Outilex*
`patex.wrapper.shell`	console virtuelle
`patex.wrapper.unitex`	classes gérant l'ensemble des fonctionnalités d'*Unitex*

patex.xml	
`patex.xml.reader`	classes permettant la lecture d'un document XML
`patex.xml.transform`	classes relatives aux transformations XSL

Tableau C.1 : Description du module PatEx

Conformément à la philosophie *Python*, chaque élément (texte, dictionnaire, MWU, *etc.*) est défini sous la forme d'un objet (au sens informatique). Nous proposons ci-après un petit exemple de code source illustrant, de manière très réduite, la solution d'apprentissage lexicale que nous avons développée.

```
1  from patex.wrapper.unitex import UTagText
2  from patex.stats.mwu import MWUList
3
4  def load(txt, grf, lang):
5      utxt = UTagText(lang)
6      utxt.open(txt, 'nt')
7      utxt.config("ttout", "lemma")
8
9      xtr = utxt.tt_extract(grf, 0)
10
11     utxt.close()
12
13     lst = MWUList()
14     lst.load(xtr)
15
16     return lst
17
18 # PROCEDURE PRINCIPALE
19 lang = "French"
20
21 grf = "~/patex/data/2grams.grf"
22 ref = "~/patex/corpus/lesoir.txt"
23 txt = "~/patex/corpus/boursier.txt"
24
25 lst = load(txt, grf, lang)
26 rlst = load(ref, grf, lang)
27
28 lst.balance()
29 lst.score(rlst)
```

Les lignes 4 à 16 définissent une fonction – `load` – qui retourne une liste de fréquences rassemblant l'ensemble des occurrences, au sein du corpus (variable `txt`), de la structure lexicale définie par l'automate (variable `grf`).

L'extraction des structures est gérée par l'objet `UTagText` qui, pour une langue donnée (variable `lang`), associe le *TreeTagger* aux fonctionnalités d'*Unitex*. Cette association nous permet, d'une part, de désambiguïser les textes à analyser et, d'autre part, d'obtenir, en sortie, un ensemble de structures lemmatisées (*cf.* Section 5.2.4).

Une fois les séquences extraites, le fichier XML, résultat de l'extraction lexicale, est passé en argument à l'objet `MWUList` qui nous permet d'appliquer les deux mesures statistiques utilisées – mesure de co-occurrence et de récurrence.

Cet extrait de code montre, selon nous, la facilité de développement qui résulte de la «mise en

objet» des différents logiciels et procédures liés à *PatEx*.

C.3.2 La commande *lgx*

```
pat@patlap ~ [1] $ lgx -h
LGX - Lexicon Grammar eXtraction Tool

  $ lgx [OPTIONS] <task (xml)>

Options:
  [ -h, --help      = print this info                                    ]
  [ -d, --directory = working directory                                  ]
  [ -n, --naive     = naive version ( historic )                         ]
  [ -t, --tag       = use a tagger ( for lexer tool or naive parsing ) ]
```

La commande *lgx* implémente les deux solutions d'extraction (naïve et incrémentale) que nous proposons dans notre travail. Nous l'avons pensée de manière à rendre son utilisation la plus simple possible. Seules trois options sont proposées :

– **`-d`** : précise le répertoire de travail dans lequel sera copié l'ensemble des ressources utilisées afin d'éviter d'encombrer les répertoires de données avec les résultats des différentes étapes de l'extraction (compilation des tables, des automates et des dictionnaires ; traitement des corpus ; *etc.*) ;
– **`-n`** : pour permettre au lecteur de réitérer l'ensemble des expériences proposées dans notre travail, nous avons inclus à *lgx* la solution naïve d'extraction (*cf.* Chapitre 3) ;
– **`-t`** : impose l'utilisation du *TreeTagger*. Cette option est prise en compte au niveau lexical (apprentissage des structures complexes spécialisées ; *cf.* Chapitre 5) et au niveau syntaxique (uniquement pour la solution naïve – option `-n` ; *cf.* Chapitre 3).

L'élément principal dans l'utilisation de *lgx* est le fichier de configuration (unique argument de cette commande). Il organise, dans un format XML, l'ensemble des éléments nécessaires au fonctionnement de l'application :

– le(s) corpus,
– la définition des champs du patron d'extraction,
– les dictionnaires spécialisés et
– les grammaires (de chunks et d'extraction).

Chaque fichier de configuration circonscrit un *sous-langage* (*cf.* Chapitre 4) dont les éléments principaux sont le lexique (*cf.* Chapitre 5) et la grammaire (*cf.* Chapitre 6) qui viennent compléter les ressources générales de la langue (les dictionnaires du français standard pour le plan

lexical et la grammaire de chunks pour le plan syntaxique).

```
1 <?xml version="1.0" encoding="ISO-8859-1"?>
2
3 <patex-task lang="French" cat="finance" subcat="transaction">
4  [...]
5 </patex-task>
```

Listing C.4: Fichier de configuration de PATEX

Le fichier de configuration est, nous l'avons dit, définit dans un format XML. Le listing C.4 précise la syntaxe élémentaire de ce fichier, c'est-à-dire le prologue XML (*cf.* Ligne 1) et l'élément racine[4] (*cf.* Lignes 3 pour l'élément ouvrant et 5 pour l'élément fermant[5]), tous deux obligatoires.

Les éléments *fils* de `patex-task` sont :

– `<corpus>` : définition des sources textuelles,
– `<pattern>` : définition du patron d'extraction et
– `<sublanguage>` : définition des ressources linguistiques.

Nous abordons ces différentes articulations au point suivant.

C.4 Fichier de configuration

C.4.1 Éléments transversaux

Deux éléments sont récurrents dans le fichier de configuration. Il s'agit des éléments `<tab>` et `<grf>` qui définissent respectivement une table (lexicale ou syntaxique) et un automate. Nous utilisons ces deux entités à différents niveaux de l'analyse (pré-traitement des corpus, apprentissage et analyse lexicales et analyse syntaxique).

L'élément `<tab>`

L'élément `<tab>` définit, plus généralement, une table de lexique-grammaire. Le concept de table lexicale est abordé spécifiquement au point C.6.1 de cette annexe. Le chemin d'accès du

[4]Pour être valide, tout document XML doit contenir ces deux éléments une et une seule fois.

[5]Une balise XML est composée d'un élément ouvrant (*e.g.* `<foo>`) et d'un élément fermant (*e.g.* `</foo>`) lorsqu'elle contient du texte (*e.g.* `<foo>`**bar**`</foo>`). Si, au contraire, une balise n'introduit pas de texte, la syntaxe se limite à une unique balise dite «vide» (*e.g.* <foo/>, la présence du *slash* en fin de balise distingue les éléments vides des autres). Notons finalement que les balises vides et ouvrantes peuvent contenir un ensemble d'attributs (*e.g.* l'attribut `f` (de valeur `xyz`) au sein de `<foo f="xyz">bar</foo>` et `<foo f="xyz"/>`). Pour de plus amples renseignements, veuillez vous reporter au document de référence suivant : `http://www.w3.org/XML/`.

fichier (attribut `file`) doit être indiqué de manière absolue, c'est-à-dire de la racine du système de fichier (*/*, sous **nix* et *C :*, sous *Windows*) au fichier proprement dit. Cette dernière remarque est valable pour tous les attributs `file` des divers éléments du fichier de configuration.

```
1 <tab file="/home/pat/patex-1.0/data/tables/36dt-uni.lst"/>
```

Listing C.5: Définition des balises `<tab>`

L'élément `<grf>`

Dans sa forme la plus simple (*cf.* Ligne 1 du listing C.6), l'élément `<grf>` est vide et indique simplement l'emplacement du fichier contenant la définition d'un automate au format *Unitex* (extension *.grf*). Il est toutefois nécessaire de raffiner cette définition si le graphe est paramétré (*cf.* Lignes 3 et 4 du listing C.6) ou s'il repose sur un ou plusieurs sous-graphes (*cf.* Lignes 6 à 10 du listing C.6).

```
1  <grf file="/home/pat/patex-1.0/data/corpus/example.grf"/>
2
3  <tab file="/home/pat/patex-1.0/data/tables/36dt-uni.lst"/>
4  <grf out="@%" file="/home/pat/patex-1.0/data/corpus/example.grf"/>
5
6  <tab file="/home/pat/patex-1.0/data/tables/36dt-uni.lst"/>
7  <grf out="@%" file="/home/pat/patex-1.0/data/corpus/example.grf">
8    <grf out="@%" file="/home/pat/patex-1.0/data/corpus/sub1.grf"/>
9    <grf file="/home/pat/patex-1.0/data/corpus/sub2.grf"/>
10 </grf>
```

Listing C.6: Définition des balises `<grf>`

Un élément `<grf>` paramétré doit, bien entendu, être accompagné d'un élément `<tab>` (*cf.* Section 3.3). Nous avons vu que la compilation d'un graphe paramétré (au regard d'une table de lexique-grammaire) génère un ensemble d'automates. Afin de pouvoir utiliser ces automates dans la suite du processus d'extraction, nous devons ajouter à la définition de la balise `<grf>` un attribut `out` qui décrit la logique de nommage des fichiers générés. Les valeurs possibles de cet attribut sont décrites dans le manuel d'*Unitex* (*cf.* [Paumier, 2006]).

Les dépendances potentielles d'un graphe (*i.e.* les sous-graphes) doivent être associées à la définition du graphe principal. D'un point de vue formel, ces dépendances sont notées comme les éléments *fils* du graphe principal (qui n'est plus, dans ce cas, un élément vide mais possède une balise ouvrante et une balise fermante, *cf.* Lignes 7 et 10). Précisons qu'il ne s'agit pas de formaliser l'organisation hiérarchique des différents graphes et sous-graphes. Les dépendances sont simplement listées au niveau du graphe principal. Ainsi, si un sous-graphe contient égale-

ment un appel à un sous-graphe de niveau inférieur, ce dernier est recensé comme élément fils du graphe principal au même titre que le sous-graphe de niveau supérieur.

C.4.2 Les corpus

Un élément `<corpus>` contient un ou plusieurs éléments `<source>` mais peut également définir un élément `<grf>` (automate) et `<tab>` (table de lexique-grammaire) qui permettent d'isoler un sous-corpus avant de procéder à l'analyse et à l'extraction. La table est typiquement celle servant à l'analyse (*cf.* Section C.6.2) mais pas nécessairement. La complexité du graphe patron associé peut quant à elle varier très largement selon que l'on cherche à isoler les phrases propres à la table étudiée[6] (*cf.* Figure C.1) ou que l'on souhaite isoler certaines structures particulières.

```
<corpus>
  <tab file="/home/pat/patex-1.0/data/tables/36dt-uni.lst"/>
  <grf out="@%" file="/home/pat/patex-1.0/data/corpus/subcorpus.grf"/>
  <source file="/home/pat/patex-1.0/debug/eval/corpus/fi.txt"> Firstinvest .com</source>
  <source file="/home/pat/patex-1.0/debug/eval/corpus/b.txt">Boursier.com</source>
  <source file="/home/pat/patex-1.0/debug/eval/corpus/cf.txt">Cerclefinance.com</source>
</corpus>
```

Listing C.7: Définition des corpus

Le listing C.7 définit trois corpus dont l'emplacement est spécifié par l'attribut `file`. En ajoutant à la définition de `<corpus>` un élément `<tab>` et un élément `<grf>`, nous imposons donc une phase de pré-traitement qui a pour résultat la constitution d'un sous-corpus. Cette phase permet principalement de réduire le temps d'extraction en limitant la masse des données à analyser.

Figure C.1 : Graphe paramétré de pré-traitement

C.5 Les motifs d'extraction

Le motif d'extraction consiste d'une part en une feuille de style XSL (dirigeant la mise en formulaire des motifs extraits ; *cf.* Section 3.3) et d'autre part en une série d'éléments `<data>` établissant le parallèle entre syntagmes formels (attribut `alias`) et segments d'information.

```
<pattern stylesheet="/home/pat/patex-1.0/data/styles/lgx.xsl">
  <data alias="N0">Cessionnaire</data>
  <data alias="N1">Objet</data>
```

[6] Notons que l'ambiguïté n'est bien entendu pas évacuée et qu'une forme potentiellement prédicative peut s'avérer être extérieure au prédicat étudié.

```
    <data alias="N2">Bénéficiaire</data>
    <data alias="N3">Montant</data>
    <data alias="N4">Intermédiaire</data>
    <data alias="N5">Montant + Objet</data>
</pattern>
```

Listing C.8: Définition du motif d'extraction

Les éléments `data` du listing C.8 correspondent aux constituants de la table 36DT spécialisée. Toutefois, et bien que dans ce travail nous ayons limité nos illustrations à cette table, un sous-langage réunit généralement plus d'une classe prédicative (*cf.* Section 2.4).

36DT (N_0 *V* N_1 *à* N_2)	abandonner	abandon
	acheter	achat
	acquérir	acquisition
	céder	cession
	débourser	déboursement
	offrir	offre
	racheter	rachat
	reprendre	reprise
	revendre	revente
	vendre	vente
38LD (N_0 *V* N_1 Loc_{dest} N_2)	investir	investissement
	recentrer	recentrage
38LS (N_0 *V* N_1 Loc_{source} N_2)	désengager	désengagement
	retirer	retrait
	séparer	séparation

Tableau C.2 : Exemple partiel de classe d'extraction

Considérons brièvement le sous-langage des transactions[7]. La table 36DT peut être envisagée comme l'élément directeur de ce sous-langage. Elle contient en effet l'ensemble des éléments pertinents au regard du formulaire d'extraction (*cf.* Section 2.4). À cette première classe viennent se greffer certains autres précicats : les prédicats des tables 38LD et 38LS, par exemple (*cf.* Table C.2). L'association de ces trois classes prédicatives nous amène donc à isoler trois structures canoniques différentes :

– 36DT : N_0 *V* N_1 *à* N_2
– 38LD : N_0 *V* N_1 *Loc* N_2 (avec *Loc* =: *destination*)
– 38LS : N_0 *V* N_1 *Loc* N_2 (avec *Loc* =: *source*)

Bien que les tables 38LD et 38LS, secondaires, offrent une richesse informationnelle moindre, leur inclusion au sein de la classe d'extraction est indispensable afin de récupérer les éléments

[7]Les quelques remarques que nous proposons ici visent à mettre l'utilisateur de *PatEx* au contact des réelles difficultés soutenant la création d'un patron d'extraction qui pourrait, à tort, être considérée comme triviale. Il ne s'agit donc pas de définir une classe d'extraction complète mais d'exposer les points importants à considérer dans sa création.

informationnels manquants au niveau des actualisations des prédicats de la table 36DT. Elle offre, par ailleurs, deux avantages intéressants :

1. Le rassemblement automatique des informations au sein du formulaire d'extraction (*cf.* Section 2.1.2)[8].
2. La résolution de la plupart des cas d'anaphores (*cf.* Section 3.5.3).

Les deux dépêches reprises ci-dessous illustrent ces phénomènes.

[$_{N0}$ ZFS **]** CÈDE **[**$_{N1}$ SES ACTIVITÉS D'ASSURANCE-VIE AUX ETATS-UNIS **]**
[$_{N0}$ Zurich Financial Services (ZFS) **]** se sépare **[**$_{N1}$d'une filiale américaine **]**. **[**$_{N0}$ L'assureur helvétique **]** a vendu **[**$_{N1}$ ses activités d'assurance-vie aux Etats-Unis **] [**$_{N2}$ à Bank One Corporation **] [**$_{N3}$ pour un montant de 500 millions de dollars **]**. Regroupées dans Zurich Life, les activités concernées rassemblent les enseignes Federal Kemper Life Assurance Company, Kemper Investors Life Insurance Company, Zurich Life Insurance Company of America et Zurich Life Insurance Company of New York. L'ensemble des encours gérés par ces marques représentent 292 milliards de dollars. Cette transaction s'inscrit dans le cadre de la restructuration menée par ZFS, qui a essuyé une perte historique de 3,4 milliards de dollars l'an passé.

[$_{N0}$ NESTLÉ **]** CÈDE **[**$_{N1}$ LE CONTRÔLE DE TRINCKS **]**
[$_{N0}$ Nestlé **]** se recentre sur son activité principale, l'agroalimentaire. **[**$_{N0}$ Le groupe helvétique **]** a cédé **[**$_{N1}$ 51% du capital de Trincks, un distributeur allemand de boissons **]**. **[**$_{N1}$ Les parts **]** ont été acquises **[**$_{N2}$ par les brasseurs allemands indépendants Krombacher, Radeberger et Wasteiner **]**. La transaction, dont le montant n'a pas été communiqué, doit encore recevoir le feu vert des autorités de la concurrence avant d'être finalisée.

Tableau C.3 : Exemples de dépêches

Ces structures canoniques ne suffisent, cependant, pas à refléter l'organisation informationnelle des constituants telle qu'elle est définie par le patron d'extraction (*cf.* Listing C.8). D'une part, il n'y a pas de correspondance entre les N_i des différentes structures (*e.g.* le N_2 de la structure 36DT ne correspond pas, d'un point de vue sémantique aux N_2 des deux autres structures) ; et, d'autre part, au sein d'une même structure, selon la charge sémantique d'un prédicat, l'information associée à un syntagme donné peut varier (*e.g.* la différence entre datifs bénéfactifs et privatifs, *cf.* Section 2.4). Afin d'assurer les correspondances inter- et intra-structurelles, il nous faut donc envisager un double aménagement qui doit bien entendu se refléter au niveau de la formalisation des automates d'extraction.

[8]Notons, toutefois, que les prédicats proposés en exemple se prêtent particulièrement au processus que nous évoquons. D'autres peuvent être moins évidents à gérer. Le prédicat *détenir* (table 38LH, [Guillet et Leclère, 1992]), par exemple, ne nous permet pas, sans prise en compte du contexte, de déterminer la nature du sujet (cessionnaire ou bénéficiaire, *cf.* exemples 1 et 2).

(1) Suite à ces achats, PPR détient 6.221.407 de ses propres actions, soit 5,1% de son capital.

(2) Le spécialiste du para-pétrolier a ainsi vendu 9,6 millions d'actions Atos Origin pour 510 millions d'euros. Il détient encore 14,5% du capital de la SSII.

Ce phénomène constitue la principale perspective de développement de *PatEx*.

C.5.1 Correspondance inter-structurelle

La mise en oeuvre de la correspondance inter-structurelle est triviale. Il suffit d'identifier les constituants qui, dans les structures étudiées, renvoient à la même unité informationnelle.

Le sujet (N_0) ne nécessite pas d'aménagement. Il renvoie bien, en contexte, à la même information quelle que soit la structure étudiée (*cf.* Table C.3). Les éléments N_1 et N_2, par contre, sont problématiques et doivent être harmonisés. En effet, l'élément N_1 – objet de la transaction – de la table 36DT est noté N_2 dans la structure 38LS et est absent de la structure 38LD. Pour éviter toute confusion, nous devons donc modifier les structures initiales, telles qu'elle sont proposées par [Guillet et Leclère, 1992] :

– 36DT : N_0 *V* N_1 *à* N_2
– 38LD : N_0 *V N Loc N*
– 38LS : N_0 *V N Loc* N_1

Les éléments extérieurs au formulaire d'extraction sont notés *N* dans ces structures et ne sont pas pris en compte au niveau de l'extraction proprement dite. Leur formalisation au sein des grammaires d'extraction reste toutefois nécessaire afin de rendre compte de la complexité syntaxique des énoncés et de contraindre l'analyse syntaxique.

C.5.2 Correspondance intra-structurelle

Pour la table *36dt*, nous devons dériver deux structures selon que le datif est bénéfactif ou privatif :

– N_2 *V* N_1 *à* N_0 (*e.g.* acheter)
– N_0 *V* N_1 *à* N_2[9] (*e.g.* céder)

Cette organisation se reflète également au niveau des tables 38LD et 38LS pour lesquelles nous définissons les structures suivantes :

– 38LD : N_0 *V N Loc N* (*e.g.* recentrer)
– 38LD : N_2 *V N Loc N* (*e.g.* investir)
– 38LS : N_0 *V N Loc* N_1 (*e.g.* séparer)

[9] La table 36DT formalise ce phénomène en introduisant la propriété N_2 *bénéficiaire* que valident les datifs bénéfactifs.

C.6 Le sous-langage

Un sous-langage, introduit par la balise `<sublanguage>` se définit au départ d'un ou plusieurs lexique(s) (`<lexicon>`) et d'une grammaire (`<grammar>`) composée d'un axiome (`<axiom>`) et, potentiellement, d'une grammaire de chunks (`<chunks>`).

```
<sublanguage>
 [...]
</sublanguage>
```

Listing C.9: Élément racine d'un sous-langage

C.6.1 Les lexiques

PatEx utilise, par défaut, l'ensemble des dictionnaires proposés par *Unitex*. Les ressources lexicales définies dans le fichier de configuration relèvent, quant à elles, du langage de spécialité étudié. Cette architecture reflète la nécessaire modularité d'un système d'extraction (*cf.* Section 2.3.2).

L'élément `<lexicon>` contient deux attributs. L'attribut `name`, nom de la ressource utilisée, et l'attribut `type`, type de la ressource. Seules deux valeurs sont acceptées par ce dernier attribut :

– `dic` : dictionnaires au format DELA,
– `grf` : bibliothèques d'automates.

Dictionnaires DELA

La définition de cet élément ne présente aucune difficulté. Les éléments fils acceptés par ce type de lexique sont au nombre de trois et demandent chacun un attribut `file` spécifiant l'emplacement du ou des fichier(s) dictionnaire :

– `<dic>` : fichier dictionnaire source,
– `<bin>` : fichier binaire contenant le dictionnaire compilé (cet élément doit obligatoirement être accompagné de l'élément `<inf>` correspondant),
– `<inf>` : fichier décrivant les formes compressées associées au fichier `.bin` (*cf.* [Paumier, 2006]).

Si l'élément `<dic>` est présent, le dictionnaire sera automatiquement recompilé que les éléments `<bin>` et `<inf>` soient ou non spécifiés.

```
  <lexicon type="dic" name="entreprise">
    <dic file="/home/pat/patex-1.0/ data/dicos/entreprise . dic"/>
```

```
</lexicon>

<lexicon type="dic" name="entreprise">
  <bin file="/home/pat/patex−1.0/ data/dicos/entreprise . bin"/>
  <inf file="/home/pat/patex−1.0/ data/dicos/entreprise . inf"/>
</lexicon>
```

Listing C.10: Définition d'un lexique DELA

Bibliothèques d'automates

```
<lexicon name="dlc_n2">
  <tab file="/home/pat/patex−1.0/debug/home/pat/patex−1.0/data/lex−tables/dlc_n2 . tab"/>
  <grf out="@%" file="/home/pat/patex−1.0/data/ngrams/mwu/2grams/2grams−param.grf">
    <grf name="Insert" file="/home/pat/patex−1.0/data/ngrams/mwu/Insert. grf"/>
  </grf>
</lexicon>
```

Listing C.11: Définiton d'un lexique de collocations

Le listing C.11 définit une bibliothèque d'automates[10] telle qu'elle est décrite au chapitre 5 (*cf.* Section 5.2.7). Cette définition se compose donc d'un élément `tab`, table lexicale, et d'un élément `grf`, graphe paramétré formalisant la syntaxe des structures recensées dans la table.

```
<lexicon type="grf" name="dlc_n2">
  <tab>
    <grf file="/home/pat/patex−1.0/data/ngrams/mwu/2grams/2grams.grf">
      <grf file="/home/pat/patex−1.0/data/ngrams/mwu/A.grf"/>
      <grf file="/home/pat/patex−1.0/data/ngrams/mwu/Insert.grf"/>
      <grf file="/home/pat/patex−1.0/data/ngrams/mwu/N.grf"/>
      <grf file="/home/pat/patex−1.0/data/ngrams/mwu/PREP.grf"/>
      <grf file="/home/pat/patex−1.0/data/ngrams/mwu/prepN.grf"/>
      <grf file="/home/pat/patex−1.0/data/ngrams/mwu/2grams/A−N.grf"/>
      <grf file="/home/pat/patex−1.0/data/ngrams/mwu/2grams/NA.grf"/>
      <grf file="/home/pat/patex−1.0/data/ngrams/mwu/2grams/NN.grf"/>
      <grf file="/home/pat/patex−1.0/data/ngrams/mwu/2grams/NprepN.grf"/>
      <grf file="/home/pat/patex−1.0/data/ngrams/mwu/2grams/prep−N.grf"/>
    </grf>
    <field name="A−N"/>
    <field name="NA"/>
    <field name="NprepN"/>
    <field name="NN"/>
    <field name="prep−N"/>
```

[10] Ce type de lexique n'est pas supporté par la solution d'extraction naïve (option `-n` de la commande *lgx*).

```
20          <ref  threshold="0" file="/home/pat/patex-1.0/ debug/corpus/stuffs/soir  . txt "/>
21        </tab>
22
23        <grf out="@%" file="/home/pat/patex-1.0/data/ngrams/mwu/2grams/2grams-param.grf">
24          <grf name="Insert" file="/home/pat/patex-1.0/data/ngrams/mwu/Insert. grf "/>
25        </grf>
26      </lexicon>
```

Listing C.12: Définiton d'un lexique de collocations

Au chapitre 5, nous avons décrit un algorithme permettant l'apprentissage de structures complexes. Cet algorithme est évidemment accessible depuis la commande *lgx* et peut être utilisé en marge du processus d'extraction[11] ou intégré à celui-ci[12]. Dans ce cas, l'élément `<tab>` ne renvoie pas à un fichier mais précise l'ensemble des ressources nécessaires à l'apprentissage :

– `<grf>` : graphe responsable de l'extraction des candidats termes,
– `<field>` : l'attribut `name` de cet élément renseigne la signature syntaxique des structures lexicales dirigeant l'apprentissage (*cf.* Section 5.2.4),
– `<ref>` : indique au programme l'emplacement d'un corpus de référence (attribut `file`) activant ainsi la mesure de récurrence (*cf.* Section 5.2.4) ; l'attribut `threshold` permet de filtrer les résultats obtenus en ignorant les structures dont le score est strictement inférieur à l'entier, valeur de cet attribut.

À ce stade, il peut être nécessaire d'apporter certaines modifications à la configuration d'*Outilex* et plus particulièrement au fichier `lingdef.conf`[13]. Au chapitre 5, nous avons envisagé la spécialisation des traits pour les termes propres au sous-langage. Notre formalisation du sous-langage de la finance nous a ainsi amené à définir différents traits extérieurs à la langue standard : `+fin` (finance), `+trans` (transaction), *etc.* Par ailleurs, pour favoriser l'analyse des structures lexicales complexes, nous avons introduit un trait `+comp` (*cf.* Section 5.2.7). *Outilex* ne peut prendre en compte ces informations de sous-catégorisation si elles ne sont pas listées dans son fichier de configuration (*cf.* Listing C.13). Pour plus d'explications quant à la syntaxe de ce fichier, nous renvoyons à la documentation d'*Outilex*.

```
1 <attrtype name='nounsubcat' type='enum'>
2   [...]
3   <value name='fin' alias='finance'/>
```

[11] Pour limiter le processus au seul apprentissage lexical, le fichier de configuration doit obligatoirement être débarrassé des éléments `<grammar>` et `<pattern>` qui activent respectivement les étapes d'analyse syntaxique et d'extraction.

[12] Le résultat de l'apprentissage est alors directement injecté au processus d'extraction et peut, puisqu'il n'y a pas de supervision humaine, occasionner certaines erreurs. Il est toutefois possible de limiter ces erreurs en contraignant les résultats par le biais de l'attribut `threshold` de l'élément `<ref>` que nous décrivons plus loin.

[13] Un exemple de ce fichier est fourni avec la distribution de *PatEx* (répertoire `patex-1.0/conf/`).

```
  <value name='trans' alias='transaction'/>
  <value name='soc' alias='société'/>
   [...]
</attrtype>

<attrtype name='compound' type='bool'>
  <true alias='comp,multiword,mwu'/>
</attrtype>
```

Listing C.13: Extrait du fichier `lingdef.conf`

C.6.2 La grammaire

Une grammaire (`<grammar>`) contient un élément `<axiom>`, principal et obligatoire, et un élément `<chunk>`, optionnel.

```
    <grammar name="36dt−finance">
     [...]
    </grammar>
```

Listing C.14: Élément racine de la grammaire

Les chunks

Les chunks[14] définissent un ensemble de regroupements syntaxiques non récursifs et pré-syntagmatiques (*cf.* Section 6.2). Ils s'organisent en différents niveaux successifs dont le nombre est spécifié par l'attribut `size` de l'élément `chunks`. L'attribut `lvl` des différents éléments `grf` indique, quant à lui, le niveau auquel ressortit le graphe en question.

```
    <chunks size="8">
      <grf lvl="0" file="/home/pat/patex−1.0/data/chunks/XADV.grf">
        <grf file="/home/pat/patex−1.0/data/chunks/XR.grf"/>
      </grf>
      <grf lvl="1" file="/home/pat/patex−1.0/data/chunks/XA.grf"/>
      <grf lvl="2" file="/home/pat/patex−1.0/data/chunks/XN.grf">
        <grf file="/home/pat/patex−1.0/data/chunks/det. grf"/>
        <grf file="/home/pat/patex−1.0/data/chunks/symbol.grf"/>
        <grf file="/home/pat/patex−1.0/data/chunks/Npr.grf"/>
        <grf file="/home/pat/patex−1.0/data/chunks/Dnum.grf"/>
      </grf>
```

[14] De même que les bibliothèques d'automates exposées au point précédent, la segmentation en chunks relève de la solution incrémentale d'extraction et n'est pas supportée par la solution naïve.

```
12        <grf lvl="3" file="/home/pat/patex-1.0/data/chunks/XP.grf"/>
13        <grf lvl="4" file="/home/pat/patex-1.0/data/chunks/XVK-passive.grf"/>
14        <grf lvl="4" file="/home/pat/patex-1.0/data/chunks/XVK.grf"/>
15        <grf lvl="5" file="/home/pat/patex-1.0/data/chunks/XVI-passive.grf"/>
16        <grf lvl="5" file="/home/pat/patex-1.0/data/chunks/XVI.grf"/>
17        <grf lvl="6" file="/home/pat/patex-1.0/data/chunks/XVG-passive.grf"/>
18        <grf lvl="6" file="/home/pat/patex-1.0/data/chunks/XVG.grf"/>
19        <grf lvl="7" file="/home/pat/patex-1.0/data/chunks/XV-passive.grf"/>
20        <grf lvl="7" file="/home/pat/patex-1.0/data/chunks/XV.grf"/>
21      </chunks>
```

Listing C.15: Grammaire de chunks

L'introduction des chunks nécessite une configuration spécifique au niveau du fichier `lingdef.conf` d'*Outilex*. Nous proposons, ci-après, la définition du chunk nominal (XN), l'ensemble des structures restantes sont définies dans le fichier `lingdef.conf` fourni avec la distribution de *PatEx*.

```
1 <pos name='XN' cutename='XN'>
2   <attribute name='subcat' type='nounsubcat' shortcut='yes'/>
3   <attribute name='person' type='person' shortcut='yes'/>
4   <attribute name='gender' type='gender' shortcut='yes'/>
5   <attribute name='number' type='number' shortcut='yes' alias='num'/>
6 </pos>
```

L'axiome ou grammaire d'extraction

Finalement, l'axiome, dernier élément du fichier de configuration, définit la grammaire d'extraction proprement dite. Cet élément est commun aux solutions naïve et incrémentale et se compose d'une table et d'un graphe principal (augmenté si besoin d'un certain nombre de sous-graphes).

```
1       <axiom>
2         <tab file="/home/pat/patex-1.0/data/tables/36dt-uni. lst "/>
3         <grf out="@%" file="/home/pat/patex-1.0/data/axioms/36dt-outilex_new.grf">
4           <grf file="/home/pat/patex-1.0/data/axioms/N.grf"/>
5           <grf file="/home/pat/patex-1.0/data/axioms/N-prep.grf"/>
6           <grf file="/home/pat/patex-1.0/data/axioms/SN.grf"/>
7           <grf file="/home/pat/patex-1.0/data/axioms/SP.grf"/>
8           <grf file="/home/pat/patex-1.0/data/axioms/Ins. grf"/>
9           <grf file="/home/pat/patex-1.0/data/axioms/Inf. grf"/>
10          <grf file="/home/pat/patex-1.0/data/axioms/Par.grf"/>
11          <grf file="/home/pat/patex-1.0/data/axioms/Rel.grf"/>
12          <grf file="/home/pat/patex-1.0/data/axioms/N0.grf"/>
```

```
13          <grf file="/home/pat/patex-1.0/data/axioms/N1.grf"/>
14          <grf file="/home/pat/patex-1.0/data/axioms/N2.grf"/>
15          <grf file="/home/pat/patex-1.0/data/axioms/N3.grf"/>
16        </grf>
17      </axiom>
```

Listing C.16: Grammaire d'extraction

Plusieurs exemples de fichiers de configuration sont distribués avec *PatEx* (répertoire `data/config` de l'archive). L'utilisateur peut s'en inspirer pour s'habituer à la rédaction de ce document. Nous proposons dans listing C.17, une version continue du fichier de configuration détaillé tout au long de ce manuel.

```
1  <?xml version="1.0" encoding="ISO-8859-1"?>
2
3  <patex-task lang="French" cat="finance" subcat="transaction">
4    <corpus>
5      <tab file="/home/pat/patex-1.0/ data/tables/36dt-uni . lst "/>
6      <grf out="@%" file="/home/pat/patex-1.0/data/corpus/subcorpus.grf"/>
7      <source file="/home/pat/patex-1.0/debug/eval/corpus/fi. txt "> Firstinvest .com</source>
8      <source file="/home/pat/patex-1.0/debug/eval/corpus/b.txt ">Boursier.com</source>
9      <source file="/home/pat/patex-1.0/debug/eval/corpus/cf. txt ">Cerclefinance .com</source>
10   </corpus>
11
12   <pattern stylesheet="/home/pat/patex-1.0/data/styles/lgx.xsl">
13     <data alias="N0">Cessionnaire</data>
14     <data alias="N1">Objet</data>
15     <data alias="N2">Bénéficiaire</data>
16     <data alias="N3">Montant</data>
17     <data alias="N4">Intermédiaire</data>
18     <data alias="N5">Montant + Objet</data>
19   </pattern>
20
21   <sublanguage>
22     <lexicon type="dic" name="entreprise">
23       <dic file="/home/pat/patex-1.0/ data/dicos/entreprise . dic"/>
24     </lexicon>
25
26     <lexicon name="dlc_n2">
27       <tab file="/home/pat/patex-1.0/debug/home/pat/patex-1.0/data/lex-tables/dlc_n2 . tab"/>
28       <grf out="@%" file="/home/pat/patex-1.0/data/ngrams/mwu/2grams/2grams-param.grf">
29         <grf name="Insert" file="/home/pat/patex-1.0/data/ngrams/mwu/Insert. grf"/>
30       </grf>
31     </lexicon>
32
```

```
    <grammar name="36dt-finance">
      <chunks size="8">
        <grf lvl="0" file="/home/pat/patex-1.0/data/chunks/XADV.grf">
          <grf file="/home/pat/patex-1.0/data/chunks/XR.grf"/>
        </grf>
        <grf lvl="1" file="/home/pat/patex-1.0/data/chunks/XA.grf"/>
        <grf lvl="2" file="/home/pat/patex-1.0/data/chunks/XN.grf">
          <grf file="/home/pat/patex-1.0/data/chunks/det. grf"/>
          <grf file="/home/pat/patex-1.0/data/chunks/symbol.grf"/>
          <grf file="/home/pat/patex-1.0/data/chunks/Npr.grf"/>
          <grf file="/home/pat/patex-1.0/data/chunks/Dnum.grf"/>
        </grf>
        <grf lvl="3" file="/home/pat/patex-1.0/data/chunks/XP.grf"/>
        <grf lvl="4" file="/home/pat/patex-1.0/data/chunks/XVK-passive.grf"/>
        <grf lvl="4" file="/home/pat/patex-1.0/data/chunks/XVK.grf"/>
        <grf lvl="5" file="/home/pat/patex-1.0/data/chunks/XVI-passive.grf"/>
        <grf lvl="5" file="/home/pat/patex-1.0/data/chunks/XVI.grf"/>
        <grf lvl="6" file="/home/pat/patex-1.0/data/chunks/XVG-passive.grf"/>
        <grf lvl="6" file="/home/pat/patex-1.0/data/chunks/XVG.grf"/>
        <grf lvl="7" file="/home/pat/patex-1.0/data/chunks/XV-passive.grf"/>
        <grf lvl="7" file="/home/pat/patex-1.0/data/chunks/XV.grf"/>
      </chunks>

      <axiom>
        <tab file="/home/pat/patex-1.0/data/tables/36dt-uni. lst "/>
        <grf out="@%" file="/home/pat/patex-1.0/data/axioms/36dt-outilex_new.grf">
          <grf file="/home/pat/patex-1.0/data/axioms/N.grf"/>
          <grf file="/home/pat/patex-1.0/data/axioms/N-prep.grf"/>
          <grf file="/home/pat/patex-1.0/data/axioms/SN.grf"/>
          <grf file="/home/pat/patex-1.0/data/axioms/SP.grf"/>
          <grf file="/home/pat/patex-1.0/data/axioms/Ins. grf"/>
          <grf file="/home/pat/patex-1.0/data/axioms/Inf. grf"/>
          <grf file="/home/pat/patex-1.0/data/axioms/Par.grf"/>
          <grf file="/home/pat/patex-1.0/data/axioms/Rel.grf"/>
          <grf file="/home/pat/patex-1.0/data/axioms/N0.grf"/>
          <grf file="/home/pat/patex-1.0/data/axioms/N1.grf"/>
          <grf file="/home/pat/patex-1.0/data/axioms/N2.grf"/>
          <grf file="/home/pat/patex-1.0/data/axioms/N3.grf"/>
        </grf>
      </axiom>
    </grammar>
  </sublanguage>
</patex-task>
```

Listing C.17: Fichier de configuration complet

Bibliographie

ABEILLÉ, A. (2002). *Une grammaire électronique du français*. CNRS Éditions, Paris.

ABNEY, S. P. (1991). Parsing by chunks. *In* BERWICK, R. C., ABNEY, S. P. et TENNY, C., éditeurs : *Principle-Based Parsing : Computation and Psycholinguistics*, pages 257–278. Kluwer, Dordrecht.

ABNEY, S. P. (1996). Chunk stylebook. Unpublished article.

ACE (2006). *The ACE 2006 (ACE06) Evaluation Plan. Evaluation of the Detection and Recognition of ACE Entities, Values, Temporal Expressions, Relations, and Events*.

AHO, A., SETHI, R. et ULLMAN, J. (2000). *Compilateurs. Principes, techniques et outils. Cours et exercices*. Dunod, Paris.

AIT-MOKHTAR, S. et CHANOD, J.-P. (1997). Incremental finite-state parsing. *In Proceedings of the fifth Conference on Applied Natural Language Processing ANLP'97*.

APPELT, D., HOBBS, J., BEAR, J., ISRAEL, D. et TYSON, M. (1993). FASTUS : A finite-state processor for information extraction from real-world text. *In Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence (IJCAI-93)*, pages 1172–1178.

APPELT, D. E. et ISRAEL, D. J. (1999). Introduction to information extraction technology. *In Proceedings of the* 16^{th} *International Joint Conference on Artificial Intelligence*.

APRESJAN, J. D. (1966). Analyse distributionnelle des significations et champs sémantiques structurés. *Langages*, 1:44–74.

APRESJAN, J. D. (1969). Syntaxe et sémantique. *Langages*, 15:57–66.

ASHER, R. E. (1993). *The Encyclopedia of Language and Linguistics*. Pergamon Press, Oxford, New York, Séoul & Tokyo.

BALVET, A. (2002). *Approches catégoriques et non-catégoriques en linguistique de corpus spécialisés. Application à un système de filtrage d'information*. Thèse de doctorat, Université Paris X-Nanterre.

BARNBROOK, G. et SINCLAIR, J. M. (2001). Specialised corpus, local and functional grammars. *In* Ghadessy *et al.* [2001], pages 237–278.

BAUMGARTNER, R., FLESCA, S. et GOTTLOB, G. (2001). Visual web information extraction with lixto. *In Proceedings of the* 27^{th} *VLDB conférence*, pages 119–128.

BERTELS, A., SPEELMAN, D. et GEERAERTS, D. (2006). Analyse quantitative et statistique de la sémantique dans un corpus technique. *In* MERTENS, P., FAIRON, C., DISTER, A. et WATRIN, P., éditeurs : *Verbum ex machina. Actes de la* 13^{e} *conférence sur le traitement automatique des langues naturelle (TALN'06)*, volume 1 de *Cahiers du Cental*, pages 73–82. Presses universitaires de Louvain.

BIBER, D. (1993a). Representativeness in corpus design. *Literary & Linguistic Computing*, 8(4):243–257.

BIBER, D. (1993b). Using register-diversified corpora for general language studies. *Computational Linguistics*, 19(2):219–241.

BIBER, D. et FINEGAN, E. (1994). *Sociolinguistic perspectives on register*. Oxford University Press, New York.

BLANC, O. et CONSTANT, M. (2005). Lexicalization of grammars with parameterized graphs. *In Proceedings of the international conference Recent Advances in Natural Language Processing (RANLP'05)*, pages 117–121.

BLANC, O., CONSTANT, M. et LAPORTE, E. (2006). Outilex, plate-forme logicielle de traitement de textes écrits. *In* MERTENS, P., FAIRON, C., DISTER, A. et WATRIN, P., éditeurs : *Verbum ex machina. Actes de la* 13^{e} *conférence sur le traitement automatique des langues naturelle (TALN'06)*, volume 1 de *Cahiers du Cental*, pages 83–92. Presses universitaires de Louvain.

BOGURAEV, B. et PUSTEJOVSKY, J., éditeurs (1996). *Corpus processing for lexical acquisition*. MIT Press.

BONELLI, E. T. (2000). Functionally complete units of meaning across english and italian : Towards a corpus-driven approach. *In* Granger et Altenberg [2000], pages 73–96.

BOONS, J.-P., GUILLET, A. et LECLÈRE, C. (1976). *La structure des phrases simples en français : constructions intransitives*. Droz, Genève.

BOUDREAU, S. et KITTREDGE, R. (2005). Résolution des anaphores et détermination des chaines de coréférences. différences entre variétés de texte. *Traitement automatique des langues*, 46(1):41–69.

BRILL, E. (1992). A simple rule-based part-of-speech tagger. *In Proceedings of ANLP-92, 3rd Conference on Applied Natural Language Processing*, pages 152–155, Trento, IT.

CALIFF, M. et MOONEY, R. (1997). Relational learning of pattern-match rules for information extraction. *In Proceedings of the ACL-97 Workshop in Natural Language Learning*, pages 9–15.

CHOMSKY, N. (1969). *La linguistique cartésienne : un chapitre de l'histoire de la pensée rationaliste suivi de La nature formelle du langage*. Seuil, Paris.

CHOMSKY, N. (1970). *Current Issues in Linguistic Theory*. Mouton, The Hague & Paris.

CHOMSKY, N. (1972a). Deep structure, surface structure and semantic interpretation. *In* Chomsky [1972c], pages 62–119.

CHOMSKY, N. (1972b). Remarks on nominalization. *In* Chomsky [1972c], pages 11–61.

CHOMSKY, N. (1972c). *Studies on Semantics in Generative Grammar*. Mouton, The Hague & Paris.

CHOMSKY, N. (1976). *Aspects of the theory of syntax*. MIT Press, Cambridge.

COURTOIS, B. (1990). Un système de dictionnaires électroniques pour les mots simples du français. *Langue Française*, 87:11–22.

CRESCENZI, V., MECCA, G. et MERIALDO, P. (2001). Roadrunner : Towards automatic data extraction from large web sites. *In Proceedings of 27th International Conference on Very Large Data Bases*, pages 109–118.

CUNNINGHAM, H., MAYNARD, D., BONTCHEVA, K. et TABLAN, V. (2002). GATE : A framework and graphical development environment for robust NLP tools and applications. *In Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*.

DAILLE, B. (1995). Combined approach for terminology extraction : lexical statistics and linguistic filtering. Rapport technique, Lancaster University.

de MARNEFFE, M.-C. et DUPONT, P. (2004). Comparative study of statistical word sense discrimination techniques. *In* PURNELLE, G., FAIRON, C. et DISTER, A., éditeurs : *Le poids des mots. Actes des 7e Journées internationale d'Analyse statistique des Données textuelle (JADT'04)*, volume 1, pages 270–281. Presses universitaires de Louvain.

DISTER, A. (2001). Levée d'ambiguïtés sur les mots lexicaux et grammaticaux. *LinguisticæInvestigationes*, 24(1):105–126.

DISTER, A., FAIRON, C. et WATRIN, P. (2006). Recherche d'expressions figées du français en belgique. méthode et observations. *Cahiers de linguistique de Louvain*. à paraitre.

DOMINGUÈS, C. (1998/1999). Traitement de la coordination à l'intérieur des groupes nominaux. *LinguisticæInvestigationes*, 22:33–57.

DUNHAM, G. (1986). The role of syntax in the sublanguage of medical diagnostic statements. *In* Grishman et Kittredge [1986], pages 175–194.

DUNNING, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

EVANS, D. A., HANDERSON, S. K., LEFFERTS, R. G. et MONARCH, I. A. (1991). A summary of the CLARIT project. Rapport technique, Carnegie-Melon University.

FAIRON, C. (1999). Glossanet : Parsing a web site as a corpus. *LinguisticæInvestigationes*, 22(1):327–340.

FAIRON, C. (2006). Corporator : A tool for creating rss-based specialized corpora. *In Proceedings of the Workshop Web as corpus. EACL 2006*.

FAIRON, C. et HO, N.-D. (2004). Quantité d'information échangée : une nouvelle mesure de similarité des mots. *In* PURNELLE, G., FAIRON, C. et DISTER, A., éditeurs : *Le poids des mots. Actes des 7^{e} Journées internationale d'Analyse statistique des Données textuelle (JADT'04)*, volume 1, pages 423–433. Presses universitaires de Louvain.

FAIRON, C., PAUMIER, S. et WATRIN, P. (2005). Can we parse without tagging ? *In Proceedings of the 2nd Language & Technology Conference*, pages 473–477.

FAIRON, C. et WATRIN, P. (2003). From extraction to indexation. collecting new indexation keys by means of IE techniques. *In Proceedings of the Workshop on Finite-State Methods in Natural Language Processing (EACL-2003)*, pages 113–118.

FEDERICI, S., MONTEMAGNI, S. et PIRELLI, V. (1996). Shallow parsing and text chunking : A view on underspecification in syntax. *In Proceedings of the ESSLLI'96 Workshop on Robust Parsing*.

FELDMAN, R., ROSENFELD, B. et FRESKO, M. (2006). TEG - a hybrid approach to information extraction. *KAIS 9(1)*, 9(1):1–18.

FININ, T. W. (1986). Constraining the interpretation of nominal compounds in a limited context. *In* Grishman et Kittredge [1986], pages 163–174.

FIRTH, J. R. (1969a). Modes of meaning. *In* Firth [1969b], pages 190–215.

FIRTH, J. R. (1969b). *Papers in Linguistics. 1934–1951*. Oxford University Press, London.

FITZPATRICK, E., BACHENKO, J. et HINDLE, D. (1986). The status if telegraphic sublanguages. *In* Grishman et Kittredge [1986], pages 39–52.

FREITAG, D. (1998). *Machine Learning for Information Extraction in Informal Domains*. Thèse de doctorat, Carnegie Mellon University.

FRIEDMAN, C. (1986). Automatic structuring of sublanguage information. *In* Grishman et Kittredge [1986], pages 85–102.

FUCHS, C. (1996). *Les ambiguités du français*. Ophrys, Paris.

GAATONE, D. (1998). *Le passif en français*. Duculot, Bruxelles.

GALE, W. A., CHURCH, K. W. et YAROWSKY, D. (1992). One sense per discourse. *In Proceedings of the* 4^{th} *DARPA Speech and Natural Language Workshop*, pages 233–237.

GHADESSY, M., HENRY, A. et ROSEBERRY, R. L. (2001). *Small Corpus Studies and ELT. Theory and practice*. John Benjamins, Amsterdam & Philadelphia.

GRANGER, S. et ALTENBERG, B., éditeurs (2000). *Lexis in Contrast*. John Benjamins, Amsterdam & Philadelphia.

GREFENSTETTE, G. (1992). Sextant : Exploring unexplored contexts for semantic extraction from syntactic analysis. *In Proceedings of the* 30^{th} *Annual Meeting of the Association for Computational Linguistics (ACL'92)*, pages 324–326.

GREFENSTETTE, G. (1993). Evaluation techniques for automatic semantic extraction : Comparing syntactic and window-based approaches. Rapport technique, University of Pittsburgh.

GRISHMAN, R. (1995). The NYU system for MUC-6 or where's the syntax ? *In Proceedings of the Sixth Message Understanding Conference*, pages 167–176.

GRISHMAN, R. (1997). Information extraction : Techniques and challenges. *In* Maria Teresa Pazienza et Carbonell [1997], pages 10–27.

GRISHMAN, R. (2001). Adaptive information extraction and sublanguage analysis. *In Proceedings of the Workshop on Adaptive Text Extraction and Mining (IJCAI-2001)*.

GRISHMAN, R. et KITTREDGE, R., éditeurs (1986). *Analyzing Language in Restricted Domains : Sublanguage Description and Processing*. Lawrence Erlbaum Associates, New Jersey.

GROSS, G. (1981a). Lexicographie et grammaire. *Cahiers de lexicologie*, 39:35–46.

GROSS, G. (1994). Classes d'objets et description des verbes. *Langages*, 115:15–31.

GROSS, G. (1996). *Les expressions figées en français. Noms composés et autres locutions*. Ophrys, Paris.

GROSS, G. et GEUNTHNER, F. (2002). Comment décrire une langue de spécialité. *Cahiers de lexicologie*, 80(1):179–199.

GROSS, M. (1968). L'emploi des modèles en linguistique. *Langages*, 9:3–50.

GROSS, M. (1975). *Méthode en syntaxe : régime des constructions complétives*. Hermann, Paris.

GROSS, M. (1979). On the failure of generative grammar. *Language*, 55(4):859–885.

GROSS, M. (1981b). Les bases empiriques de la notion de prédicat sémantique. *Langages*, 63:7–52.

GROSS, M. (1982). Une classification des phrases « figées » du français. *Revue québécoise de linguistique*, 11:151–185.

GROSS, M. (1984). Lexicon-grammar and the syntactic analysis of french. *In Proceedings of the* 10^{th} *International Conference on Computational linguistics*, pages 275–282.

GROSS, M. (1986a). *Grammaire transformationnelle du français. 3 - Syntaxe de l'adverbe*. Asstril, Paris.

GROSS, M. (1986b). Methods and tactics in the construction of a lexicon-grammar. *Linguistics in the Morning Calm 2*, pages 177–197.

GROSS, M. (1988). Les limites de la phrase figée. *Langages*, 90:7–22.

GROSS, M. (1989). La construction de dictionnaires électroniques. *Annales de Télécommunications*, 44:4–19.

GROSS, M. (1990). Sur la notion harrissienne de transformation et son application au français. *Langages*, 99:39–56.

GROSS, M. (1993). Lexicon grammar : Application to french. *In* Asher [1993], pages 2195–2205.

GROSS, M. (1997). *Notes de cours d'informatique linguistique*. Université Paris 7, Paris.

GROSS, M. (2001). Les ambiguïtés. *LinguisticæInvestigationes*, 24(1):3–41.

GROSS, M. et LENTIN, A. (1967). *Notions sur les grammaires formelles*. Gauthier-Villars, Paris.

GUENTHNER, F. (1998). Constructions, classes et domaines : concepts de base pour un dictionnaire électronique de l'allemand. *Langages*, 131:45–55.

GUILLET, A. (1990). Fondements formels des classes sémantiques dans un lexique-grammaire. *Langages*, 98:70–78.

GUILLET, A. et LECLÈRE, C. (1992). *La structure des phrases simples en français : constructions transitives locatives*. Droz, Genève.

HARRIS, Z. S. (1961). *Structural Linguistics*. University of Chicago Press, Chicago.

HARRIS, Z. S. (1962). *String analysis of sentence structure*. Mouton & co, The Hague.

HARRIS, Z. S. (1964). From morpheme to utterance. *Language*, 22:161–183.

HARRIS, Z. S. (1971). *Structures mathématiques du langage*. Dunod, Paris.

HARRIS, Z. S. (1974). Discourse analysis. *Language*, 28(1):1–30.

HARRIS, Z. S. (1976a). Co-occurrence and transformation in linguistic structure. *Language*, 33(3):283–340.

HARRIS, Z. S. (1976b). *Notes du cours de syntaxe*. Seuil, Paris.

HARRIS, Z. S. (1981a). Discontinuous morphemes. *In* Hiz [1981], pages 36–44.

HARRIS, Z. S. (1981b). Distributional structure. *In* Hiz [1981], pages 3–22.

HARRIS, Z. S. (1981c). The elementary transformations. *In* Hiz [1981], pages 211–235.

HARRIS, Z. S. (1981d). Morpheme alternants in linguistic analysis. *In* Hiz [1981], pages 23–35.

HARRIS, Z. S. (1981e). Transformational theory. *In* Hiz [1981], pages 236–281.

HARRIS, Z. S. (1988). *Language and Information*. Columbia University Pres, New York.

HINDLE, D. et ROOTH, M. (1998). Structural ambiguity and lexical relations. *Computational Linguistics*, 19(1):103–120.

HIRSCHMAN, L. (1986). Discovering sublanguage structures. *In* Grishman et Kittredge [1986], pages 211–234.

HIRSCHMAN, L. et SAGER, N. (1982). Automatic information formatting of a medical sublanguage. *In* Kittredge et Lehrberger [1982], pages 27–80.

HIZ, H., éditeur (1981). *Papers on Syntax*. Reidel Publishing Company, Dordrecht.

HOBBS, J. R. (1986). Sublanguage and knowledge. *In* Grishman et Kittredge [1986], pages 53–68.

HOBBS, J. R., STICKEL, M. E., APPELT, D. E. et MARTIN, P. (1993). Interpretation as abduction. *Artificial Intelligence*, 63:69–142.

HUFFMAN, S. B. (1996). Learning information extraction patterns from examples. *Lecture Notes In Computer Science*, 1040:246–260.

HUNSTON, S. et FRANCIS, G. (2000). *Pattern Grammar. A corpus-driven approach to the lexical grammar of English*. John Benjamins, Amsterdam & Philadelphia.

IDE, N. et VÉRONIS, J. (1998). Word sense disambiguation : The state of the art. *Computational Linguistics*, 24(1):1–40.

ISABELLE, P. (1987). Machine translation at the taum group. *In* King [1987].

JACOBS, P. S., KRUPKA, G., RAU, L., MAULDIN, M., MITAMURA, T., KITANI, T., I, I. S. et CHILDS, L. (2001). GE-CMU : Description of the SHOGUN system used for MUC-5. *In Proceedings of the Fifth Message Understanding Conference*, pages 109–120.

JACOBS, P. S. et RAU, L. F. (1990). The GE NLToolset : A software foundation for intelligent text processing. *In Proceedings of the* 13^{th} *International Conference on Computational Linguistics*, pages 373–375.

JAMES D. BENSON, M. J. C. et GREAVES, W. S., éditeurs (1988). *Linguistics in a Systemic Perspective*. John Benjamins, Amsterdam & Philadelphia.

KEVERS, L. (2006). L'information biographique : modélisation, extraction et organisation en base de connaissances. *In* MERTENS, P., FAIRON, C., DISTER, A. et WATRIN, P., éditeurs : *Verbum ex machina. Actes de la* 13^{e} *conférence sur le traitement automatique des langues naturelle (TALN'06)*, volume 2 de *Cahiers du Cental*, pages 680–689. Presses universitaires de Louvain.

KIM, J. et MOLDOVAN, D. (1995). Acquisition of linguistic patterns for knowledge-based information extraction. *IEEE Transactions on Knowledge and Data Engineering*, 7(5):713–724.

KING, M., éditeur (1987). *Machine Translation Today : The State of the Art*. Edinburgh University Press.

KITTREDGE, R. (1982). Variation and homogeneity of sublanguage. *In* Kittredge et Lehrberger [1982], pages 107–137.

KITTREDGE, R. et LEHRBERGER, J. (1982). *Sublanguage. Studies of Language in Restricted Semantic Domains*. Walter de Gruyter, Berlin – New York.

KOZIMA, H., KOZIMA, A. et FURUGORI, T. (1993). Similarity between words computed by spreading activation on an english dictionary. *In Proceedings of the 6th Conference of the European Chapter of the Association for Computational Linguistics (EACL-93)*, pages 232–239.

KUSHMERICK, N., WELD, D. S. et DOORENBOS, R. B. (1997). Wrapper induction for information extraction. *In Proceedings of the Intl. Joint Conference on Artificial Intelligence (IJCAI)*, pages 729–737.

LAMIROY, B. (1998). Le lexique-grammaire. essai de synthèse. *Travaux de linguistique*, 37:7–23.

LAMIROY, B., KLEIN, J., LECLÈRE, C. et LABELLE, J. (2003). Expressions verbales figées dans quatre variétés du français : le projet BFQS. *Cahiers de lexicologie*, 83:153–172.

LAPORTE, E. (2001). Reduction of lexical ambiguity. *LinguisticæInvestigationes*, 24(1):67–103.

LAPORTE, E. et MONCEAUX, A. (1999). Elimination of lexical ambiguities by grammars. the elag system. *LinguisticæInvestigationes*, 22:341–367.

LECLÈRE, C. (1978). Sur une classe de verbes datifs. *Langue française*, 39:66–75.

LECLÈRE, C. (1990). Organisation du lexique-grammaire des verbes français. *Langue française*, 86:112–122.

LECLÈRE, C. (1998). Travaux récents en lexique-grammaire. *Revue internationale de linguistique française*, 37:155–186.

LECLÈRE, C. (1999). Une approche syntaxique de la synonymie. *Cahiers de l'institut de linguistique de Louvain*, 2:77–94.

LEFEUVRE, F. (1999). *La phrase averbale en français*. L'Harmattan, Paris/Montréal.

LEHRBERGER, J. (1982). Automatic translation and the concept of sublanguage. *In* Kittredge et Lehrberger [1982].

LEHRBERGER, J. (1986). Sublanguage analysis. *In* Grishman et Kittredge [1986], pages 19–38.

LENHERT, W. G. (1991). Symbolic-subsymbolic sentence analysis : Exploiting the best of two worlds. *In Proceedings of the Third Message Understanding Conference*, pages 135–164. Morgan Kaufmann.

LESK, M. (1997). Automatic sense disambiguation : How to tell a pine cone from an ice cream cone. *In Proceedings of the 1986 SIGDOC Conference*, pages 24–26. Association for Computing Machinery.

MANNING, C. D. et SCHÜTZE, H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge.

MARIA TERESA PAZIENZA, J. S. et CARBONELL, J. G., éditeurs (1997). *Information Extraction : A Multidisciplinary Approach to an Emerging Information Technology*, volume 1299 de *Lecture Notes In Computer Science*. Springer-Verlag, London.

MARSH, E. (1986). General semantic patterns in different sublanguages. *In* Grishman et Kittredge [1986], pages 103–128.

MATHIEU-COLAS, M. (1996). Essai de typologie des noms composé français. *Cahiers de lexicologie*, 69(2):71–125.

MAUREL, D. (1989). *Reconnaissance de séquences de mots par automate : adverbes de date du français*. Thèse de doctorat, Université Paris 7.

MAUREL, D. (2004). Les mots inconnus sont-ils des noms propres ? *In* PURNELLE, G., FAIRON, C. et DISTER, A., éditeurs : *Le poids des mots. Actes des* 7^e *Journées internationale d'Analyse statistique des Données textuelle (JADT'04)*, volume 2, pages 776–784. Presses universitaires de Louvain.

MAUREL, D. et FRIBURGER, N. (2002). Finite-state transducer cascade to extract proper names in texts. *In Proceedings of th* 6^{th} *International Conference on Implementation and Application of Automata (CIAA-2001)*, pages 115–124.

MCDONALD, D. (1996). Internal and external evidence in the identification and semantic categorization of proper names. *In* Boguraev et Pustejovsky [1996], pages 21–39.

MEL'CUK, I., ARBATCHEWSKY-JUMARIE, N., ELNITSKY, L., IORDANSKAJA, L. et LESSARD, A. (1984). *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques*, volume 1. Les presses de l'Université de Montréal, Montréal.

MILLER, G. A. (1995). WordNet : a lexical database for English. *Communications of the ACM*, 38(11):39–41.

MITKOV, R. (2002). *Anaphor Resolution*. Longman.

MONTGOMERY, C. A. et GLOVER, B. C. (1986). A sublanguage for reporting and analysis of space envents. *In* Grishman et Kittredge [1986], pages 129–162.

MOON, R. (1987). The analysis of meaning. *In* Sinclair [1987b], pages 86–103.

NAKAMURA, T. (2004). Analyse automatique d'un discours spécialisé au moyen de grammaires locales. *In* PURNELLE, G., FAIRON, C. et DISTER, A., éditeurs : *Le poids des mots. Actes des 7^e Journées internationale d'Analyse statistique des Données textuelle (JADT'04)*, volume 1, pages 837–847. Presses universitaires de Louvain.

NEVIN, B. E. et JOHNSON, S. M., éditeurs (2002). *The Legacy of Zellig Harris*. John Benjamins, Amsterdam & Philadelphia.

ONYSHKEVYCH, B., OKUROWSKI, M. E. et CARLSON, L. (1993). Tasks, domains, and languages. *In Proceedings of the fifth Message Understanding Conference (MUC-5)*, pages 7–17.

PAUMIER, S. (2003). *De la reconnaissance de formes linguistiques à l'analyse syntaxique*. Thèse de doctorat, Université de Marne-la-Vallée.

PAUMIER, S. (2006). *Unitex 1.2. Manuel d'utilisation*.

PEREIRA, F. (2002). Formal grammar and information theory. *In* Nevin et Johnson [2002], pages 13–32.

PESANT, D. L. et MATHIEU-COLAS, M. (1998). Introduction aux classes d'objets. *Langages*, 131:6–33.

PHILLIPS, M. (1985). *Aspects of Text Structure : An investigation of the lexical organisation of text*. Elsevier, Amsterdam.

POIBEAU, T. (2003). *Extraction d'information : du texte brut au web sémantique*. Hermès, Paris.

RAYSON, P. et GARSIDE, R. (2000). Comparing corpora using frequency profiling. *In Proceedings of the workshop on Comparing Corpora*, pages 1–6.

RENOUF, A. (1993). A word in time : First findings from the investigation of dynamic text. *In Papers from the thirteenth International Conference on English Language Research on Computerized Corpora*, pages 279–288.

REY-DEBOVE, J. et REY, A. (1994). *Le nouveau petit Robert. Dictionnaire alphabétique et analogique de la langue française*. Dictionnaires Le Robert, Paris.

RILOFF, E. (1993). Automatically constructing a dictionary for information extraction tasks. *In Proceedings of the Eleventh Annual Conference on Artificial Intelligence*, pages 811–816.

RILOFF, E. (1996). Automatically generating extraction patterns from untagged text. *In Proceedings of the Thirteenth Annual Conference on Artificial Intelligence*, pages 1044–1049.

RILOFF, E. et JONES, R. (1999). Learning dictionaries for information extraction by multi-level bootstrapping. *In Proceedings of the* 16^{th} *National Conference on Artificial Intelligence (AAAI-1999)*, pages 474–479.

ROCHE, E. (1993). Une représentation par automate fini des textes et des propriétés transformationnelles des verbes. *LinguisticæInvestigationes*, 17(1):189–222.

ROMAINE, S. (1993). Empiricism. *The Encyclopedia of Language and Linguistics*, 3:1111–1114.

SAGER, N. (1982). Syntactic formatting of science information. *In* Kittredge et Lehrberger [1982], pages 9–26.

SAGER, N. (1986). Sublanguage : Linguistic phenomenon, computational tool. *In* Grishman et Kittredge [1986], pages 1–18.

SAGER, N. et NHÀN, N. T. (2002). The computability of strings, transformations, and sublanguage. *In* Nevin et Johnson [2002], pages 79–120.

SALKOFF, M. (1979). *Analyse syntaxique du Français : Grammaire en chaîne*. John Benjamins, Amsterdam & Philadelphia.

SCHMID, H. (1994). Probabilistic part-of-speech tagging using decision trees. *In International Conference on New Methods in Language Processing*, Manchester, UK.

SENELLART, J. (1999). *Outils de reconnaissance d'expressions linguistiques complexes dans des grands corpus*. Thèse de doctorat, Université Paris 7.

SERETAN, V., NERIMA, L. et WEHRLI, E. (2003). Extraction of multi-word collocations using syntactic bigram composition. *In Proceedings of th* 4^{th} *International Conference on Recent Advances in NLP (RANLP-2003)*, pages 424–431.

SILBERZTEIN, M. D. (1993). *Dictionnaires électroniques et analyse automatique de textes. Le système INTEX*. Masson, Paris.

SINCLAIR, J. M. (1987a). Grammar in the dictionary. *In* Sinclair [1987b], pages 104–115.

SINCLAIR, J. M., éditeur (1987b). *Looking Up. An account of the COBUILD Project in lexical computing*. Collins Cobuild, London & Glasgow.

SINCLAIR, J. M. (1987c). The nature of evidence. *In* Sinclair [1987b], pages 150–159.

SINCLAIR, J. M. (1988). Sense and structure in lexis. *In* James D. Benson et Greaves [1988], pages 71–97.

SINCLAIR, J. M. (1997). *Corpus, Concordance, Collocation*. Oxford University Press, Oxford.

SLOCUM, J. (1986). How one might automatically identify and adapt to a sublanguage : An initial exploration. *In* Grishman et Kittredge [1986], pages 195–210.

SMADJA, F. (1993). Retrieving Collocations from Text : Xtract. *Computational Linguistics*, 19(1):143–177.

SMADJA, F., HATZIVASSILOGLOU, V. et MCKEOWN, K. R. (1996). Translating Collocations for Bilingual Lexicons : A Statistical Approach. *Computational Linguistics*, 22(1):2–38.

SODERLAND, S. (1999). Learning information extraction rules for semi-structured and free text. *Machine Learning*, 34(1-3):233–272.

SODERLAND, S., FISHER, D., ASELTINE, J. et LEHNERT, W. (1995). CRYSTAL : Inducing a conceptual dictionary. *In Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pages 1314–1319.

SPAGGIARI, L., BEAUJARD, F. et CANNESSON, E. (2005). A controlled language at airbus. *LinguisticæInvestigationes*, 28(1):107–122.

SPYNS, P., NHÀN, N. T., BAERT, E., SAGER, N. et MOOR, G. D. (1998). Medical language processing applied to extract clinical information from dutch medical documents. *In Proceedings of the* 9^{th} *World Congress on Medical Informatics (MEDINFO'98)*, pages 685–689.

SUDO, K., SEKINE, S. et GRISHMAN, R. (2004). Cross-lingual information extraction system evaluation. *In Proceedings of the* 20^{th} *International Conference on Computational Linguistics*.

WALKER, D. E. et AMSLER, R. A. (1986). The use of machine-readable dictionaries in sublanguage analysis. *In* Grishman et Kittredge [1986], pages 69–84.

WEBSTER (2000). The online plain text english dictionary.

WEHRLI, E. (1997). *L'analyse syntaxique des langues naturelles : problèmes et méthodes*. Masson, Paris.

WEISCHEDEL, R. (1995). BBN : Description of the PLUM system as used for MUC-6. *In Proceedings of the Sixth Message Understanding Conference*, pages 55–69.

YANGARBER, R. et GRISHMAN, R. (1998). NYU : Description of the proteus/PET system as used for MUC-7 ST. *In Proceedings of the Seventh Message Understanding Conference*.

YANGARBER, R., GRISHMAN, R., TAPANAINEN, P. et HUTTUNEN, S. (2000). Automatic acquisition of domain knowledge for information extraction. *In Proceedings of the* 18^{th} *International Conference on Computational Linguistics (COLING 2000)*, pages 940–946.

YAROWSKY, D. (1992). Word-sense disambiguation using statistical models of roget's categories trained on large corpora. *In Proceedings of the* 33^{th} *Annual Meeting of the Association for Computational Linguistics*, pages 189–196.

YAROWSKY, D. (1993). One sense per collocation. *In Proceedings of the ARPA Human Language Technology Workshop*, pages 266–271.