Université Paris 7 – Denis Diderot
École Doctorale Sciences du Langage

Année :

N° attribué par la bibliothèque

THÈSE

Pour l'obtention du Diplôme de
Docteur de l'université Paris 7
en Linguistique

Construction et exploitation d'un corpus syntaxiquement annoté pour le français

Présentée et soutenue publiquement le 21 juin 2001
par

Lionel Clément

Jury :
Anne Abeillé – Professeur Université Paris 7 (Directrice)
Claire Blanche Benveniste – Professeur Université de Provence (examinatrice)
Benoît Habert – Professeur Université Paris X Nanterre (examinateur)
Laurent Romary – Chargé de recherche LORIA (Rapporteur)
Jean Véronis – Professeur Université de Provence (Rapporteur)

Remerciements

Je tiens à remercier Anne Abeillé d'avoir dirigé ce travail. Beaucoup de choix théoriques et d'idées présentés ici lui reviennent alors qu'elle me faisait redécouvrir ou deviner discrètement leurs motivations. C'est au fond une belle initiation que celle-là. Je la remercie particulièrement pour son soutien, ses compétences, sa rigueur et son infaillible patience.

Je garderai toujours un grand souvenir de l'accueil qui m'a été réservé par Jean Véronis et son équipe au CILSH (Centre Informatique pour Lettres et Sciences Humaines) de l'Université de Provence. J'ai passé une année riche et agréable durant cette première année d'enseignant-chercheur à Aix-en-Provence. Merci pour les discussions et les nombreux commentaires qui m'ont permis d'améliorer ce travail.

Laurent Romary m'a également beaucoup aidé en corrigeant et critiquant une première version de cette thèse ; je l'en remercie chaleureusement.

Je remercie Benoît Habert et Claire Blanche-Benveniste d'avoir accepté d'examiner ce travail et d'avoir fait partie du jury.

Merci enfin à Laurence Danlos qui m'a accueilli depuis tant d'années dans le laboratoire TALaNa. Je dois beaucoup à la qualité scientifique de ce laboratoire qui embrasse de nombreux aspects du TAL et à la qualité des personnes qui y travaillent.

Merci à l'équipe de TALaNa, Alexandra Kinyon, Marianne Desmet, Kim Gerdes, Karen Ferret, Rodrigo Reyes, Moustafa Krasem et de nombreux autres collaborateurs avec qui j'ai partagé des idées en griffonnant sur des nappes de papier.

Le projet présenté dans ce mémoire n'aurait pas été réalisable sans le concours averti de nombreux étudiants stagiaires.

Ces stagiaires étaient : Antonio Balvet, Nicolas Barrier, Sébastien Barrier, Grégory Bichot, Elizabeth Bresson, Nathalie Briot, Martine Cheradame, Vanessa Combet, Muriel Delahaye, Morgane Erenati, Véronique Gendner, Nathalie Lafon, Mariamma Leib, Éric Marty, Nicolas Montessuit, Virginie Nanta, Marie Pasquier, Louis-Gabriel Pouillot, François Toussenel et Isabelle Znamenski.

Qu'ils soient remerciés de m'avoir supporté.

Enfin merci aux membres de ma famille pour leur soutien de tous les instants.

Table des matières

Introduction

La linguistique de corpus : une nouvelle discipline ?

La linguistique de corpus semble émerger comme une discipline à part entière depuis quelques décennies à tel point qu'une littérature lui est nouvellement attribuée, y compris en langue française depuis quelques années ([Habert *et al.*, 1997], [Bilger, 2000a], revue *TAL* 1995, *Revue Française de Linguistique Appliquée* 1996 et 1999), ce qui ne fait que rattraper un retard.

L'utilisation des sources textuelles informatisées comme Frantext, et surtout l'enrichissement systématique de ces données avec des informations linguistiques internes ou externes et l'usage d'outils informatiques permettant d'exploiter ces données semblent modifier la pratique des linguistes.

On écartera cependant rapidement l'idée que cela soit une résurgence du structuralisme qui dominait la linguistique américaine depuis les travaux de L. Bloomfield jusqu'aux critiques de N. Chomsky. Le principe d'immanence adopté par le structuralisme américain conférait aux corpus l'objectif même de l'étude linguistique. Le corpus était par définition clos, et devait induire, pour autant qu'il soit représentatif, toutes les propriétés de la langue. Dans cette définition précise, sa représentativité devraient être telle qu'une théorie du corpus soit une théorie de la langue.

Les travaux récents en science du langage ne supposent pas cette définition stricte du corpus. L'étude empirique des faits de langue porte sur des exemples inventés grammaticaux ou agrammaticaux jugés par des informateurs ou sur un *corpus* vu comme une source d'informations. Pour autant qu'elle soit empirique, la science du langage trouve sa falsifiabilité dans l'agrammaticalité ou le contre-exemple. Le corpus ne peut pas être l'objet même de la science du langage dans cette conception, il n'informe que sur l'attesté ou le contre-

exemple mais ne peut en aucune manière prédire l'agrammatical. De plus, si le corpus informe sur l'attesté, il n'informe en rien sur le possible de langue. Il n'informera donc pas sur la compétence du locuteur. Mais il peut apporter des informations pertinentes sur la fréquence relative des constructions et sur leur disponibilité. Il permet par exemple de savoir si elles sont spécialisées pour un petit nombre de mots ou si elles sont productives ([Blanche-Benveniste, 1996]).

La discussion portant sur la légitimité de l'exemple construit plutôt que de l'exemple choisi dans les textes n'est pas nouvelle. La grammaire de Port-Royal et ses successeurs ont fortement été contestés (par F. Brunot, Damourette & Pichon, Grévisse par exemple) pour ne pas avoir été construits grâce à des exemples choisis dans la littérature dont la valeur normative rend *indiscutable* l'usage. On retrouve cette discussion avec l'utilisation qui est faite des corpus représentatifs. Elle ne porte pas sur la valeur normative d'une grammaire comme c'était le cas dans le passé mais sur la qualité de la description d'une langue. Un corpus du français parlé comme celui constitué par le GARS (Groupe Aixois de Recherche en Syntaxe [Blanche-Benveniste *et al.*, 1991]) informe le linguiste autrement que ne pourrait le faire son intuition ou un locuteur natif sur les propriétés de ce type de production langagière. A la qualité normative de l'exemple littéraire, on peut rapporter la qualité de la description linguistique issue de corpus.

Le recours aux corpus n'a réellement jamais cessé comme source d'information dans les années 1960-1980 même si les partisans de cet empirisme se sont fait opposer un faux procès d'intention, celui du retour au positivisme.

En langue anglaise, le *Brown Corpus* a été construit dès 1961 par W.N. Francis et H. Kučera aux États Unis, le *Survey of English Usage* en 1959 par R. Quirk en Grande-Bretagne. En France, bien avant le lancement de la construction de gros corpus destinés à l'édition de dictionnaires de langue anglaise des années 1980 (*Cobuild Corpus*), l'INaLF a constitué une base de textes littéraires gigantesque (160 millions de mots) pour construire le *Trésor de la Langue Française* (T.L.F.). Les textes ont été compilés depuis les années 60 et ont été informatisés dans la base Frantext.

Ce qui est nouveau aujourd'hui, c'est l'abondance d'informations que contiennent les corpus informatisés récents. Ces corpus deviennent de plus en plus représentatifs par leur taille mais également par leur qualité d'échantillonnage ou d'annotation. Ils donnent aux linguistes une nouvelle mesure de l'attesté en rapport au possible de langue. Sans prétendre jamais à l'exhaustivité, les corpus informatisés récents peuvent montrer la productivité d'une configuration. Les résultats statistiques sur ces corpus rendent légitimes des

hypothèses que l'on n'aurait pas pu émettre par la seule intuition.

La linguistique de corpus, au sens où nous l'entendons, n'est pas une théorie linguistique mais une technologie construisant ou exploitant de nouvelles ressources. Les méthodes d'investigation ont certainement changé et les résultats sont différents mais la science du langage, qu'elle porte ou non sur l'étude de corpus, ne change pas son objet : la modélisation et la description des langues.

Linguistique de corpus et TAL

Les corpus annotés, dont la construction est issue comme nous le verrons des techniques du Traitement Automatique des Langues, lui reviennent comme source première. Ainsi, les corpus informatisés peuvent faire l'objet d'études statistiques pour entraîner des systèmes stochastiques. Les systèmes de reconnaissance vocale, d'étiquetage automatique, d'analyse syntaxique automatique par exemple utilisent des données qui ont été recueillies sur des corpus volumineux. Ils peuvent donc induire des grammaires, mais entendons que ces grammaires ne sont pas des théories linguistiques, il n'y a aucun principe d'immanence ; la grammaire induite est le résultat de l'application d'une théorie sur un corpus, rien d'autre.

Les corpus annotés servent également à évaluer la qualité de tels systèmes. Il faut pour cela qu'ils aient été corrigés à la main afin de les confronter avec les résultats attendus.

La construction d'un corpus annoté revient au Traitement Automatique des Langues aussi comme expérience d'analyse automatique. La technologie mise en œuvre pour annoter ou échantillonner un corpus suppose, pour l'économie du processus, un ensemble de traitements automatiques portant sur des productions langagières dont on ne sait rien *a priori*. Il s'agit donc d'une procédure d'analyse *robuste*.

Enfin, pour que les données soient exploitables, il faut qu'elles soient normées. C'est-à-dire qu'un corpus informatisé de grande taille doit être constitué grâce à un standard d'annotation permettant sa diffusion, sa cohérence et son utilisation sur des procédures automatiques. Cette norme suppose une formalisation des informations internes et externes. Nous verrons en 4.1.4 une proposition de norme allant dans le sens d'une telle formalisation : celle de la TEI (*Text Encoding Initiative*). Signalons également le projet Genelex comme proposition de norme pour l'annotation morpho-syntaxique, syntaxique et sémantique ainsi que le projet Multext/Grace.

Gros corpus et corpus représentatifs

Il est très aisé actuellement de compiler une collection de textes d'une langue comme le français et d'y apporter différentes sortes d'informations internes comme la lemmatisation, l'étiquetage morpho-syntaxique ou différents types d'analyses syntaxiques de surface. Toutes ces opérations d'annotation étant entièrement automatisables, on peut construire à faible coût un corpus de plusieurs centaines de millions de mots annotés sans aucune révision manuelle. Bien sûr, un tel corpus contient un taux d'erreurs correspondant à ce qu'affichent les différents outils utilisés (tagger, lemmatiseurs, *shallow-parser*, etc.). Les étiqueteurs et lemmatiseurs affichent un taux d'erreur proche de 95% par mot avec un jeu relativement réduit d'étiquettes, les *shallow-parsers* et *chunkers* des taux variables avoisinant 90%.

En revanche, la correction d'un corpus annoté est très coûteuse puisque les erreurs des systèmes automatiques ne sont généralement pas prédictibles et il faut une inspection longitudinale de tout le corpus pour les identifier.

La correction d'un corpus d'un million de mots en morpho-syntaxe s'estime à quelques 4 hommes-années, sa correction en syntagmes environ à la même quantité de travail. Nous devons donc expliquer les motivations d'une telle entreprise qui ne peut produire que des corpus peu volumineux (au regard des volumes de textes non annotés disponibles) pour un coût élevé.

La qualité d'un corpus annoté mais non corrigé est discutable. Ou plus exactement l'apport de l'annotation dans ce cas est discutable. Les régularités observables d'un tel corpus ne fournissent guère que les connaissances linguistiques qui ont été mises en œuvre lors de l'annotation automatique. Même si la qualité de l'annotation peut sembler suffisante pour une étude statistique (ce qui apparaît avec des scores proches de 100%[1]), les erreurs résiduelles sont cruciales puisqu'elles sont précisément présentes là où le modèle formel du système automatique n'a pu rendre compte avec succès d'un phénomène linguistique.

En revanche, les connaissances apportées lors de la correction et la validation humaine de l'annotation portent précisément sur ces cas et peuvent enrichir les études linguistiques.

1. Nous verrons cependant que ces scores élevés cachent la réalité comme le souligne [Abeillé, 1996]. Avec 98% de mots bien analysés dans un texte dont les phrases ont en moyenne 25 mots, seulement une phrase sur deux possède une analyse correcte.

Types d'annotation

L'annotation d'un corpus porte sur deux types d'informations : les informations *externes* au texte et les informations *internes*. Les informations externes sont relatives à la manière par laquelle le texte a été produit (le lieu, la date, le type d'enregistrement, la méthode de retranscription, l'âge et l'origine du locuteur, etc.). Elles sont censées informer l'utilisateur du corpus pour des études diachroniques, sociolinguistiques, ou toute autre étude linguistique exploitant directement le type de production langagière. Les informations internes portent sur une description linguistique des éléments du discours (la segmentation en unités, l'étiquetage de ces unités, leur articulations syntaxiques ou sémantiques, etc.). Les deux types d'informations servent à classer par genres des échantillons de corpus ([Biber, 1991]). Les informations internes servent plus spécifiquement à l'étude empirique des régularités de la langue, ou ainsi que nous l'avons déjà dit de source pour induire des grammaires.

Les informations internes se limitent généralement à l'annotation morpho-syntaxique (nous dirons également *l'étiquetage*), la lemmatisation, l'annotation syntaxique (en constituance et dépendance), l'annotation sémantique et plus marginalement à l'annotation des anaphores et des *traces*.

Nous nous focaliserons dans le présent mémoire sur l'étiquetage morpho-syntaxique d'un corpus français.

Types d'interrogation

L'étude empirique d'un phénomène linguistique grâce à un corpus étiqueté commence par le dépouillement des données. La tâche est grandement simplifiée depuis l'usage de l'outil informatique par les linguistes bien que la méthode soit la même que celle qui était éprouvée *à la main*. L'outil informatique et la qualité des corpus annotés permettent de meilleures réponses à des questions classiques et permettent de poser de nouvelles questions. Les linguistes qui travaillent sur corpus informatisés utilisent des filtres, c'est-à-dire des programmes qui sélectionnent l'ensemble des occurrences d'un texte qui subsument une spécification d'un mot ou de toute autre unité de texte. Les résultats peuvent alors être présentés selon l'étude menée pour lire l'ensemble des concordances des occurrences trouvées (l'outil s'appelle alors un «concordancier») ou sous forme de mesures à l'origine de résultats statistiques.

Le type d'interrogation d'un corpus syntaxiquement ou sémantiquement

annoté n'est techniquement guère différent, il s'agit toujours de procéder à la recherche d'une spécification (correspondant à une figure syntaxique par exemple ou à un contexte précis) pour connaître les occurrences (ou leurs quantités, leurs propriétés). On peut s'intéresser par exemple à la longueur d'un groupe nominal qui est modifié par une relative, aux types d'adverbes que l'on trouve en tête de phrase, aux types d'adverbiales qu'on trouve entre le verbe et ses compléments, etc. L'outil informatique doit alors s'adapter au type de formalisation de l'annotation. Le filtre doit s'appliquer sur une suite linéaire de mots et/ou d'étiquettes mais également sur des arbres, des treillis, des forêts ou encore des graphes acycliques. La formalisation de l'annotation du corpus doit être conduite par une réflexion sur la faisabilité et le coût d'un tel filtre. Par exemple, le *Penn TreeBank*, un corpus annoté pour la syntaxe que nous présentons succinctement en 1.1.6 est associé à un outil d'interrogation (tgrep) qui permet de filtrer les descriptions syntagmatiques sous forme d'arbres telles qu'elles ont été représentées dans le corpus.

Plan

Dans le premier chapitre, nous présenterons les principaux corpus informatisés bruts et annotés existants pour l'anglais et le français qui intéressent notre étude. Puis nous aborderons différents aspects de la construction des corpus annotés en Traitement Automatique des Langues.

Le second chapitre présentera le projet du Corpus Annoté de Paris 7 dans son ensemble. Nous exposerons les objectifs de l'entreprise mais également les méthodes et les choix retenus pour l'annotation.

Le troisième chapitre passera en revue les choix linguistiques de segmentation et d'annotation du corpus. On y abordera les définitions des catégories retenues puis quelques cas délicats comme les frontières floues entre les participes et les adjectifs ou encore les catégories retenues pour une liste de mots difficiles.

Le quatrième chapitre porte sur les choix informatiques. Nous verrons quels types d'algorithmes ont été choisis pour différentes procédures automatiques et comment les annotations du corpus doivent être encodées pour être compatibles avec ces procédures. Nous exposerons également quelques implémentations et critiques de programmes existants.

Le cinquième chapitre expose des interrogations portant sur des phénomènes linguistiques qui ont été faites sur le Corpus Annoté de Paris 7. Nous verrons en quoi les informations contenues dans ce corpus permettent d'ob-

tenir des résultats intéressants. Nous verrons que l'informatisation du corpus permet une grande économie de moyens et que des explorations intéressantes du corpus sont possibles alors qu'elles n'étaient pas envisageables par un dépouillement manuel des données.

Le sixième chapitre aborde la suite du projet : l'annotation syntaxique du corpus. Nous présentons d'abord une expérience personnelle d'annotation grâce à des grammaires locales, puis nous exposons l'annotation en syntagmes proprement dite. Comme pour l'annotation morpho-syntaxique, nous présentons les méthodes et choix retenues avant d'aborder quelques difficultés connues comme l'annotation des catégories vides ou des syntagmes *sans tête*.

Chapitre 1

État de l'art

Les corpus informatisés existent depuis le début des années 1960 tant pour l'anglais que pour le français. Le projet de l'INaLF de constitution d'un corpus littéraire volumineux pour informer le Trésor de la Langue Française (TLF) a donné lieu à une base informatisée de textes (Frantext). Ces textes sont principalement des romans du XX^{e} et début du XXI^{e} siècle. Construits au même moment, les corpus d'anglais (*The Survey of English Usage*, *The survey of Spoken English*) étaient motivés par la même démarche : constituer une base représentative (et normative pour le TLF) de la langue en vue de renseigner une étude fondée sur corpus. Les lexicographes, premiers utilisateurs de ces corpus, voyaient dans ces bases textuelles la matière à l'exhaustivité de l'usage, la couverture de tous les régionalismes et genres de la langue.

Cet aspect de la linguistique de corpus a été suivi de techniques d'échantillonnage des corpus. Le but était de produire des corpus représentatifs d'un genre ou d'une variété régionale sans prétendre à l'impossible exhaustivité.

L'informatisation (qui n'a commencé que dans les années 1970 et a vraiment éclaté dans les années 1980) a contribué à faciliter la lourde tâche de constitution des corpus (aujourd'hui la préparation à la numérisation d'un livre est l'étape la plus longue de sa mise au format électronique). L'informatisation contribue également à la diffusion de ces corpus et à leur exploitation par des procédures automatiques.

Les corpus annotés disponibles pour l'écrit

Les corpus n'ont commencé à être annotés que plus récemment. Plusieurs projets sont en cours actuellement pour des langues variées (cf. [Abeillé *et al.*, 2000a]). Nous allons présenter simplement les corpus disponibles pour l'anglais et le français.

1.1 Les premiers corpus électroniques annotés

Pour une présentation de l'histoire de la linguistique de corpus, on peut citer une introduction par Tony McEnery et Andrew Wilson [McEnery & Wilson, 1996]. Un historique détaillé est dressé dans *Text and Corpus Analysis* par Leech [Leech, 1991].

1.1.1 Le Brown corpus

Souvent cité comme première expérience de corpus linguistique annoté de grande envergure, Le *Brown Corpus* (*Brown University Standard Corpus of Present-Day American English*) est un corpus d'un million de mots en anglais américain développé à l'université de Brown en 1964 par W.N. Francis et H. Kučera [Francis & Kučera, 1989].

Il est constitué de 500 extraits de 2000 mots chacun provenant de textes publiés en 1961 aux États-Unis d'Amérique. Les auteurs du Corpus devaient constituer une base textuelle représentative de l'anglais américain écrit contemporain. 18 genres différents de prose écrite publiée (presse, écrits scientifiques, romans, biographies, etc.) ont été sélectionnés et l'ensemble est censé couvrir l'anglais américain écrit d'où l'appellation de "corpus de référence" donnée parfois à ce type det ressource linguistique.

Le Brown Corpus a successivement été revisé entre 1971 et 1979. Lors de ces refontes, une version étiquetée semi-automatiquement (Brown Corpus version C) marque chaque mot comme appartenant à l'une des 81 classes suivantes :

- Les parties du discours en distinguant les classes ouvertes (nom commun, nom propre, verbe, adjectif et adverbe) des classes fermées (déterminant, préposition, conjonction, pronom).

- Quelques mots distingués (*not*, there, particule verbale *to*, les formes verbales *do*, *be* et *have* aaa)
- Les ponctuations
- Quelques traits morphologiques (marque du pluriel et du possessif sur les noms, comparatif et superlatif sur les adjectifs, marque du passé, du participe, de la troisième personne ou du gérondif sur les verbes)

La mise à disposition de ce corpus informatisé dans le domaine public a largement contribué au renouveau de la linguistique de corpus et à l'exploitation nouvelle de l'informatique dans ce domaine au début des années 80. A cet égard, le retard de la linguistique de corpus français est certainement lié à une moins grande diffusion des ressources quand elles existent.

Il est remarquable que ce corpus ait été constitué précisément lors de l'émergence des théories génératives. La volonté de compiler une source textuelle représentative de l'état d'une langue se heurtait à la conception contemporaine du modèle de la compétence de Chomsky [Chomsky, 1957].

1.1.2 LOB

Le *Lancester Oslo Bergen Corpus* est l'équivalent en anglais britanique du Brown Corpus. Il comprend également 1 million de mots étiquetés extraits de textes édités en 1961. Il a été conçu, lui aussi, comme un *instantané* de l'anglais écrit publié au début des années 60. La compilation des textes à été faite selon les mêmes critères, en sélectionnant 15 catégories (ou "genres littéraires") du *Brown Corpus* représentatives de l'anglais publié.

Le corpus a été construit entre 1970 et 1978 à l'Université de Lancaster et à l'université d'Oslo en collaboration avec le *Norwegian Computing Centre for the Humanities at Bergen.*

LOB a donc bénéficié de l'expérience du *Brown Corpus* et a eu l'avantage des avancées techniques de l'informatique depuis cette première expérience.

Tout comme le *Brown Corpus*, le *Lancester Oslo Bergen Corpus* a été soigneusement étiqueté avec un jeu d'étiquettes marquant les parties du discours, la morphologie et quelques sous-catégories ou classes distributives comme les pré-déterminants ou pré-quantifieurs. Quelques classes fermées de mots comme des locutions fonctionnelles *in order to, in that, so as to, as to, in spite of, etc.*, les particules verbales et des formes distinguées principalement verbales (*does, did, do, have*, etc.) sont marquées comme telles.

Le jeu d'étiquettes du *Lancester Oslo Bergen Corpus* est assez hétérogène :

y sont codées et distinguées des informations graphiques, morphologiques, syntaxiques, distributionnelles, et sémantiques.

1.1.3 Cobuild - Bank of English

Le *Cobuild Corpus* ou *Bank of English* est un très gros corpus qui a été compilé dans le but de servir de base de connaissance à un dictionnaire. Le projet a été conduit en 1980 à l'université de Birmingham par l'équipe de John Sinclair (COBUILD), en collaboration avec l'éditeur de dictionnaires Collins.

Les textes du *Corpus Cobuild* n'ont pas été choisis au hasard mais avec précision pour qu'ils puissent être représentatifs de l'anglais britannique et américain avec une répartition des genres littéraires, de l'oral et de l'écrit, du sexe et de l'âge des auteurs. L'expérience historique du Brown Corpus a donc été exploitée pour ce projet.

Le but de ce corpus était d'améliorer les descriptions lexicographiques en exploitant les variations des contextes de chaque mot. La quantité de textes devait être importante et le *Cobuild Corpus* est souvent cité comme première expérience de très gros corpus (*mega-corpus*).

Le corpus avait atteint la taille de vingt millions de mots d'anglais auxquels s'ajoutaient vingt millions de mots de textes spécialisés. Aujourd'hui, le *Bank of English* comporte 200 Millions de mots ([Järvinen, 2000]).

1.1.4 British National Corpus

Le *British National Corpus* (BNC) contient cent millions de mots étiquetés.

Il a été conçu dans les années 1995 par un consortium constitué de plusieurs éditeurs et centres de recherche des universités de Lancaster et d'Oxford.

Le but du BNC était de proposer un corpus multi-usages à la communauté linguistique. Le corpus est représentatif de l'anglais contemporain écrit et en partie oral (10%), balisé avec la norme TEI et étiqueté avec des catégories morpho-syntaxiques. Le BNC bénéficie des avancées de l'informatique de réseau puisqu'il est consultable sur Internet.

Ce corpus a été intégralement étiqueté automatiquement pour la morpho-syntaxe avec un taux d'erreurs de 2%. 61 étiquettes morpho-syntaxiques ont été retenues.

1.1.5 Susanne

Extrait du Brown Corpus, le Corpus Susanne (*Geoffrey Sampson's Suzanne*) [Sampson, 1994] de 128 000 mots a été annoté manuellement avec des informations morpho-syntaxiques et syntaxiques. Chaque phrase du corpus est assortie d'une description syntaxique détaillée contenant les constituants nuls et des indices coréférenciels.

La description de la typologie utilisée qui accompagne le corpus tient une grande place dans le projet. Idéalement, celle-ci devrait permettre que deux annotateurs fassent exactement les mêmes descriptions sur les même textes sans se concerter. Ainsi, les *guidelines* du Corpus Susanne sont moins des descriptions théoriques des catégories ou une motivation linguistique de la typologie employée que des modèles décrivant exhaustivement les choix empiriques proposés pour l'ensemble du corpus.

1.1.6 Penn Treebank

Ce corpus contient quatre millions de mots issus du *Brown Corpus* mais aussi des sources diverses d'anglais écrit (extraits du *Dow Jones News service*, *Wall Street Journal*) et d'oral retranscrit. L'intégralité du corpus a été étiqueté et arboré. Le corpus complet a été corrigé manuellement.

Le corpus est étiqueté pour la morphosyntaxe, segmenté en phrases, et chaque phrase est représentée par un arbre encodant les enchâssements syntagmatiques. Quelques fonctions grammaticales sont notées pour les syntagmes. Pour assigner ces fonctions, l'usage des *traces* a été fait, rendant parfois ce corpus inexploitable pour une étude linguistique qui ne tient pas compte des effacements ou mouvements des théories générativistes.

1.2 Les corpus disponibles en français

Le français accuse un retard dans la pratique de la linguistique de corpus qui semble se combler ces dernières années. Plusieurs raisons peuvent être évoquées à ce sujet. L'informatisation des corpus bruts français n'était pas en retard puisque très tôt la base électronique Frantext a été constituée à partir des travaux de l'INaLF. En revanche, la diffusion de corpus dans le domaine public et sur Internet a été très tardive. Encore aujourd'hui, les documents mis à disposition de tous sont peu nombreux, pauvres en annotation et assez anciens pour des raisons de droits d'auteur.

La linguistique de corpus intéresse par ailleurs la communauté francophone depuis peu. Les premiers corpus de français parlé ont par exemple été constitués hors de France (*Corpus d'Orléans* par exemple). Le GARS (Groupe Aixois de Recherche en Syntaxe) a constitué un précédent en constituant dès 1975 un corpus de français parlé et un programme de recherche en linguistique de corpus.

Le français ne bénéficie d'un corpus comparable au Brown Corpus que depuis un an (corpus Parole, corpus CLIF), soit vingt ans après l'expérience sur l'anglais. En revanche, la compilation de textes écrits français se fait depuis de nombreuses années dans le but de constituer le Trésor de la Langue Française (TLF) : un dictionnaire de français exemplifié de nombreuses citations. La partie informatisée du TLF, FRANTEXT, est un recueil de textes de littérature française des XIX^{e} et XX^{e} siècles.

Depuis longtemps les linguistes disposent donc de textes français non annotés. Aujourd'hui ces textes sont étendus aux publications sur cdroms ou Internet par des éditeurs de quotidiens (Le Monde, Le Monde Diplomatique, Libération, etc.), de la Bibliothèque Nationale de France, du Trésor de la Langue Française. A tel point que la consultation de centaines de millions de mots devient abordable pour un laboratoire de recherche linguistique.

Cependant, ces textes se limitent souvent à l'écrit, et aux publications anciennes qui échappent aux droits d'auteur.

L'équipe du GARS de l'université de Provence a accumulé un corpus important (deux millions de mots) de texte oral retranscrit.

- Le projet Multext ([Véronis & Khouri, 1995]) fournit un corpus de 300 000 mots du *Journal Officiel de la Communauté Européenne* annotés automatiquement pour les parties du discours. Ce corpus n'a pas été corrigé manuellement.
- Le projet Multitag propose un corpus d'un million de mots segmentés, annotés pour la morpho-syntaxe par cinq étiqueteurs automatiques. Un système de vote automatique complété par la correction manuelle permet de proposer une bonne qualité d'étiquetage. Ce corpus s'apparente au Corpus Annoté de Paris 7 mais diffère sur plusieurs points dont les choix linguistiques des catégories et la segmentation des mots composés (cf 3.4). Ce projet s'inscrit dans le cadre de *l'action GRACE*([Adda *et al.*, 1999].)
- Le projet Parole offre un gros corpus dont une partie est annotée. Il contient 14 millions de mots issus du journal *Le Monde* dont 250 000 segmentés, étiquetés pour la morpho-syntaxe et partiellement corrigés.

– Le projet CLIF contient trois millions de mots dont 300 000 sont communs avec la partie morpho-syntaxique du Corpus Annoté de Paris 7 dont il est question dans ce présent mémoire.

1.3 L'annotation de corpus

Le terme d'annotation de corpus recouvre différentes réalités. La terminologie proposée traditionnellement ([Sinclair, 1996]) donne une acception précise et limitée à l'annotation qui consiste à documenter les textes pour indiquer leur origine, leur date de parution et toute information externe au texte qui pourrait intéresser un linguiste (le sexe de l'auteur, son âge, etc.)

Nous utiliserons ce terme de façon ambiguë, soit pour désigner les corpus documentés, que nous préférons nommer ainsi, soit plus souvent pour désigner des documents linguistiques écrits ou oraux (voir retranscrits) auxquels toute sorte d'informations ont été apportées dont voici une liste non-exhaustive :

– Alignement de mots, de paragraphes ou d'autres unités définies entre textes traduits ([Véronis & Langlais, 2000])
– Annotation de diverses informations sur la linéarité du texte oral : chevauchements de parole, reprises, etc. ([Blanche-Benveniste, 1999])
– Segmentations en unités diverses (mots, phrases, clauses)
– Annotation morpho-syntaxique des lemmes, traits flexionnels, parties du discours et autres catégories syntaxiques
– Annotation des dépendances syntaxiques, segmentation des constituants, annotation des étiquettes des constituants (éventuellement vides).
– Annotation d'indices de coréférence et d'anaphores
– Annotation sémantique

Les textes électroniques[1] volumineux représentent une matière première en ingénierie des langues. Ils permettent de valider un processus automatique en le confrontant à une réalité de la production langagière. Un système d'analyse supposé robuste doit être capable de s'appliquer sur une grande quantité de textes sans défaillance. Ainsi l'on teste tous les correcteurs grammaticaux, les indexeurs, les résumeurs, les traducteurs automatiques et autres

1. Nous appelons *textes électroniques* comme *dictionnaires électroniques* des documents enregistrés sur des supports de mémoire de masse sans rien supposer des traitements informatiques qui leur seront appliqués.

logiciels sur les textes bruts depuis que les textes électroniques existent. Les textes bruts fournissent également la matière à construire automatiquement des ressources linguistiques pour le TAL. Les listes de formes, de mots composés, de terminaisons sont extraites depuis longtemps par les concepteurs de correcteurs orthographiques par des méthodes très simples de consultation de documents très volumineux. Nous verrons plus loin qu'il est également possible d'utiliser ces textes en quantité pour établir des probabilités d'apparition d'occurrences de formes dans leur contexte. Cependant un corpus de textes qui ne contient aucune information linguistique reste d'un usage limité.

L'annotation plus récente des textes électroniques donne une plus grande valeur en Traitement Automatique des Langues à la linguistique de corpus.

L'annotation morpho-syntaxique de gros corpus corrigés manuellement donne la possibilité d'appliquer les méthodes des statistiques pour valider les systèmes comme les annotateurs morpho-syntaxiques automatiques (*Taggers*). Il est en effet devenu possible de mesurer la qualité de tels outils, de comparer une répartition aléatoire avec un taux d'erreurs mesuré. La constitution des corpus arborés permet de telles mesures sur les analyseurs syntaxiques.

L'annotation des corpus permet également de les utiliser comme ressources en TAL. Ce qui n'était possible que sur les formes fléchies devient possible sur d'autres types d'informations. Il est ainsi possible de générer automatiquement des dictionnaires terminologiques à partir de textes annotés en genres et domaines. Nous ne pourrons dresser une liste des possiblités offertes par les corpus annotés en TAL. Disons qu'il est possible d'induire automatiquement des ressources comme des dictionnaires et grammaires, d'entraîner des systèmes stochastiques, d'évaluer des étiqueteurs, routeurs[2] et autres analyseurs syntaxiques.

Nous pouvons citer une application française célèbre dans le domaine de la recherche en TAL : l'action GRACE ([Adda *et al.*, 1999]) dont le but était l'évaluation de plusieurs étiqueteurs morpho-syntaxiques. Ce projet a été possible après une réflexion sur la mise en correspondance des étiquettes utilisées en morpho-syntaxe.

On le voit, l'usage des corpus en TAL doit passer par une reflexion méthodologique sur leur adéquation avec d'autres ressources et techniques

2. Nous appelons *routeur* un système automatique qui extrait des textes en fonction d'une requête. Ce système s'inscrit dans le domaine de la recherche automatique d'informations.

en ingénierie des langues.

Les travaux industriels exploitant les corpus annotés pour induire des connaissances ou pour valider des systèmes ne passent généralement pas par une telle réflexion car l'enjeu théorique est bien souvent négligé. Cependant, les industriels et universitaires ont depuis peu admis qu'une réflexion sur l'encodage des ressources était devenu nécessaire. Ceci a donné lieu au consortium Genelex dans le cadre du projet Eureka par exemple dans les années 1992-1994.

Une réflexion plus profonde a été amorcée depuis les années 1995 (Marie-Paule Pery-Woodley [Péry-Woodley, 1995], Anne Abeillé [Abeillé, 1996], Anne Condamines et al. [Condamines *et al.*, 1999] sur l'adéquation des corpus annotés avec le Traitement Automatique des Langues.

Cette réflexion porte sur plusieurs points :

A - La taille des corpus

Quelle quantité de données est suffisante pour qu'un corpus puisse être représentatif d'un phénomène particulier ? Il est peu de travaux en linguistique de corpus informatisés qui s'articulent autour d'une étude statistique suffisamment précise pour évaluer la quantité minimale et nécessaire de données. Or il est crucial de connaître cela pour effectuer une étude fiable sur corpus. Le risque est d'obtenir des résultats médiocres alors qu'un faible coût aurait pu les améliorer sensiblement, ou au contraire de déployer un coût excessif pour obtenir un résultat qui n'en demandait pas tant. Au pire, les résultats peuvent s'approcher d'une répartition aléatoire rendant l'étude sur corpus absolument inutile.

Livrons-nous à une expérience sur le Corpus de Paris 7 dans le but de savoir quelle quantité de mots sont suffisants pour obtenir une base statistique représentative de trigrammes. Nous présenterons les méthodes stochastiques d'étiquetage en 4.2.2. Il n'est question ici que d'exemplifier cette question.

Nous allons dresser la liste des 10 trigrammes de parties du discours les plus fréquents à partir d'extraits de N mots du corpus ; N suivant une progression géométrique d'ordre 8.

Nous obtenons le résultat que nous résumons dans le tableau de la figure 1.1. On y voit que la répartition des 10 trigrammes les plus fréquents est sensiblement la même dans un corpus de 8 192 mots que dans un corpus de plus de 65 536 mots. De plus, la répartition est telle que les 10 trigrammes les plus fréquents couvrent 26% de leur total et que les trigrammes qui ne font pas partie des 20 plus fréquents couvrent moins de 1% de ce total.

Nombre de mots	Effectif	Trigramme	Pourcentage
1024	66	P/D/N	6.4%
	33	N/P/D	3.2%
	31	D/N/PONCT	3.0%
	28	D/N/P	2.7%
	26	N/P/N	2.5%
	22	D/N/A	2.1%
	21	N/PONCT/N	2.0%
	16	PONCT/D/N	1.5%
	16	P/N/PONCT	1.5%
	16	N/PONCT/D	1.5%
8192	562	P/D/N	6.8%
	350	D/N/P	4.2%
	319	N/P/D	3.8%
	223	N/P/N	2.7%
	196	D/N/PONCT	2.3%
	169	PONCT/D/N	2.0%
	165	D/N/A	2.0%
	137	P/N/PONCT	1.6%
	131	N/PONCT/D	1.5%
	118	V/D/N	1.4%
65536	4319	P/D/N	6.5%
	2715	D/N/P	4.1%
	2338	N/P/D	3.5%
	1668	D/N/PONCT	2.5%
	1532	N/P/N	2.3%
	1344	D/N/A	2.0%
	1283	PONCT/D/N	1.9%
	1060	V/D/N	1.6%
	909	P/N/PONCT	1.3%
	882	N/PONCT/D	1.3%
524288	32803	P/D/N	6.2%
	20778	D/N/P	3.9%
	17856	N/P/D	3.4%
	13386	D/N/PONCT	2.5%
	11326	N/P/N	2.1%
	10041	PONCT/D/N	1.9%
	9805	D/N/A	1.8%
	8324	V/D/N	1.5%
	7077	P/N/PONCT	1.3%
	6694	N/PONCT/D	1.2%

FIG. 1.1 – *Étude des trigrammes de POS en fonction de la taille du corpus*

Le résultat de cette courte étude montre qu'il est inutile de doubler voire de tripler un corpus de plus de 10 000 mots pour espérer améliorer sensiblement un étiqueteur probabiliste se basant sur les trigrammes de parties du discours. Il faudrait bien sûr approfondir une telle étude pour savoir quelle quantité est requise pour une analyse statistique, mais nous voyons ce que peut coûter l'absence d'une telle étude.

B - La représentativité des corpus.

Un corpus doit être représentatif de l'étude qu'il est censé couvrir. Marie-Paule Péry-Woodley ([Péry-Woodley, 1995] p. 218) remarque qu'un corpus représentatif d'une langue n'est évidemment par exhaustif ni équilibré, ce qui supposerait de façon aussi irréaliste qu'il couvre la «langue générale» ; notion insaisissable. Nous présentons ci-dessous ce que nous entendons par corpus de référence avec plus de précision.

C - Réutilisabilité

Pour qu'un corpus soit utilisable en TAL comme dans les autres domaines de la linguistique, il faut qu'il puisse livrer assez de données dans le domaine spécifique de l'étude. C'est-à-dire que son organisation doit permettre d'en extraire une partie en fonction d'un besoin spécifique. Le Corpus de Paris 7 permet par exemple de fournir des suites de parties du discours, des suites de lemmes, du texte segmenté en phrases, du texte segmenté en mots composés ou en mots simples pour ne prendre que ces exemples. Ce corpus est donc réutilisable par la communauté linguistique en fonction de l'étude qui est menée.

1.4 Construction de corpus

Corpus de référence

Par *corpus de référence* nous pouvons entendre plusieurs choses. Le plus souvent on dit qu'un corpus est un corpus de référence quand il couvre de façon représentative tous les domaines et niveaux de langue. C'est-à-dire qu'il contient des textes écrits mais également oraux, des textes issus d'une langue de registres divers mais également d'argot et autres sophistications de la langue employée.

Nous pouvons entendre également que le corpus est représentatif d'un état de langue. C'est-à-dire qu'il contient à peu près tous les types de constructions, de domaines, de registres, un lexique couvrant, etc. Cette notion de corpus a été largement critiquée par les générativistes à la suite de Chomsky. Ces corpus censés induire une grammaire sont nécessairement finis, alors que la grammaire doit rendre compte de la compétence à produire une infinité de phrases et contiendraient des «accidents de performance». Cette critique du *positivisme* linguistique qui consisterait à proposer qu'un corpus puisse être l'objet même de la linguistique comme relevant de toutes les productions lan-

gagières a été étendue à une critique sur l'idée générale que le corpus puisse être l'objet du linguiste. Dans cette dernière optique structuraliste, le corpus est une donnée à partir de laquelle le linguiste vise à retrouver la compétence du locuteur.

La critique des génératives est fondée. Dans le premier cas, il n'y a aucune raison pour que l'étude d'un corpus clos puisse rendre compte de la capacité infinie de parole. Dans le deuxième cas, le principe d'immanence est supposé pour que les mécanismes de la compétence du sujet parlant puissent se retrouver dans sa production. Or ce principe n'a aucun fondement théorique.

On le voit, il faut prudemment considérer un corpus comme un corpus de référence. Nous ne considérons pas qu'un corpus de référence, au sens où nous l'entendons, doit induire une grammaire ni être l'objet de la linguistique. Nous entendons qu'il puisse fournir, de façon représentative, des indications sur des faits de langue.

Les méthodes de travail sur corpus depuis quelques années ont permis de produire des résultats qu'il n'était pas envisageable d'obtenir à la main. Nous proposons au chapitre 5 trois études qui ont été réalisées grâce aux outils informatiques.

Pour que ces résultats aient une valeur théorique, il faut que le corpus soit représentatif de l'étude. C'est-à-dire, qu'au delà de la seule collection de textes, le corpus doit fournir des indications linguistiques validées en relation avec l'étude menée. C'est à ce prix qu'un corpus peut être vu comme un corpus de référence.

Le Corpus de Paris 7 est entièrement corrigé à la main pour la morpho-syntaxe et les annotations morpho-syntaxiques ont été motivées et documentées dans des guides d'annotation [Abeillé & Clément, 1997] que nous avons détaillés au chapitre 3. Il constitue donc un corpus de référence pour une étude exploitant ces informations.

Chapitre 2

Le projet de Corpus Annoté de Paris 7

Nous présentons le Corpus Annoté de Paris 7 dans ce chapitre. Avant d'expliquer comment le corpus a été validé par des correcteurs, nous expliquerons les choix qui ont été faits sur les textes. Le corpus doit être représentatif de l'usage qui lui est destiné, ces choix doivent donc être explicites et motivés.

L'annotation morpho-syntaxique de l'ensemble du corpus a nécessité des procédures manuelles et automatiques en nombre. Nous détaillerons la méthode complète en expliquant les différentes étapes qui ont permis la segmentation, la lemmatisation, l'étiquetage morpho-syntaxique et l'annotation des mots simples et des mots composés. En outre, nous expliquerons les trois jeux d'étiquettes morpho-syntaxiques qui ont été définis pour trois procédures particulières qui ne se recouvrent qu'en partie.

Enfin après avoir défini dans les grandes lignes les objectifs de la construction d'un tel corpus et des outils qui l'accompagnent, nous présenterons le projet en pratique dans le but d'en faire un antécédent exploitable en ingénierie linguistique.

2.1 Représentativité du corpus

Les textes que nous avons choisi d'étiqueter sont extraits d'un ensemble d'articles du journal Le Monde de 1989 à 1993 compilés sous l'égide de l'ELSNET et distribués par l'ELSNET et Linguistic Data Consortium (LDC). Nous avons sélectionné 1 million de mots (soit environ 875 000 mots gra-

phiques sans ponctuation, 920 000 mots graphiques composés) en puisant de façon aléatoire dans l'ensemble de la base pour faire varier les dates de parution des article, leurs auteurs et leurs thèmes.

Ces textes journalistiques constituent une source de français contemporain écrit. Nous n'entendons cependant pas dire que le français utilisé par les journalistes de ce quotidien est représentatif des différentes productions langagières écrites du monde francophone. Ce corpus ne convient pas à l'étude des genres et registres car bien que l'on ait conservé des informations externes (nom de l'auteur, date de parution, mots clés choisis par l'éditeur, titre, etc.) aucun critère interne n'a été exploité pour découper le texte en échantillons représentatifs [Biber, 1991]. Par ailleurs, l'intégralité du texte provient de la même source de diffusion et ne peut embrasser des registres de langues différents.

En particulier les différentes réalisations régionales y sont absentes. Par *standard*, nous entendons que les textes ne sont pas particulièrement marqués d'un niveau de langue, que les figures propres à la littérature, au discours oral ou au dialogue par exemple sont absentes. Le style journalistique peut sembler marqué par quelques aspects : l'emploi des terminologies sportives ou économiques y sont abondantes, des figures moins fréquentes dans d'autres domaines de la langue se voient particulièrement usitées. De manière générale toute figure qui permet de passiver ou d'élider un agent est abondamment utilisée par les journalistes qui ne désirent pas toujours être explicites sur les acteurs des faits rapportés ou parce qu'ils veulent simplement se désengager du propos rapporté (*Une population se voit reconduite à la frontière*, *Le président du club a été limogé*).

Bien que ces textes ne soient pas marqués du point de vue du discours, la représentativité de ce corpus n'est guère assurée par sa couverture en domaines ou en niveaux de langue comme c'était idéalement le cas pour l'anglais britannique du Brown Corpus (voir 1.1.1) et de ses successeurs.

En fait, il est difficile de choisir un texte représentatif d'une langue sans évaluation statistique sur le lexique ou quelques formes choisies pour échantillonnage. Nous nous sommes contentés d'exclure un certain nombre de textes pour ne garder que la littérature qui semblait convenir à l'usage dont nous pensions qu'il serait fait du corpus.

Nous avons exclu :

- Les textes purement littéraires, les romans et nouvelles, pamphlets, poésies, etc. Ces textes sont très marqués par quelques figures stylistiques n'apparaissant que rarement en français «standard». Nous pen-

sons que de tels textes ne peuvent apparaître dans un corpus de français écrit dont l'exploitation n'intéresse pas directement l'étude littéraire. Il reste que le texte du *Monde* est styliquement marqué et le choix de ce quotidien en particulier est discutable de ce seul point de vue.

- les textes qui nous semblaient contenir une terminologie (textes techniques, médicaux et scientifiques dont les noms composés sont en grand nombre des termes du domaine).
- Les textes de langue contrôlée. Ces textes permettent de fournir à des automatismes d'analyse un matériau moins complexe. La méthode employée généralement est l'élimination de figures difficiles à modéliser comme les dislocations, le passif, les dépendances hors îlot, etc. Ces textes sont peu représentatifs de la richesse de la langue.
- La retranscription de textes parlés.
  - le texte oral ne se projette pas à l'écrit simplement de façon linéaire. Ces textes contiennent des reprises, des chevauchements de parole, des ambiguïtés de retranscription [Blanche-Benveniste, 1996].
  - les textes oraux et les textes écrits ont des points de divergence, dont fait mention [Blanche-Benveniste, 1997]. On remarque particulièrement des phénomènes lexicaux se manifestant très différemment à l'oral et à l'écrit. Il n'était donc pas raisonnable d'inscrire ce type de textes dans un corpus se voulant homogène et monolithique.
- Le courrier électronique.
  Les forums de discussion, seuls courriers électroniques accessibles en nombre et non privés sont des textes dont le style est propre au mode de diffusion. Ces textes rapidement écrits contiennent souvent des abréviations, une typographie pauvre et un style *proche de l'oral* (phrases courtes avec peu d'enchâssements, temps verbaux de l'oral).
- Les pages Internet.
  Les textes très volumineux que l'on peut trouver sur l'Internet, sont une source assez exploitée comme corpus électronique. Le seul avantage de ces textes est d'être en nombre et libre de droits d'auteur. Le français de ces données électroniques est souvent le résultat d'un traducteur automatique dont nous savons les limites actuelles[1]. Nous avons fait une recherche très massive sur l'Internet grâce à un «robot» pour extraire des textes français (filtrés par de simples mots clefs propres au français).

1. A moins de restreindre les recherches au domaine de moins en moins utilisé *fr* ; ce qui garantit partiellement que les textes aient été produits en France

> Le résultat de cette recherche était inexploitable : un grand nombre de ces textes était incompréhensible par un locuteur francophone et les concordances que l'on risquait d'extraire étaient celles des quelques systèmes de traductions disponibles sur l'Internet (Reverso, Systran, etc.).

Si quelques phrases sont «bizarres à entendre», ou «mal écrites» dans le Monde, comme le disaient parfois les annotateurs, la très grande majorité du texte couvert par le corpus suit la compétence d'un locuteur francophone natif. En cela, il est représentatif des constructions du français écrit «standard» même si une partie que nous aurions certainement dû écarter appartient manifestement au domaine politique et économique.

Les textes couvrent différents domaines puisque les pages culturelles, les pages économiques ou sportives du quotidien n'ont pas été exclues et que les articles éditoriaux traitent de sujets très variés.

Ce corpus est relativement important en taille ; les informations linguistiques qu'il peut contenir peuvent faire l'objet d'études statistiques. Un million de mots est la limite que nous avons fixé pour que la correction manuelle soit possible dans le temps imparti.

2.2 Méthodologie

Comme les projets *Penn TreeBank* [Marcus *et al.*, 1993] et *Prague Dependency Treebank* [Hajicova *et al.*, 1998]), le projet est organisé en deux étapes : un étiquetage automatique suivi d'une validation manuelle et systématique de la totalité des textes.

Le corpus a d'abord été segmenté et annoté automatiquement avec un étiqueteur morpho-syntaxique[2] couplé à un dictionnaire de mots simples et de mots composés.

Le corpus a été corrigé dans sa totalité par des annotateurs qui devaient valider la segmentation des mots et leur étiquetage morpho-syntaxique en suivant les recommandations que nous présenterons au chapitre 3.

La figure 2.1 représente la méthode générale de l'étiquetage morpho-syntaxique. Nous y distinguons principalement deux étapes. La première étape est entièrement automatique. Le corpus est segmenté en phrases et

2. dit également *tagger*

en mots puis chacun de ces mots est annoté avec une étiquette morpho-syntaxique simplifiée. La deuxième étape est semi-automatique, elle consiste à anticiper automatiquement certaines informations ambiguës qui n'étaient pas prédites par l'étiqueteur automatique puis à valider manuellement la segmentation et l'étiquetage. Enfin les informations non-ambiguës de chaque mot étaient projetées du lexique vers le corpus comme nous allons le voir.

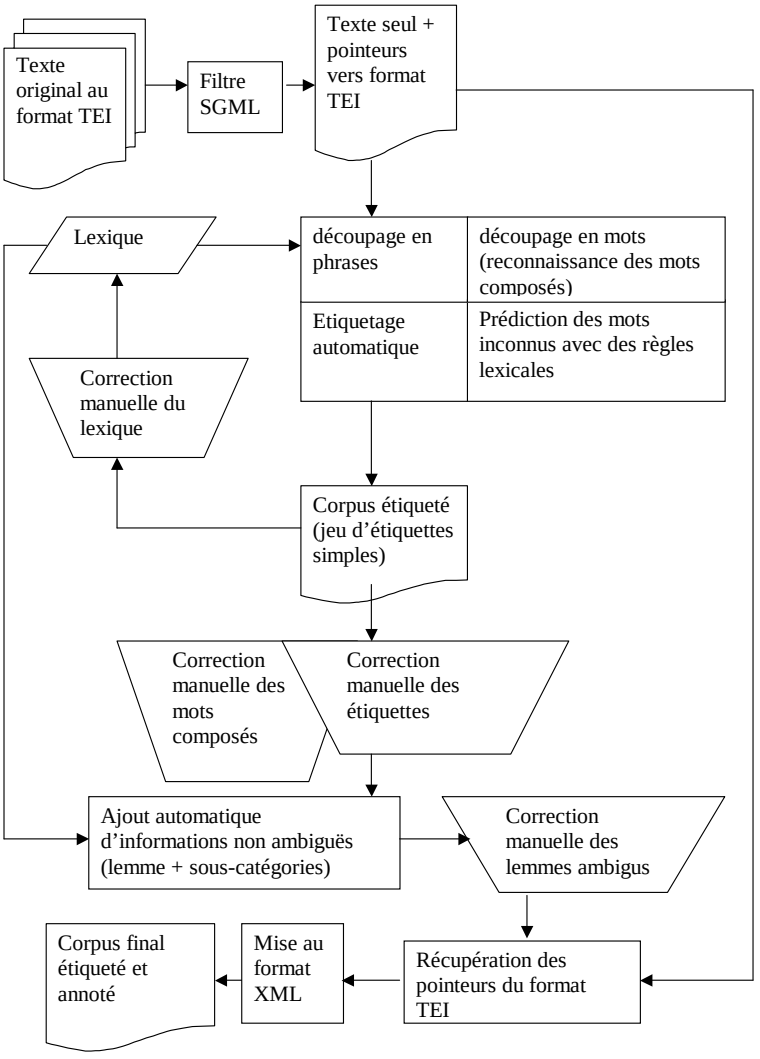


FIG. 2.1 – *Annotation morpho-syntaxique du corpus*

Trois jeux d'étiquettes

Nous avons défini trois jeux d'étiquettes adaptés aux différentes tâches: étiquetage automatique, validation manuelle et édition finale. Le corpus de référence est livré avec un jeu d'étiquettes complet encodant la partie du discours, la sous-catégorie grammaticale, la morphologie flexionnelle, les lemmes et enfin les composants des mots composés. L'ensemble de ces informations n'est pas nécessaire lors de la validation de l'étiquetage ; le lemme et la plupart des sous-catégories sont calculés à partir d'un lexique. De plus, l'étiqueteur automatique ne fait pas certaines distinctions que nous avons retenues.

Version simplifiée Cette version est simplifiée pour l'étiqueteur automatique. Elle contient 110 étiquettes et ne distingue pas, par exemple pronom relatif et interrogatif comme en (a), ou les cas des pronoms clitiques comme en (b) et (c). Ce jeu d'étiquettes contient la catégorie et la morphologie flexionnelle. Il contient en plus la sous-catégorie des noms et des conjonctions.

(a) Jean demande à la fille **qui** viendra si elle n'a besoin de rien.
(b) Jean **nous** a obtenu une place de cinéma.
(c) Jean **(lui + *le)** a obtenu une place de cinéma.

La désambiguïsation de la phrase (a) réclame une exploration du corpus qui dépasse la capacité d'un étiqueteur morpho-syntaxique tel que présenté en 4.2. En effet, sans connaissance des propriétés de sous-catégorisation du verbe *demander*, sans représentation de constituance et dépendance permettant d'identifier *si elle n'a besoin de rien* comme complément oblique du verbe, il ne sera pas possible de préférer l'une des lectures de la phrase. Et même avec ces éléments, il faudra encore identifier la coréférence du pronom *elle* et exploiter des connaissances pragmatiques pour exclure le complément circonstanciel *si elle n'a besoin de rien* et ainsi identifier un pronom relatif en *qui*. En (b), pour identifier *nous* comme un pronom datif et non comme un pronom accusatif, il est nécessaire de connaître la sous-catégorisation du verbe *obtenir*. Ceci est exclu des capacités de notre étiqueteur, cette distinction est simplement éliminée de l'étiquetage.

Version de granularité moyenne (122 étiquettes) Cette version du jeu d'étiquettes est destinée à la validation manuelle. Elle permet d'annoter, en plus des informations morpho-syntaxiques de la version simplifiée,

Catégorie ou Partie du discours	Description	Personne	Genre	Nombre	Temps et mode verbal
NC	nom commun	–	+	+	
NP	nom propre	–	+	+	
A	adjectif	–	+	+	
ADV	adverbe	–	–	–	
P	préposition	–	–	–	
D	déterminant	+	+	+	
CL	pronom clitique	+	+	+	
PRO	autres pronoms	+	+	+	
CC	conjonction de coordination	–	–	–	
CS	conjonction de subordination	–	–	–	
I	interjection	–	–	–	
V	verbe	+	+	+	W, G, K, P, I, J, F, T, C, S, Y
ET	mot étranger	–	–	–	
PONCT	ponctuation	–	–	–	

FIG. 2.2 – *Jeu d'étiquettes du tagger*

celles qui doivent être désambiguïsées manuellement faute d'un traitement automatique le permettant.

En revanche, les informations qui ne sont pas ambiguës hors contexte sont annotées automatiquement et n'ont pas besoin d'être avalisées par l'annotateur. La version de granularité moyenne est donc un jeu d'étiquettes

simplifiées destiné à l'annotation du corpus.

La sous-catégorie **cardinal** accompagnant tous les nombres n'est pas indiquée généralement car cette information n'est pas ambiguë et peut être ajoutée automatiquement en consultant le lexique. En revanche, l'adjectif *neuf* sera noté **cardinal** ou **qualificatif** selon son contexte. La projection automatique des informations ambiguës exploite des résultats statistiques pour proposer l'étiquette la plus probable hors contexte. Pour reprendre notre exemple, *neuf* sera brutalement étiqueté **adjectif cardinal** dans l'ensemble du corpus, car sa fréquence comme qualificatif est moindre. Nous n'avons pas choisi de faire un traitement sophistiqué pour mieux étiqueter ces termes ambigus. Ce traitement aurait réclamé des heuristiques pour des termes qui sont en nombre relativement faible. Ils doivent donc être corrigés à la main.

Aux sous-catégories des noms et des conjonctions a été ajoutée la sous-catégorie des pronoms relatifs.

La forme fléchie associée à l'étiquette morpho-syntaxique permet de trouver dans le lexique le lemme de chaque terme. Nous considérons donc que cette information n'est pas ambiguë et ne doit pas apparaître dans la version de granularité moyenne.

Il reste cependant quelques cas d'homonymie entre des termes qui partagent la même partie du discours et les mêmes traits flexionnels (*fils, cuisinière*, etc.), voir *infra* quelques autres exemples. Nous en profitons pour dire que cette homonymie porte uniquement sur le lemme et non sur le lexème, le vocable ou les variations d'écriture. Le lemme est une abstraction sur les mots. Il s'agit d'une classe de formes graphiques qui ne se différencient que par leurs variations flexionnelles. C'est-à-dire que deux formes partageant le même lemme sont équivalents à leurs traits flexionnels près.

Pour désigner le lemme, l'usage est de choisir une forme flexionnelle non marquée[3] : l'infinitif d'un verbe, le masculin singulier des noms, pronoms, adjectifs, déterminants et participes passés. Le lemme et les informations flexionnelles permettent de déduire la forme graphique. Ainsi le nom commun *fraises* n'est pas ambigu sur le lemme *fraise* que ce nom désigne en contexte l'outil du dentiste ou le fruit, puisque toutes les variations flexionnelles des deux acceptions sont les mêmes pour les deux lexèmes. *Rengrènent* ne correspond qu'à un seul lemme *rengrener* bien qu'il y ait une variation libre de l'écriture de l'infinitif *rengréner ou rengrener* (dans ce cas, l'infinitif

3. La discussion portant sur la marque flexionnelle d'une catégorie (genre, nombre, temps verbal, etc.) n'est pas nécessaire au choix arbitraire de l'une des formes d'un même lemme comme représentant de ce lemme.

rengréner correspondra au lemme *rengrener* si le choix du lemme a été porté sur cette écriture arbitrairement.

Pour quelques 176 formes, l'étiquette morpho-syntaxique ne suffit pas à trouver le lemme (contrairement aux formes *savons*, *portes*, etc. que la partie du discours permet de désambiguïser). Voici quelques termes qu'on trouve en nombre dans le corpus :

(nous indiquons la forme fléchie ambiguë en italique)

- croire/croître *crût*
- (dé)peigner/(dé)peindre *(dé)peignait*
- falloir/faillir *faille*
- foi/fois *fois*
- mouler/moudre *moule*
- ouvrir/ouvrer *ouvrait*
- (re)couvrir/(re)couvrer *(re)couvre*
- (re)fondre/(re)fonder *refondait*
- plaire/pleuvoir *plu*
- être/suivre *suis*
- surfer/surfaire *surfera*

La différence de fréquence entre les lemmes de ces formes ambiguës est importante en contexte dans pratiquement tous les cas. Les lemmes les plus fréquents en contexte sont proposés par le système et sont systématiquement corrigées par les annotateurs.

Le jeu de granularité moyenne encode :

- La catégorie
- Les sous-catégories ambiguës, les sous-catégories des noms, conjonctions et des pronoms relatifs
- La morphologie
- Les seuls lemmes ambigus

Nous présentons en 2.3 les codes des étiquettes utilisées pour la validation manuelle de la morpho-syntaxe du corpus.

Version complète (202 étiquettes) La version complète est automatiquement enrichie grâce à une projection du lexique. Toutes les sous-catégories, les lemmes et les composants des mots-composés sont automatique-

[123] désigne la 1re, 2e ou 3e personne [mf] désigne le genre masculin ou féminin [sp] désigne le nombre singulier ou pluriel		
Code	**Description**	**Sous-catégories**
NC[mf][sp]	Nom commun	
NP[mf][sp]	Nom propre	
A[mf][sp]	Adjectif	Cardinal, indéfini
ADV	Adverbe	Interrogatif, exclamatif
P	Préposition	
D[mf][sp]	Déterminant	Partitif
CL[123][mf][sp]	Pronom clitique	Sujet, Objet, Réflexif
PROR[123][mf][sp]	Pronom relatif	
PRO[123][mf][sp]	Autre pronom	Interrogatif
CC	Conjonction de coordination	
CS	Conjonction de subordination	
I	Interjection	
VW	Verbe infinitif	
VG	Verbe participe présent	
VK[mf][sp]	Verbe participe passé	
VP[123][sp]	Verbe présent de l'indicatif	
VI[123][sp]	Verbe imparfait de l'indicatif	
VJ[123][sp]	Verbe passé simple	
VF[123][sp]	Verbe futur	
VT[123][sp]	Verbe subjonctif passé	
VC[123][sp]	Verbe conditionnel	
VS[123][sp]	Verbe subjonctif présent	
VY[12][sp]	Verbe impératif	
ET	Mot étranger	
PONCT	Ponctuation	

FIG. 2.3 – *Jeu d'étiquettes pour la validation du corpus annoté*

ment notés à partir des autres informations morpho-syntaxiques et de la forme fléchie.

Nous présentons en 2.4 le jeu complet d'étiquettes du corpus. Ces informations morpho-syntaxiques, ainsi que les composants des mots composés sont encodées grâce à un balisage XML spécifique (voir 4.1.4). Nous n'avons donc pas défini de code pour chaque étiquette morpho-syntaxique complète

comme c'était le cas dans le projet Multext [Véronis & Khouri, 1995]. Le tableau montre les sous-catégories et traits morphologiques de chaque catégorie sans tenir compte de quelques exceptions (les verbes *voici*, *voilà* n'ont pas de variation morphologique, l'adverbe *tout* peut varier en nombre et en genre). Nous renvoyons vers les sections spécifiques à chaque description des parties du discours pour une discussion sur ces exceptions.

Catégorie ou Partie du discours	Description	Sous-Catégorie	Personne	Genre	Nombre	Temps et mode verbal
N	Nom	commun, cardinal propre	−	+	+	
A	adjectif	cardinal, ordinal, poss., qualif., indéf., interr.	−	+	+	
ADV	adverbe	interr., exclam., négatif	−	−	−	
P	préposition		−	−	−	
D	déterminant	card., dém., déf., indéf., exclam., négatif, poss.	+	+	+	
CL	pronom clitique	suj., obj., refl.	+	+	+	
PRO	autres pronoms	interr., pers., négatif, poss. rel., indéf., démonstr., cardinal	+	+	+	
C	coord., subord.	−	−	−	−	
I	interjection	−	−	−	−	
V	verbe	−	+	+	+	W, G, K, P, I, J, F, T, C, S, Y
ET	mot étranger	−	−	−	−	
PONCT	ponctuation	forte, légère	−	−	−	

FIG. 2.4 – *Jeu d'étiquettes complètes du corpus de référence*

2.3 Étiquetage automatique

L'annotation automatique du corpus et la segmentation en mots et en phrases ont été réalisées grâce à un étiqueteur automatique. Nous présenterons cet outil comme ressource informatique en 4.2.3.

Le résultat de cet étiquetage à été automatiquement enrichi vers le jeu d'étiquettes de granularité moyenne grâce à des outils développés pour l'occasion et grâce au logiciel Lexed que nous présenterons en 4.3. Ensuite, le corpus a été formaté pour assurer une bonne lisibilité et sa réutilisabilité.

L'ensemble de ces opérations a réclamé l'écriture d'un grand nombre de petits programmes, le plus souvent écrits avec un langage de commandes UNIX du type *shell*. Pour assurer une reproductibilité des opérations, nous avons typé les fichiers en fonction de leur contenu en ajoutant un suffixe à leur nom comme c'est l'usage sous Unix et avons développé des dépendances grâce à l'utilitaire **Make**.

Le corpus destiné aux annotateurs dépend du corpus étiqueté automatiquement qui dépend lui-même du lexique, mais également des listes de lemmes ambigus, etc. C'est au prix de la description du réseau complet des dépendances entre fichiers que nous pouvions assurer que les fichiers soient mis à jour en fonction des dernières modifications et corrections des lexiques, listes d'exceptions, logiciels et textes sources.

De nombreuses corrections du corpus ont été apportées dans les lexiques pour être projeté sur l'ensemble des occurrences du corpus automatiquement. Ces corrections portaient sur la segmentation des mots composés et sur l'étiquetage des mots grammaticaux.

La nature des composants des mots composés et les lemmes ont été validés et corrigés directement sur les lexiques produits à partir du corpus avant d'être projeté automatiquement.

2.4 Validation du corpus étiqueté

Comme la correction du *British Component of the International Corpus of English (ICE-GB)*[Wallis, 2000], deux méthodes ont été employées pour valider le Corpus Annoté de Paris 7. L'une transversale et semi-automatique garantissait la cohérence de l'étiquetage, l'autre longitudinale et entièrement manuelle garantissait qu'une décision a été prise pour chaque cas problématique sur l'intégralité du corpus.

La correction transversale s'intéresse à des types de mots ou d'ambiguïtés rendant problématique une décision isolée. Nous présentons une étude de ces problèmes en 3.9. Les lexiques ont également fait l'objet d'études, notamment sur les compositions et les locutions que nous voulions systématiser sur tout le corpus.

La correction longitudinale fait intervenir les compétences des linguistes pour prendre une décision isolée visant à rapporter les critères d'une classe à un terme, ou pour décider qu'un terme est ou non un composé *in fine*.

Un guide d'annotation très précis a été écrit et complété durant la correction manuelle du corpus ([Abeillé & Clément, 1997]). L'intégralité du corpus a été relue deux fois par deux personnes différentes et de petits programmes (ou macros d'éditeurs de texte) permettant de repérer des suites d'étiquettes rares ont été utilisés pour corriger semi-automatiquement des erreurs résiduelles.

2.4.1 Validation des mots composés

Un grand nombre de mots composés sont potentiellement ambigus avec une suite homographe de mots simples. C'est par exemple le cas de la locution *en fait* en (a) et (b).

(a) Jean *en fait* toute une histoire
(b) Jean n'est jamais venu *en fait*

L'étiqueteur de Paris 7 donnait systématiquement une préférence pour les mots composés. L'annotateur devait alors valider le mot composé ou l'éclater en mots simples selon le contexte.

2.4.2 Validation de la morphologie

Le corpus ne contient aucune ambiguïté morphologique. Certains noms propres ne possèdent intrinsèquement pas de genre (comme le nom de certaines villes ou certaines marques), si le contexte ne contient aucune marque d'accord, l'information morphologique est sous-spécifiée avec le seul nombre. Dans les autres cas, l'étiquette morphologique est notée avec la plus grande finesse. Ainsi, même lorsque le terme ne porte pas de marque de flexion, il porte les traits d'accords comme illustré en (a), (b) et (c).

(a) La roue avant (*Adjectif féminin singulier*).

(b) Les roues avant (*Adjectif féminin pluriel*).

(c) Moi qui (*Pronom relatif féminin singulier*) lui (*Pronom fort masculin singulier*) ai donné un fils.

2.4.3 Validation des catégories

Le corpus ne contient aucune ambiguïté résiduelle portant sur les catégories. Ainsi nous ne trouvons dans le corpus aucune phrase comme les exemples fabriqués : *Le boucher sale la tranche* ou *La bonne voile la porte* ou encore *la belle ferme le voile.*

Il existe bien sûr des phrases ambiguës dans le corpus (comme les ambiguïtés de rattachement prépositionnel ou de portée de quantificateur) mais aucune ambiguïté n'a été relevée par les annotateurs comme relevant d'une impossibilité de choisir entre telle ou telle étiquette d'un mot comme c'est le cas dans les phrases fabriquées que nous avons cité. Lorsque l'annotateur se trouvait en difficulté pour apposer la bonne étiquette, ce n'était pas tant que le mot était ambigu mais qu'il possédait les propriétés de plusieurs classes.

Les mots comme *sale*, *ferme*, *bonne* qui sont polycatégoriels en langue ne posaient pas de problème particulier car jamais ambigus en discours[4]

En revanche les extraits qui suivent posent des problèmes de choix. Non parce que les termes que nous avons surlignés en gras soient ambigus dans leur contexte, mais parce que le contexte (dans le sens le plus large possible) propre à chaque catégorie retenue ne se laisse pas aisément capter.

(a) À l'exemple de Jean-Noël Gaviot, pilote d'un **Toyota** (*nom commun ou nom propre ?*), tout surpris d'être doublé à quelques kilomètres de l'arrivée.

(b) ...pour récupérer les paillettes de sperme **congelé** (*adjectif ou verbe ?*) de son mari décédé.

En (a), le nom *Toyota* possède les propriétés que nous avons retenues pour les noms communs. Le nom ne dénote pas le fabricant lui-même mais une voiture en particulier. Ce qui se laisse supposer par la présence du déterminant indéfini.

4. À moins que l'auteur joue avec la langue précisément pour obtenir un effet en forçant l'ambiguïté de deux homographes. Nous n'avons heureusement pas rencontré de tels jeux de mots dans le Corpus de Paris 7.

En forçant le trait il est vrai, nous pourrions dire que l'article indéfini est celui qui est employé pour désigner une propriété. Lequel détermine parfois les noms propres comme dans «Un Mitterrand plus déterminé que jamais à convaincre son auditoire.»

En (b), le mot *congelé* est un verbe dans la mesure où la structure argumentale du verbe peut être reconstruite. Un complément d'agent du verbe *congelé* marquerait sûrement les propriétés verbales, notamment l'aspect , le temps ou la structure actancielle ([...] pour récupérer les paillettes de sperme **congelé** par le laboratoire Bioglac [...].) En absence de cet argument verbal, il est difficile de supposer son effacement et l'on peut analyser la phrase tout autrement en considérant non pas l'événement ou le procès mais une propriété inaliénable du nom (le sperme congelé est une entité telle dans la littérature journalistique médicale ou juridique qu'on peut imaginer la lexicalisation prochaine du terme). Dans ce cas, **congelé** sera un adjectif.

2.5 Le projet en pratique

Le projet a mobilisé le travail de plus d'une vingtaine de personnes durant les mois d'été des quatre années passées. Les stagiaires qui ont participé à ce travail à différentes périodes on eu des fonctions différentes dont la principale était de corriger les mots composés et les étiquettes morpho-syntaxiques, mais aussi de corriger les lexiques, d'écrire les règles de l'étiqueteur morpho-syntaxique du Corpus Annoté de Paris 7, de participer au développement de l'étiqueteur et de nombreux autres outils qui ont été mis en œuvre lors de ce projet.

Outil de correction des étiquettes morpho-syntaxiques Nous avons développé un environnement de travail à l'usage des annotateurs sous l'éditeur de texte *emacs* qui permettait de corriger les étiquettes rapidement et de façon ergonomique.

Contrairement à l'équipe du DFKI de Sarrebruck [Brants *et al.*, 2001] ou à ce qui a été fait pour le Penn treebank [Marcus *et al.*, 1993], nous n'avons pas présenté d'interface graphique sophistiquée permettant de choisir l'élément à remplacer dans un menu déroulant ou autres *widgets*. Ce choix a été fait en considérant qu'un utilisateur entraîné manipule des objets informatiques bien plus lentement et laborieusement avec la souris qu'il ne fait avec un clavier et des procédés mnémoniques.

Nous n'avons pas non plus, pour les mêmes raisons, choisi de proposer un dialogue d'interface en clair, mais une codification des étiquettes à corriger. L'ensemble des informations morpho-syntaxiques a donc été encodé avec un jeu d'étiquettes mnémoniques.

Néanmoins, pour rester cohérent avec les travaux du LADL, nous avons choisi en partie, la codification des dictionnaires DELA [Silberztein, 1993].

De plus, comme nous l'avons expliqué en 2.2, l'annotateur ne devait pas corriger l'ensemble des informations morpho-syntaxiques, mais seulement celles qui étaient ambiguës.

L'annotateur ne devait donc retenir que la liste des codes présentée en 2.3, ce qu'il parvenait à faire sans difficulté en quelques heures et faisait ainsi abstraction de ces codes pour ne plus voir que les incohérences d'annotation.

Lors d'une session de travail, le clavier de la machine était redéfini pour permettre de se déplacer dans le texte avec aisance, mais également pour corriger, soit la partie du discours soit la morphologie. Nous avons également défini des fonctions du clavier permettant d'annoter la segmentation des mots composés ou d'éclater une suite de termes rassemblés à tort comme un composé.

Ce travail était ponctué de réunions hebdomadaires permettant de mettre en évidence les principaux problèmes rencontrés lors de la correction de l'étiquetage. La correction manuelle du corpus a été coûteuse en temps de travail (45 hommes mois soit 500 mots à l'heure environ). Cela est dû essentiellement à la correction des mots composés, mais également à la finesse de la granularité de l'étiquetage.

Chapitre 3

Les choix linguistiques pour le corpus annoté

Dans cette section, nous présenterons les choix linguistiques retenus pour l'annotation morpho-syntaxique du Corpus Annoté de Paris 7. Les deux types d'apports qui constituent l'annotation du corpus sont la segmentation du texte en unités et l'assignation d'étiquettes à ces unités. Nous devrons donc définir ces deux objets linguistiques du point de vue qui nous intéresse : l'utilisation du corpus par des linguistes à des fins de recherche et l'inscription de ce travail dans le cadre du Traitement Automatique des Langues.

L'annotation d'un corpus doit être un compromis entre le souci de mettre à jour et d'exprimer des généralisations linguistiques – ce qui peut conduire à innover par rapport à la tradition grammaticale – et un souci de conservatisme terminologique et typologique pour rendre plus naturelle l'interrogation du corpus.

Le corpus doit pouvoir être utilisé par la communauté linguistique pour des études très diverses en présupposant le moins possible des éléments de la théorie dans laquelle s'inscrit cette étude. En conséquence, l'annotation du corpus ne pourra pas être l'application d'une théorie linguistique comme le programme minimaliste de Chomsky ou une théorie basée sur une grammaire d'unification ou encore les grammaires de dépendance de Tesnière. Or cela permettrait de définir l'ensemble de nos choix linguistiques avec le seul souci de ne pas être contradictoire avec la théorie retenue.

Pour cette raison, nous ne nous satisfaisons pas de définitions formelles se rapportant à un aspect unique d'une science du langage. Les «catégories» ne seront pas définies par leurs seules propriétés distributionnelles mais égale-

ment en fonction de diverses propriétés notionnelles, sémantiques, morphologiques ou syntaxiques par exemple. Les définitions seront alors remplacées par un faisceau de propriétés propres à chaque objet de telle sorte qu'une classification demeurera possible.

Il reste cependant que le statut scientifique des généralisations linguistiques que nous proposons est fragile si nous refusons toute théorie. Ce n'est pas le cas, nous plaçons notre étude dans le cadre d'une science du langage et les paires minimales, distributions complémentaires et autres tests classiques seront par exemple employés pour identifier les classes *clitique* et *déterminant*. Ces classes seront alors définies précisément par ces tests comme le veut la théorie.

Le compromis est de rendre cohérent une étude qui lie des critères sur les mêmes objets (comme les parties du discours ou les mots composés) touchant à des aspects de la linguistique qui ne se recoupent qu'en partie. Les noms communs seront distingués des noms propres par leur signifié, les déterminants seront distingués des adjectifs par leurs propriétés combinatoires, les verbes seront distingués des adjectifs par des propriétés syntaxiques. Mais une théorie linguistique unifiée qui mettrait en relation ces distinctions respectives n'est pas supposée dans cette étude. Il vient donc qu'une typologie des objets permettant de classer formellement de façon univoque les objets linguistiques est impossible à réaliser. Les frontières floues entre mots simples et mots composés, entre catégories proches seront alors discutées une à une. C'est précisément cette étude qui est abordée dans ce chapitre et qui doit nécessairement accompagner l'annotation du corpus pour que celui-ci puisse être utilisé.

Nous allons présenter les choix de segmentation en mots et mots composés dans une première partie où nous décrirons également comment s'articulent le lexique et le corpus, avant de décrire les choix linguistiques retenus pour étiqueter ces mots.

3.1 Lexique et corpus

L'ensemble des mots[1] d'un corpus constitue un vocabulaire. Ce vocabulaire tend vers un lexique dans la mesure où le corpus tend vers l'exhaustivité. Un lexique sera donc vu comme l'ensemble des *mots* d'une langue. Bien qu'idéalement infini, nous considérons qu'un lexique s'oppose à un vocabu-

1. Nous préciserons *infra* ce que nous entendons par «mot»

laire non pas parce que l'un est fini et l'autre ne l'est pas, mais parce que le vocabulaire est l'ensemble des mots du discours.

Il est nécessaire d'utiliser un lexique donné, c'est-à-dire une liste finie de mots assortis d'informations linguistiques indépendante de tout discours pour annoter un texte. Sans cela, il ne serait pas possible de marquer les mots de façon autonome pour les propriétés intrinsèques qu'ils possèdent. De plus, l'adéquation de l'annotation à un lexique assure que l'information ainsi apportée est cohérente. La catégorie (nous entendons la partie du discours mais également les traits flexionnels et autres lemmes) d'un mot se donne en langue.

Par ailleurs, l'annotation d'un corpus par l'attribution d'une catégorie morpho-syntaxique à chaque mot ne peut pas être informative si elle n'est pas donnée en discours.

Une catégorie se donne donc en langue et en discours. En langue, elle assure que l'annotation du corpus apporte une information externe et cohérente, en discours, elle assure qu'elle sera une donnée empirique exploitable du corpus.

On voit que le risque est que ces deux approches ne soient pas compatibles ; qu'une catégorie donnée en discours soit contradictoire avec le lexique. Les phénomènes de recatégorisation sont à examiner de ce point de vue.

Il est par exemple possible de construire une épithète qualificative *nominale* (c'est-à-dire dont l'entrée du lexique indique qu'elle se rapporte à un nom) et la figure est productive (*une soirée cinéma*, *un dossier clef*, etc.). Reprenons la définition que donne Michèle Noailly du substantif épithète : «tout substantif intervenant en position de N_2 dans un groupe nominal de type (Art) N_1 N_2» [Noailly, 1990]. Il est clair que l'attribution de l'étiquette N à cet élément ne dépend pas de la figure elle-même mais qu'elle est donnée par les propriétés intrinsèques de N. En d'autres termes, N est un élément du lexique tel qu'il est un nom.

Dans ce cas, nous nous trouvons devant un choix possible. Soit la fonction syntaxique d'épithète indique que N est recatégorisé en adjectif, catégorie possible pour cette position syntaxique, soit la position d'une épithète est possible pour un nom. Dans le premier cas, cela veut dire que le lexique devra contenir, pour chaque nom qui peut être mis en position d'épithète, un adjectif ambigu avec lui. Le lexique sera alors inconsistant puisque les propriétés intrinsèques du substantif seront confondues avec celles de l'adjectif.

Notre position sera donc toujours de conserver la cohérence du lexique comme donnée externe au corpus. Une catégorie sera donnée en discours

dans la seule mesure où elle se peut en langue ; qu'elle est compatible avec un lexique cohérent.

3.2 Segmentation en «mots» et «mots composés»

3.2.1 Le «mot»

Dans ce travail, nous avons étiqueté un corpus écrit, c'est-à-dire que nous avons assorti chaque unité du texte, chaque «mot», d'une étiquette. Nous allons étudier les discussions qui ont porté sur une définition du mot avant de voir ce que nous pourrons en tirer pour notre propre étude.

Les informaticiens ont eu recours à des définitions qui se prêtent plus à un traitement automatique robuste des textes informatisés.

[Silberztein, 1993] par exemple, propose une définition purement graphique et très formelle pour définir les «mots» simples du dictionnaire électronique du LADL : le DELAS.

Un mot simple est une séquence de lettres délimitée par deux séparateurs.

Cette définition brutale suppose une liste fermée de lettres dont certaines sont distinguées comme séparateurs. Ainsi, à considérer l'apostrophe, le trait d'union, les espaces, alinéas et ponctuations comme des séparateurs, *Aujourd'hui* et *court-circuit* se décomposent en 2 mots, mais *courtcircuit* en 1, *c'est-à-dire* en 4, *faits divers* en 2, *afin presque de* en 3 et *auquel* en un seul. Au vu de ces quelques exemples, nous remarquons, d'une part que cette définition peut choquer l'idée qu'on se fait habituellement du mot, nous reviendrons sur ce point, d'autre part qu'elle ne permet pas d'assigner une étiquette et une seule à chaque «mot». Plus grave, le statut linguistique du «mot» respectant cette définition n'a aucune base théorique ; le «mot» est une suite arbitraire de lettres ; c'est une unité typographique.

Tout au plus, cette définition permet d'appliquer des automatismes sur les textes comme la projection d'un lexique (liste ouverte de ces «mots»), le parcours d'un automate fini où chacune des transitions correspond à l'une de ces unités ou le comptage d'occurrences. On voit combien la simplicité de l'objet peut séduire l'informaticien. Mais le «mot» ainsi défini ne permettra pas de mettre la main sur la segmentation d'un quelconque niveau d'analyse.

Nous tenterons de trouver une base plus linguistique à la définition de l'unité choisie. En particulier, l'assignation d'une étiquette et le marquage de chaque «mot» devront être légitimes. On pourra même considérer que ces critères soient définitoires.

3.2.2 Notre approche

Nous allons présenter ce que nous appellerons «mot» pour le seul travail qui nous intéresse par facettes d'une notion qui ne se laissera pas définir par une simple formule mais par un ensemble de propriétés : lexicales, sémantiques ou syntaxiques.

Le mot comme unité lexicographique

C'est certainement l'unité qui vient le plus intuitivement à l'esprit. C'est aussi la définition proposée par les grammaires classiques qui voyaient dans le «mot» une unité sémantique.

Chacun des mots doit obtenir une définition dans le lexique et en discours. Chaque entrée du dictionnaire correspond à un lexème auquel un sens idiomatique est associé. Ainsi *pomme de terre* est un «mot» ou plus exactement correspond à une entrée lexicale. En discours, ces lexèmes (on parle alors de **vocable** comme occurrence de lexème) sont porteur d'une fonction dans la phrase et sont caractérisés par une appartenance à une «partie du discours».

Nous avons privilégié cette partie de la définition dans le lien qu'entretient le corpus avec le lexique. Le texte constitué par le corpus peut être vu comme une mise en discours de faits de langue décrits dans le lexique. Or l'analyse de ce discours en contexte ne peut être infirmé par un phénomène linguistique décrit par une théorie falsifiable sans contradiction. C'est-à-dire que le lexique doit pouvoir contenir toutes les informations du corpus annoté tout en étant cohérent avec une science du langage.

C'est pour cette raison que nous ne voulions pas recatégoriser systématiquement les mots qui avaient les emplois d'autres catégories. Un nom commun qui a la fonction d'épithète n'est pas adjectif dans le Corpus Annoté de Paris 7 non pas parce qu'il occupe une position que ne peut occuper un adjectif mais parce qu'il ne peut être adjectif dans le lexique (cette partie doit être démontrée toutefois). Par ailleurs, la fonction d'épithète n'est pas contradictoire avec la catégorie nom commun.

Ainsi, la reconnaissance d'un mot en contexte ne peut être contradictoire

avec le dictionnaire.

Une date comme *le 21 avril 1987*, par exemple, peut être vue en discours comme une suite textuelle libre dans la phrase, indécomposable, substituable par un adverbe comme *hier*. Pour autant ce ne sera pas un mot sans que le lexique ne soit incohérent. Nous n'avons donc pas fait d'unités lexicographiques de ces «locutions» mais avons proposé de les identifier grâce à une grammaire locale pour se garder d'en analyser la structure au même niveau syntaxique que les syntagmes.

Le mot comme unité graphique

Nous avons privilégié également une définition du mot comme unité graphique. Les propriétés graphiques du mot ne sont pas suffisantes en soi, puisque nous n'avons aucun critère graphique pour distinguer un mot composé d'une suite de mots simples. Nous n'avons donc pas affaire à une propriété définitoire.

Cependant nous refusions de voir plus d'un mot dans une suite de lettres non séparées par des blancs hormis une liste fermée d'amalgames qui concentrent en une unité graphique plusieurs «mots».

Cette liste est la suivante :

- au
- audit
- auquel
- aux
- auxdits
- auxquels
- auxquelles
- des
- desdits
- desquelles
- desquels
- du
- dudit
- duquel
- duquel
- ès

L'unité graphique est donc utilisée comme minimum, à l'exception des agglutinés, mais non comme un maximum puisque les mots composés sont composés de plusieurs unités graphiques.

Cette définition n'est pas forcément acceptable en grammaire générale. Par exemple les langues romanes agglutinent le pronom conjoint après le verbe (*vas a* ***darme*** *la mano (tu me donneras la main), cántamelo (chante-le-moi)* et les mots composés sont écrits en une seule graphie en allemand. Nous la rejetterons donc comme définition générale.

En français, elle est commode et permet de ne pas devoir faire une analyse morphématique des mots. Ainsi, même s'il est passionnant de savoir que *vinaigre* peut s'analyser, nous n'en faisons rien et prenons le mot comme tout autre nom commun issu d'un seul lexème.

Ainsi, nous ne disons rien des phénomènes de dérivations et de compositions qui conduisent à la formation graphique d'un terme non séparé par des blancs ou des tirets. Les graphies fluctuantes avec ou sans tiret pour les mots composés (*contrordre, contre-ordre*) ne nous inspirerons pas plus de commentaires, les uns seront «mots simples», les autres «mots composés» sans qu'il ne soit rien supposé de la composition du terme. Les composants de la forme *contrordre* ne sera pas analysée comme ce sera le cas de *contre-ordre*.

Le mot comme unité syntaxique

Nous utiliserons le terme «mot» pour désigner une unité linguistique non décomposable à un niveau d'analyse syntaxique. Nous reprenons donc la définition proposée par Maurice Grévisse à notre compte.

Le «mot» selon cette définition peut occuper à lui seul une position ou une fonction syntaxique dans la phrase. C'est-à-dire qu'il est susceptible de jouer un rôle relativement aux autres éléments.

Nous avons composé en une unité morpho-syntaxique des «mots» dans la mesure où l'ensemble de leurs constituants s'articulent par des règles de composition, de dérivation morphologiques et de flexion. Ces mots ne sont pas nécessairement continus (i.e. la locution prépositive *afin de*) mais sont atomiques du point de vue de la composition syntaxique.

Mais il n'est pas simple de savoir si les tours restrictifs en *ne...que* font intervenir deux «mots» ou si *ne, que* sont deux opérateurs morphématiques d'un même «mot». La situation est la même pour les clitiques et auxiliaires

verbaux (i.e. l'auxiliaire *aller* comme morphème de temps).[2]

L'étude de Christian Touratier ([Touratier, 1996]) met clairement en évidence des généralités morphématiques de la forme verbale. Dans son étude, les formes du singulier *je, tu, il*, etc., sont des marques formelles du verbe qui correspondent aux personnes dont les signifiants sont je...(s), tu...(s), il...(t). Les clitiques sujets, avec la finale sonore du verbe sont alors analysés comme des morphèmes de personne.

De même, les pronoms conjoints (clitiques) peuvent être analysés comme des affixes verbaux (c.f. [Miller, 1991]). Ainsi le clitique accusatif *le* dans *Jean le voit* peut être analysé comme un morphème conférant à la forme verbale les propriétés de non transitivité. Dans ce cas, aucune fonction n'est assignable au clitique et la catégorie de clitique comme «partie du discours» est discutable.

Cependant cette analyse rendrait rédhibitoire l'interrogation du corpus intéressant les compléments du verbe ou le verbe lui-même. Nous discuterons les motivations d'une classe clitique infra. en 3.3.7.

Les formes verbales composées soit avec les auxiliaires *être et avoir* soit avec les autres auxiliaires de temps (*aller, venir*) ou modaux (*devoir, falloir, pouvoir, savoir, vouloir*) n'ont pas été reconnues comme des «mots composés» dans le Corpus Annoté de Paris 7. Le mécanisme d'auxiliation contribue à la construction du passif qui échappe à l'analyse morphologique, mais également à l'apport d'informations aspectuelles, temporelles ou modales que nous n'avons pas choisi de représenter par les catégories grammaticales retenues en morpho-syntaxe. Nous devions donc écarter une analyse dont nous ne pouvions rendre compte avec le matériau choisi.

Nous avons également distingué les auxiliaires négatifs de la forme *ne, n'* non pas parce que nous considérions que les deux formes ne devaient pas être liées, mais parce que nous pensons que ce mécanisme de dépendance ne devait pas non plus être fait à ce niveau d'analyse.

2. L'étude diachronique des formes futures des langues romanes est intéressante de ce point de vue([Marchello-Nizia, 1999] p.109). Le futur se construit avec une périphrase du type *je vais chanter* en français depuis le XVI^{e} siècle ou sous un forme synthétique *je chanterai* plus anciennement. Mais la forme synthétique vient elle aussi d'une construction avec un auxiliaire (cantare habeo) et le processus est vraisemblablement cyclique. On trouve en espagnol contemporain de la région de Mexico *vadormir* pour *voy a dormir*. Le statut de l'auxiliation du verbe dans les formes futures est donc fluctuante en diachronie.

3.2.3 Mots composés

Les caractères de figement des compositions ont largement été décrits dans la littérature. Dans le cadre des lexiques-grammaires du LADL, on peut citer par exemple le recueil d'articles parus dans le numéro 90 de *Langages* [Danlos, 1988], les compositions nominales ont été particulièrement été étudiées par Gaston Gross dans [Gross, 1996]. Citons enfin l'article de Max Silberztein ([Silberztein, 1993]).

Figement

Nous écarterons de notre étude les expressions figées[3] et les collocations qui seront reconnues telles au niveau de l'analyse syntaxique.

Les expressions figées telles que *apporter de l'eau au moulin, casser sa pipe, la moutarde lui monte au nez*, etc. sont des expressions qui ne peuvent en effet se réduire à un «mot». Les propriétés de figement sont telles qu'il n'est parfois pas possible de de projeter l'ensemble des figures sur une expression figée comme nous le voyons en (d). Cependant, nous voyons que ces expressions figées ne peuvent pas se réduire à une seule catégorie ; il n'est pas possible par exemple d'appliquer la négation en (h).

(a) Jean apporte de l'eau au moulin.
(b) C'est apporter de l'eau au moulin de Luc.
(c) Jean apporte de l'eau à son moulin
(d) *C'est à son moulin que Jean a apporté de l'eau.
(e) Jean retourne souvent sa veste
(f) Sa veste, il l'a retourné souvent.
(g) Jean ne meurt pas
(h) *Jean ne casse sa pipe pas

Par ailleurs, les expressions figées ont un caractère figé sur une partie de leurs composants. La détermination, le nombre ou d'autres propriétés morpho-syntaxiques de ceux-ci peuvent être fixées par l'expression figée (par

3. Nous entendons par *expressions figées* seulement les idiotismes qui ne nous intéressent pas du point de vue de la grammaire comparée mais par leur construction. Nous rejetons donc en bloc toute autre locution, mot composé, etc., qui procèdent également de figements d'une partie de leurs constituants mais qui ne constituent pas des idiotismes

exemple le nom commun *pipe* est nécessairement au possessif singulier dans l'expression figée *casser sa pipe*)

(a) Jean casse sa pipe.

(b) • Jean casse ta pipe[4].

Mais les expressions figées ne mettent pas en œuvre une grammaire qui leur est propre. Leurs propriétés de figement devront être captées à un autre niveau d'analyse, nous pensons qu'elles ne doivent pas être représentées en morpho-syntaxe.

Les collocations comme *gros fumeur*, ne sont pas non plus des «mots» mais bien des paires de mots qui fonctionnent indépendamment en syntaxe et limitent ou modifient la portée sémantique. L'adjectif *gros* prend un sens particulier quand il modifie *fumeur*, mais la position remplie par l'adjectif n'est pas discutable. Nous pouvons modifier l'adjectif seul par exemple (*un très gros fumeur*).

Il peut se trouver que la frontière qui sépare les mots composés des collocations et des expressions figées soit fragile. Nous retiendrons que les mots composés n'ont pas de composant indépendant en syntaxe. Ce qui n'est jamais vrai des collocations et expressions figées par définition.

A nouveau les phénomènes diachroniques ne nous permettent pas de décider si un terme est un mot composé ou un phénomène de figement dans l'absolu. La plupart des termes lexicalisés aujourd'hui étaient des collocations autrefois. Par ailleurs, la terminologie propre à un domaine ajoute au lexique des termes non lexicalisés dans la langue générale (i.e. *accord salarial*, *téléphone portable*, etc.).

Caractères de figement des mots composés

Les critères que nous présentons dans cette section ne sont ni nécessaires ni suffisants pour définir les mots composés. Ils permettent simplement d'établir un faisceau de caractères du mot composé.

4. Nous notons avec un «•» une phrase grammaticale dont l'interprétation n'est pas celle normalement attendue du locuteur (ici, il s'agit de l'acception idiomatique et non littérale du verbe *casser*).

Morphologie

- Le mot composé n'est pas une expression discontinue. Certains verbes composés sont ainsi distingués d'une construction syntaxique.

  Rendre grâce à son courage

  ∗**rendre** une **grâce** à son courage

  Avoir faim

  Avoir une **faim** de loup

  en effet

  par ailleurs

  Mais la discontinuité peut se limiter à une liste très fermée d'insertions pour certains termes :

  sans doute

  sans aucun **doute**

  sans nul **doute**

  ∗sans un doute

  Cette condition n'est pas nécessaire, on trouve quelques cas de discontinuité pour des mots composés.

  compte tenu notamment **de**

  afin, toujours selon les formules de Mr Cõng, **d'**élargir la démocratie...

  font désormais **partie** de cette communauté

- L'un des composants n'apparaît que dans l'expression complexe.

  au **fur** et à mesure

  aujourd'hui

  à l'**insu** de

  faire **fi** de

  par monts et par **vaux**

- Le genre ou le nombre d'un composant change dans l'expression complexe.

  une **deux-chevaux**

  un **peau-rouge**

- Présence d'un tiret ou d'une apostrophe.

  entr'ouvert

  garde-barrière

Actualisation

Ce critère est fondamental. [Gross, 1996] le présente comme propre à définir la notion de *locution*. Il s'énonce ainsi : «on peut parler de suite composée quand aucun des éléments lexicaux constitutifs ne peut être actualisé.»

Par actualisé, il est entendu ici le passage de la langue au discours d'un élément. Il s'agit donc de quantifier les substantifs, de localiser dans l'espace et le temps les événements ou d'apporter des éléments du discours (déictiques, anaphores, etc.) dans la phrase.

Pratiquement, cela revient le plus souvent à déterminer les noms et à renseigner la flexion des verbes.

Voyons par exemple la distinction entre *verre à vin* et *verre de vin* que nous reprenons de [Abeillé, 1993].

Intuitivement nous avons affaire à une locution dans le premier cas et à la structure classique d'un nom avec son complément dans le deuxième cas. Nous pouvons faire l'hypothèse de l'effacement du déterminant *de* pour éviter la cacophonie *un verre de de vin* dans *verre de vin* puisque nous avons *un verre de ce vin*. En revanche la présence du déterminant n'est pas acceptable dans le premier exemple (**Un verre à ce vin*).

Avant de voir les possibilité du nom *vin* d'être ou non actualisé dans les deux expressions, observons que les propriétés lexicales ne sont pas les mêmes comme le montrent les exemples suivants :

(a) j'ai bu du vin
(b) j'ai bu un verre de vin
(c) ??j'ai bu un verre à vin
(d) ??j'ai cassé un verre de vin
(e) j'ai cassé un verre à vin

Il est impossible de modifier *vin* dans l'expression *verre à vin* et il est absolument impossible de l'actualiser avec un déictique ou avec un article permettant une reprise anaphorique comme le montrent les exemples suivants :

(a) un verre de ce vin-là
(b) un verre du vin qui est frelaté, ce vin...
(c) ∗un verre à ce vin-là
(d) un verre de vin est posé sur la table, le verre.../∗le vin...

(e) ∗un verre à vin rouge
(f) un verre de vin rouge

Nous pouvons ainsi identifier des locutions verbales (mais également avec un prédicat nominal ou adjectival) lorsque le complément ne peut pas porter de déterminant, non pas parce que le déterminant est effacé pour des raisons euphoniques par exemple, mais parce qu'il est impossible de quantifier le complément. Ajoutons que le déterminant peut être présent sans que le complément ne soit quantifié. Dans ce cas, le déterminant est figé. C'est le cas d'une expression comme *jouer du piano* ou *faire du vélo.*

(a) jouer du piano
(b) ∗jouer de ce piano
(c) ∗jouer d'un piano à queue
(d) faire du vélo
(e) ∗faire de ce vélo-là
(f) faire du vélo tout terrain
(g) ∗faire du vélo qui peut rouler sur les chemins

(a) Prendre une veste
(b) • Prendre sa veste
(c) à la mode
(d) • à ma mode
(e) un vice de forme
(f) • un vice de cette forme

Grammaticalisation

La grammaticalisation est un phénomène diachronique qui fait passer un élément du lexique à la grammaire. Ceci a plusieurs conséquences. Comme nous l'avons déjà vu, l'élément peut être entièrement figé dans l'expression à tel point qu'il n'existe pas indépendamment (*Dans son* ***for*** *intérieur, avoir le cœur qui bat la* ***chamade***). Il peut y avoir également un figement syntaxique d'une expression telle qu'elle ne corresponde plus à la syntaxe du français contemporain. Par exemple, les adjectifs de couleur sont toujours postposés en français contemporain, alors qu'on trouve *rouge-gorge rouge-queue* comme éléments lexicalisés. Les toponymes, qui peuvent traverser les siècles sans modification, sont souvent des exemples de constructions qui ne correspondent plus à la langue contemporaine.

Comme tout phénomène diachronique, les phénomènes de grammaticalisation suivent un continuum à tel point qu'il n'est pas toujours aisé de savoir si l'on a affaire à un élément lexicalisé ou encore entièrement productif. Par exemple, Claire Blanche-Benveniste remarque dans [Blanche-Benveniste, 1996] que le choix lexical des compléments du verbe *voir* ne se limite pas à une liste fermée alors que d'autres constructions comme *je trouve ça + adjectif + d'attendre* sont liées à une petite liste d'adjectifs.

Critères syntaxiques

Quelques critères syntaxiques permettent d'identifier un mot composé.

- S'il est toujours possible de substituer un constituant à un élément lexical selon le mécanisme d'analyse en constituants immédiats, il se trouve des substitutions vers des syntagmes qui ne sont ni complémentables ni modifiables. La catégorie syntaxique du mot composé est alors une catégorie terminale. S'il est possible de substituer un groupe de mots en une préposition par exemple ce groupe de mots est un mot composé. Avec certaines précautions, on peut faire ce test de commutation plus largement. Ces précautions consistent à analyser la structure du probable mot composé.

  En vertu de, *près de* sont des prépositions tout comme *à*, *de*.

  un tantinet, *un peu* sont des adverbes comme *environ*
- En faisant l'hypothèse qu'un syntagme projette une tête syntaxique (comme un nom pour un groupe nominal), certaines expressions sont repérées comme des mots composés car la catégorie syntaxique de l'ensemble ne correspond pas à la séquence de catégories des constituants.

  peut-être:ADV (verbe + verbe)

  c'est_à_dire:CC (clitique+verbe+préposition+verbe)

  rendez-vous:N (verbe clitique)

  garde_côtes:N (verbe + nom)
- La séquence des catégories des constituants d'un mot composé n'existe pas par ailleurs.

  y_compris:ADV (Clitique - Verbe au participe passé)

  à la va vite:ADV (Prep - Det - Vconj - Adv)
- Les composants sont contigus. Seules quelques petites insertions sont possibles et sont limitées à quelques catégories (en général un petit adverbe ou adjectif ou alors une incise phrastique).

  à force de

un maillot **doré** deux pièces
*un maillot que ma sœur porte doré deux pièces
à l'insu **justement** de ses voisins
*à l'insu, comme on dit ici, de ses voisins
afin, comme le dit son professeur, de lui donner sa chance
??afin, son professeur le dit souvent, de lui donner sa chance

– Les règles d'accord ne sont pas respectées

Une grand mère
*Une grande mère
La grand-rue
A grand peine

Critères sémantiques

– La sémantique du mot composé n'est pas la composition sémantique des composants.

une sage-femme n'est pas une femme qui est sage.

– On ne peut pas remplacer un des constituants par un synonyme ou un antonyme.
  (a) à_bas
  (b) *à_haut
  (c) à_la_va_vite
  (d) *à_la_va_rapidement
– Le mot composé peut parfois correspondre à une seule unité dans d'autres langues.

une pomme_de_terre ⇒ a potato / una patata
en_arrière ⇒ behind / backward / indietro

3.3 Classes de «mots»

Il s'agit de proposer une typologie des *mots* tels qu'ils ont été définis plus haut. La contrainte imposée ici est que cette typologie soit classificatrice, c'est-à-dire que chaque acception de chaque terme soit associée à une classe unique. C'est à ce prix qu'il sera possible d'étiqueter systématiquement et de façon univoque le corpus.

Les classes distributionnelles de Léonard Bloomfield appliquées à l'analyse de discours comme l'a proposé Zellig Harris sembleraient proposer le terrain formel requis, car ces classes sont, à proprement parler, des classes d'équivalence. Les termes qui les composent sont les segments substituables dans des contextes identiques. Et deux termes sont équivalents s'ils participent au même paradigme.

Une classe distributionnelle correspond au regroupement de distributions établies par le contexte, sans qu'il ne soit jamais question d'autres facteurs, en particulier du sens des termes.

Ainsi, la classe des déterminants[5] en français regroupera des termes qui appartenaient à des parties du discours différentes de la grammaire *Générale et raisonnée* ou grammaire de Port-Royal [Arnauld & Lancelot, 1676] : pronom (et adjectif) pour «mon», adjectif pour «ces», «tel», «aucun», article pour «le». Cette classe sera définie par l'ensemble des positions possibles d'un déterminant français et rien d'autre.

Il est remarquable que ces classes recoupent dans leur plus grand nombre les parties du discours issues de la tradition aristotélicienne. Les définitions notionnelles[6], flexionnelles[7] et fonctionnelles[8] qui permettent d'établir la *nature* d'un terme sont le plus souvent corrélées à une position syntaxique. Nous n'entrerons pas dans une discussion disant dans quelle mesure cette corrélation est possible ou même nécessaire, ni dans celle qui décrit une position syntaxique du point de vue d'une classe distributionnelle. Mais comme le remarque Wilmet ([Wilmet, 1997]) dans *Grammaire critique du français*, les neuf parties du discours traditionnelles (nom, adjectif, article, pronom, verbe, adverbe, préposition, conjonction, interjection) correspondent en partie à des classes distributionnelles.

La refonte de la grammaire du Bon Usage [Grévisse, 1964] par Goosse [Grévisse, 1993] a contribué à aller dans le sens de cette correspondance en proposant, comme c'est aujourd'hui l'usage, d'éclater des classes de mots appartenant manifestement à des classes distributionnelles multiples.

– Les articles et adjectifs appartenant aux mêmes distributions ont été classés parmi les déterminants, rendant la classe des articles obsolète comme partie du discours.

5. Telle que nous entendons qu'elle constitue une classe distributionnelle

6. Qui permet de distinguer par exemple le nom (onoma) du verbe (rhemma) comme dénotant respectivement la substance (y compris celle d'une idée) et le procès ou l'état

7. Qui permet de distinguer les mots «variables» des mots «invariables»

8. Qui décrivent les relations des termes composant la phrase

- Les conjonctions ont été distinguées comme coordonnants et subordonnants
- La classe des introducteurs a été établie.

La grammaire du Bon Usage conserve toutefois une définition principalement notionnelle des classes de mots et rejoint la tradition grammaticale des parties du discours. Nous reprendrons, avec quelques ajustements, ces parties du discours pour établir notre jeu d'étiquettes assorties aux mots du corpus.

3.3.1 Catégories retenues

Personne

La personne marque les pronoms, les déterminants personnels et les verbes. En langue, cette catégorie est intrinsèque aux mots personnels. La modification des noms par ceux-là permet de construire une référence en discours relative à la situation de communication. Cette information est également pertinente en discours puisqu'elle suffit parfois à construire les liens de dépendances entre le verbe et son sujet.

Genre

Le genre est une catégorie arbitraire du nom. Même s'il subsiste une alternance des genres pour les animés qui correspond à un signifié propre à désigner le sexe (*un*+*une* camarade, *un acteur*+*une actrice*), cela n'est pas nécessaire (*une sentinelle*, *un manequin*).

Cette catégorie marque l'adjectif, le verbe participe passé ou le déterminant. L'accord en genre permet d'articuler le nom et ses compléments et modifieurs au sein du groupe nominal. Il participe à l'économie du système également en marquant l'accord entre le verbe et l'attribut, l'objet (pour les participes transitifs) ou le sujet (pour les participes intransitifs ou pronominaux) tout comme le nombre et la personne. La forme non marquée en genre est le masculin. Nous rejetons l'idée que le neutre puisse exister en français comme dans d'autres langues (le latin par exemple). Les pronoms *ça, il* (impersonnel) seront identifiés comme des termes non marqués en genre ; c'est-à-dire masculin.

Catégorie ou Partie du discours	Description
Nom	Nom commun, propre et cardinal
Adjectif	Adjectif qualificatif, indéfini, interrogatif, ordinal et possessif
Adverbe	Adverbes et divers connecteurs et négatifs
Préposition	Préposition
Déterminant	Déterminant, dont article.
Clitique	Pronom clitique (ou atone, ou conjoint)
Pronom	Pronom personnel. Pronom relatif. Pronom indéfini. Pronom interrogatif
Conjonction	Conjonction de coordination, Conjonction de subordination
Interjection	Interjection, mot-phrase
Verbe	Verbe - présentatifs *voici voilà*
Mot étranger	Mot d'une langue étrangère
Ponctuation	Ponctuations forte et faible

FIG. 3.1 – *Parties du discours retenues*

Nombre

Le nombre participe également de l'accord entre le nom ou du verbe et leurs satellites comme nous venons de le dire. Contrairement au genre, cette catégorie n'est pas intrinsèque au nom (sauf quelques exceptions : *des ciseaux*, *des lunettes*, *des obsèques*), et est porteuse de sens. Nous ne décrirons pas ce sens ni ses emplois, nous nous limiterons à l'annotation du nombre lorsque les marques morphologiques le permettent. L'annotation de cette catégorie

dans le Corpus Annoté de Paris 7 se limitera d'ailleurs à l'identification d'une forme marquée en nombre et non au sens de pluralité. *Ce* sera identifié comme un pronom singulier dans les tours *Ce:CL3ms sont des amis de toujours* par exemple.

Temps et Mode

Le verbe porte une marque permettant de situer un évènement ou un état dans le temps et la modalité. Cette marque n'est cependant ni suffisante ni nécessaire. Les auxiliaires verbaux, les types de détermination des compléments du verbes, la diathèse sont par exemples acteurs dans ses propriétés énonciatives. Au niveau d'annotation qui nous intéresse, nous marquerons seulement les informations morphologiques de temps et mode qui accompagnent les verbes.

Cas

Le français conserve quelques formes casuelles sur les pronoms. Les cas de l'ancien français portant sur les noms n'a absolument plus de pertinence. Les anciennes variations en cas régime/cas sujet, quand elle apparaissent encore, font apparaître aujourd'hui des entrées lexicales différentes avec une variation de sens (*gars/garçon*). En revanche, les pronoms clitiques et relatifs portent les cas nominatif (ou ergatif) (i.e. *il, je, tu*), accusatif (i.e. *le, nous*), datif (i.e. lui, nous), génitif (i.e. *dont, en*) et locatif (i.e. *y, en*). Nous limiterons notre annotation sur les seuls pronoms clitiques aux sujets (pour nominatif ou ergatif), objets (pour accusatif et datif). De plus nous marquerons le clitique réflexif comme tel.

3.3.2 Nom

Le nom est régulièrement précédé d'un déterminant et est modifié par des compléments et épithètes ou relatives. Le nom peut être sujet, attribut, complément ou apposition. On le trouve également comme épithète (*Un ingénieur maison.*)

Le nom est porteur d'un genre qui lui est intrinsèquement attribué et peut varier en nombre.

La définition du nom comme «partie du discours» fait apparaître la notion de «substance». Un nom est ce par quoi on désigne les choses, idées, êtres

animés. Nous présenterons la différence entre nom commun et nom propre en 3.7.4.

Dans une acception classique ([Arnauld & Lancelot, 1676]), le nom est ce qui est substance ou accident (la manière des choses). Les noms sont alors distingués comme *noms substantifs* et *noms adjectifs*. Cette distinction logique fondée sur le sens s'est accompagnée d'une autre notion logique. L'adjectif (le nom adjectif) porte une «connotation» distincte de la substance ; une «signification confuse d'une chose à laquelle ces substances se rapportent»(ibid. p.26)

«Car, parce que la substance est ce qui subsiste par soi-même, on a appelé noms substantifs tous ceux qui subsistent par eux-même dans le discours, sans avoir besoin d'un autre nom, encore même qu'ils signifient des accidents. Et au contraire on a appelé adjectifs ceux même qui signifient des substances, lorsque par leur manière de signifier ils doivent être joints à d'autres noms dans le discours.» (ibid. p. 26)

Cette distinction a été retenue pour faire des noms deux parties du discours : les substantifs et les adjectifs. On voit qu'une définition notionnelle de ces classes est fragile : l'adjectif pouvant dénoter la substance, le nom une propriété aliénable.

Nous retiendrons cette distinction mais nous mettrons de coté sa valeur logique. Ainsi, comme c'est l'usage aujourd'hui, nous distinguerons les noms et adjectifs par leurs propriétés fonctionnelles. Une difficulté se pose alors pour les noms épithètes (*la branche information, une élection grand-public, un niveau record, les pays frères, un scénario catastrophe, une jupe culotte, reine fantôme*)

Remarquons que les noms épithètes ne s'accordent pas en genre avec le nom auquel il se rattache (*un scénario catastrophe*, *une température record*, *une reine fantôme*) et ne s'accordent pas toujours en nombre (*des gateaux maison*, *des températures record*, mais *des femmes enfants*). D'autres s'accordent en nombre et en genre tout comme les adjectifs (*Des pays frères*, *des républiques (sœurs+*frères)*. La forme adjectivale n'existe pas pour ces termes, ce qui veut dire qu'ils correspondent à des noms dans le lexique et nous n'avons pas tenu à proposer un homonyme adjectival de ces noms.

Cette décision peut être motivée par l'emploi épithète de noms qui ne sont pas ambigus avec leur forme adjectivale. Par dérivation du nom *catastrophe* on fait l'adjectif *catastrophique*. Il existe donc bien une construction *Un scénario catastrophique* où le mot *catastrophique* a le sens de *catastrophe* ajouté d'une «connotation» par laquelle on dit le rapport des deux termes. Il

existe par ailleurs une construction en complément de nom *Un scénario de catastrophe.* Mais la «connotation» par laquelle s'exprime l'emploi de l'épithète nominale n'est pas couverte par ces deux constructions.

Remarquons que certaines de ces épithètes sont lexicalisées et n'ont plus le sens du substantif (*un gateau maison*). Nous n'avons pas fait le choix de les recatégoriser sans apporter d'argument à ce choix qui devient alors arbitraire.

3.3.3 Adjectif

Nous avons déjà présenté une définition des adjectifs avec les noms.

L'adjectif est attribut ou épithète. Il ne peut jamais se trouver seul, c'est-à-dire sans dépendance syntaxique. Il peut être modifié par un adverbe.

L'adjectif, comme le nom, peut varier en nombre. Il s'accorde en genre avec le nom qu'il modifie ; c'est-à-dire le nom ou pronom auxquels il sert d'épithète ou du sujet ou du complément d'objet auxquels il sert d'attribut ([Grévisse, 1993] §548). Certains adjectifs sont invariables, c'est le cas de quelques adjectifs issus d'adverbes (*la roue avant, la porte arrière, une femme bien*) ou d'emprunts (*des gens incognito.*)

Nous présentons en 3.6.1 la fragile frontière qui sépare les adjectifs des participes passés.

La définition logique des adjectifs faisant apparaître leur qualité «nominale» doit être élargie pour considérer des adjectifs indéfinis, ordinaux, cardinaux, possessifs et interrogatifs. Une distinction devait être explicite entre les adjectifs «noms adjectifs» et les adjectifs qui n'ont plus rien de commun avec les noms. Nous avons désigné par «qualificatif» les «noms adjectifs».

Les adjectifs sont alors des quantifiants-caractérisants selon la terminologie de Marc Wilmet ([Wilmet, 1997]) en place de prédéterminant (*Tous les hommes, Trois de ces hommes, Aucun de ces hommes*).

3.3.4 Adverbe

Régulièrement les adverbes n'ont pas de trait d'accord. Il existe une fausse exception : l'adverbe *tout* se note *toute* devant un son consonantique (*Des pommes toutes mouillées (très mouillées, dans leur ensemble), Des femmes tout habillées, Des femmes toutes joyeuses*). Il faut moins voir un cas d'accord ici que la manifestation graphique d'une prononciation de /tut/ devant une

forme au féminin[9]. Une règle a été érigée pour rendre compte de cette prononciation faisant intervenir également le nombre. On peut voir en effet qu'il n'y a aucun accord quand la forme au féminin ne suit pas immédiatement *tout* :

(a) Elles étaient toutes joyeuses (ambigu : sens 1 et 2)
(b) Elles étaient toutes aussi joyeuses (sens 1)
(c) Elles étaient tout aussi joyeuses (sens 2)

Nous n'avons pas distingué une «valeur adverbiale» pour certains adjectifs (emploi de *seul(es)* dans les tours *seules les candidates qui ont une tenue réglementaire participeront*). Nous dirons que seul(es) dans cette position est polysémique et peut avoir le sens de *il n'y a que les candidates qui...* ou le sens (pas très naturel dans cet exemple) *solitairement, les candidates qui ont une ...*

Les adverbes sont susceptibles de modifier les phrases, les verbes, les adjectifs et les adverbes eux-mêmes.

3.3.5 Préposition

Les prépositions n'ont aucun trait d'accord. Elles se distinguent des autres parties du discours par leur propriétés syntaxiques. Elles régissent toujours un complément nominal ou phrastique infinitif ou participe.

Nous n'avons pas distingué la catégorie distributionnelle des complémenteurs car trop complexe à identifier et dépendante de théories syntaxiques comme la Grammaire Générative. Nous verrons en 3.9.3 que nous n'avons donc pas distingué les mots *de* dans (*Il promet de venir, Il rêve de venir.*) Où Hélène Huot ([Huot, 1981]) distingue un complémenteur d'une préposition comme introducteur d'infinitives.

La catégorie des prépositions est donc assez hétérogène, on y trouve des complémenteurs comme nous venons de le signaler et quelques *introducteurs*

9. Nous n'étudierons pas cette règle phonétique. Nous nous interrogeons sur la nature de l'accord qu'on peut observer dans les tours précieux *Des roses **fraîches** écloses, des fenêtre **grandes** ouvertes.* Nous observons que ces adverbes doivent précéder des sons vocaliques et qu'en absence «d'accord», nous aurions les prononciations suivantes : /frɛzecloz/ ou /frɛecloz/, /grãtuvɛrt/.

ou *présentatifs* :

(a) *de* dans les tours *Et Luc* ***de*** *s'écrier Noël! Noël!*
(b) *à* avant dans les interjections (***Au*** *feu!* ***À*** *moi!*)
(c) **Il y a** comme présentatif (***Il y a*** *un mois qu'il n'est pas venu*)
(d) **Voici** en tête de circonstancielle (***Voici*** *trois mois qu'il est parti.*)

La forme *en* du gérondif n'est pas non plus distinguée.

Les prépositions se distinguent par leur propriété syntaxique de régir un complément, Nous n'avons donc pas catégorisé les prépositions qui supposent l'analyse d'une ellipse. Ces termes sont des adverbes dans le Corpus Annoté de Paris 7 comme nous le verrons en 3.6.5.

Notons l'exception de la position d'antéposition : la préposition *durant* peut apparaître avant son complément (*Trois années durant*) dans un registre de langue assez élevé.

Les prépositions sont en très grand nombre des locutions formées d'adverbes ou de noms suivi de *de* (*afin de, près de, au bord de, à propos de, etc.*)

Notons que quelques signes propres à l'analyse de corpus ont été annotés préposition par leur propriétés d'introducteur et leur position syntaxique. Il s'agit des signes arithmétiques (+, −) et de quelques abréviations latines relatives à l'organisation des textes (*i.e.*, *c.f.*, *p.s.*)

Les signes autonomes (notes de bas de pages, renvois de texte (*§5*, *pp.6-8*, *p.8*) ont été annotés nom commun.

3.3.6 Déterminant

Les déterminants sont les articles définis, indéfinis et partitifs, les *adjectifs* possessifs et démonstratifs de la grammaire traditionnelle mais également des mots indéfinis (*quelque, autre*, etc.), des mots exclamatifs et interrogatifs (*quel(le)(s)*), les cardinaux, les déterminants négatifs (*nul, aucun*, etc.) et enfin les déterminants relatifs formés autour de *lequel*.

Le déterminant s'accorde en genre et en nombre avec le nom dont il dépend.

Nous voyons que cette classe regroupe des termes dont les propriétés sémantiques ne permettent pas toujours de les distinguer des autres parties du discours.

Cependant la propriété sémantique qui semble commune à tous les déterminants est l'actualisation du nom : « ils assurent son passage de la langue dans le discours, tout en formant avec lui des expressions référentielles qui désignent des occurrences particulières de la notion attachée lexicalement au nom. Ils spécifient notamment si cette notion renvoie à des entités massives ou comptables, saisies de manière singulière ou plurielle, partitive ou globale, etc.» ([Riegel *et al.*, 1994] p. 152.)

Sans entrer dans les détails qui réclameraient un large développement, cette propriété d'*actualisation* ne se limite pas au seuls déterminants mais à l'ensemble de la structure du groupe nominal. Cette dimension sémantique des déterminants est certainement à rapporter à la distinction que fait Jean-Claude Milner (in [Milner, 1978]) entre référence virtuelle *qui concerne les unités lexicales hors tout emploi particulier* (p.332), de la référence actuelle, *relation unissant un objet du monde et sa désignation (ibid.)*

Enfin, des déterminants peuvent être présents dans des groupes nominaux non actualisés comme ceux qui apparaissent dans les expressions idiomatiques :

(a) Jules prend **le** taureau par les cornes.
(b) Marie prend **ses** jambes à **son** coup.
(c) Jule joue **du** piano.

Le niveau d'analyse qui nous intéresse ici ne nous permet pas de considérer un groupe déterminant défini par ses propriétés syntaxiques comme les syntagmes que nous surlignons en gras (**Toutes les** pommes, **Trois de mes** amis, **Entre deux et trois** francs). Ces groupes qui jouent le rôle de *spécifieur* de nom ne seront pas identifiés. En revanche, nous annoterons des déterminants composés (Nombres composés, *Beaucoup de, La plupart des*, etc.)

Le déterminant sera donc défini par ses seules propriétés distributionnelles : il précède un nom et n'est jamais seul.

Nous excluons les suites de déterminants. Les termes qui participent à l'actualisation et à la quantification du nom et qui précèdent ou suivent le déterminant seront annotés adjectifs.

(a) Les/Dét. trois/Adj. amis.
(b) Trois/Adj. de/Prép. mes/Dét. amis.
(c) Tous/Adj. les/Dét. amis.

En conséquence, la classe des *prédéterminants* participant à la syntaxe du groupe déterminant sera ignoré.

3.3.7 Clitique

Les pronoms conjoints ou clitiques ne sont classiquement pas distingués des autres pronoms. Cependant leurs propriétés prosodiques, syntaxiques et distributives nous conduisent à en faire une classe à part. En outre, les clitiques n'ont pas toujours été analysés comme des groupes nominaux mais comme des affixes verbaux. Nous avons déjà signalé l'analyse morphématique des verbes par Christian Touratier qui inclut dans la conjugaison verbale le clitique nominatif comme morphème de personne. Cette analyse est particulièrement motivée dans les langues romanes comme l'espagnol qui ne possèdent pas de pronom sujet clitique mais un morphème agglutiné au verbe : un suffixe verbal.

Rappelons les tests de Zwicky et Pullum ([Zwicky & Pullum, 1983]), repris par Philippe Miller ([Miller, 1991]) qui ont permis d'analyser ces «mots» comme des affixes verbaux et non comme des pronoms. Nous reprenons ces tests de Anne Abeillé (communication non publiée.)

1. Les affixes sélectionnent la catégorie de leur hôte. Les mots sont moins sélectifs.
   (a) Il faut tout lui dire
   (b) *Il faut lui tout dire
   (c) donne-moi la pomme
   (d) *donne la pomme -moi
2. Les affixes ont une distribution idiosyncratique ; les mots sont plus réguliers.
   (a) Pierre (y+*∅) ira.
   (b) Pierre (*y+∅) va.
3. Les affixes ont un ordre fixe ; les mots sont plus libres.
   (a) Je le lui donne.
   (b) *Je lui le donne.
   (c) Pierre donne une pomme à Léa
   (d) Pierre donne à Léa une pomme.
4. Les mots ont une indépendance phonologique, pas les affixes.
   (a) Pierre en a (/ãna/+*/ãa/)

5. Les affixes n'ont pas portée large sur des hôtes coordonnés, tandis que les mots le peuvent.
   (a) Pierre regarde et écoute les enfants.
   (b) Pierre les regarde et (les+$*\emptyset$) écoute.

Ajoutons qu'il est bien connu que les clitiques ne peuvent pas être accentués contrairement aux autres groupes nominaux et aux pronoms forts.

Nous n'avons pas fait d'analyse prosodique, morphologique et phonologique des données textuelles pour annoter le corpus de Paris 7. Il était donc naturel de ne pas considérer le clitique verbal comme un affixe ainsi que le suggèrent les linguistes sus-nommés. Les propriétés positionnelles mentionnées comme argument en faveur de l'analyse morphologique mettent en évidence la distribution particulière de ces «mots». L'annotation des clitiques comme tels permettra d'en faire une analyse morphologique. L'un des buts du projet d'annotation du corpus de Paris 7 est de rendre le corpus distribuable, y compris pour des analyses morphématiques.

Nous avons donc classé les pronoms clitiques selon ces critères. L'analyse transformationnelle voit dans ces formes les réalisations de surface issue du déplacement de groupes nominaux ((c.f. Richard S. Kayne in [Kayne, 1975] §2.2.) Cette analyse met en évidence les propriétés syntaxiques des clitiques comme leur montée avec certains verbes.

Exemple repris de [Kayne, 1975]

(a) Il les fera manger à son fils
(b) *Il fera les manger à son fils

En grammaire transformationnelle, les pronoms clitiques gardent un trait relatif à leur fonction syntaxique (le cas). Cette fonction est marquée par une préposition ou simplement la position syntaxique (sujet ou objet) du groupe nominal déplacé en français contemporain. Sans même adopter l'analyse transformationnelle, les clitiques sont marqués par les cas nominatif, accusatif, datif, génitif et locatif. Ce trait est directement fonction de la place occupée par le clitique (un clitique datif à la troisième personne ne peut jamais précéder un clitique accusatif par exemple.) C'est pour cette raison que nous devions annoter le cas des clitique. De plus, la catégorie casuelle, qui intéresse également les groupes nominaux pleins, est une propriété lexicale dans le sens où elle est intrinsèquement liée aux pronoms clitiques.

Par souci de simplicité, nous n'avons pas annoté tous les cas mais seulement ceux qui présentent les ambiguïtés qui intéressent les études syntaxiques (sur la valence verbale par exemple.)

Les cas retenus ont été annotés :

- **sujet** pour nominatif
- **objet** pour accusatif et datif
- **réfléchis**

Nous dressons la liste des pronoms clitiques figure 3.2.

Les propriétés distributionnelles des clitiques se résument à ceci : leur place est fixe, ils sont nécessairement conjoints aux verbes qui ne se coordonnent pas isolément, ils *montent* avec certains verbes.

3.3.8 Pronom

Les pronoms, si l'on en exclut les clitiques comme nous le proposons, se comportent comme un équivalent fonctionnel d'un groupe nominal. Ils prennent alors un sens anaphorique, cataphorique ou déictique. Le pronom relatif, qui occupe souvent une place à part dans les grammaires, n'échappe pas à cette description.

Le pronom est marqué en genre, en nombre et en personne. Le genre *neutre* admis dans certaines grammaires comme celle de Martin Riegel *et al.* pour les formes *soi, ceci, cela*, etc. sera remplacé par le genre masculin. Le français ne distingue pas un genre neutre d'un genre marquant le sexe comme nous le feront remarquer à propos du genre des titres en 3.7.4. Le genre des noms qui ne désignent pas des êtres animés est arbitraire en français et le sens abstrait non animé ou non humain des formes *soi, ceci, cela* ne doit pas s'accompagner d'un genre neutre selon nous. Nous avons donc choisi d'associer le genre masculin comme forme non marquée des pronoms «neutres».

La personne des pronoms non personnels a été arbitrairement marquée par la troisième personne. Cette convention permet de rendre l'annotation plus homogène mais n'est pas motivée par un supposé trait personnel des formes nominales impersonnelles.

Notons que nous n'avons pas classé les déterminants relatifs parmi les pronoms car ceux-ci, bien qu'apportant un sens anaphorique n'ont aucune autonomie syntaxique (*J'ai perdu mon billet,* ***lequel*** *billet est indispensable pour rentrer* .)

(a) J'ai perdu mon billet, peux-tu me donner **le tien**.
(b) J'ai perdu le billet **auquel** je tenais tant.
(c) Je cherche **ceci**.

(d) Je ne cherche **personne**.
(e) Je ne cherche **rien** ni **personne**.
(f) J'en cherche **deux** qui sachent parler anglais.

Les pronoms se distinguent des clitiques par leur relative autonomie syntaxique. Il peuvent se coordonner (d) et peuvent être modifiés (f). Cependant, les pronoms se comportent comme des groupes nominaux déterminés. Quand un déterminant précède un pronom, il est restreint et ne participe pas à l'actualisation du référent comme le montrent les exemples (b) et (c).

(a) *J'ai perdu mon billet, peux-tu me donner **ce tien**.
(b) ?? J'ai perdu mon billet, peux-tu me donner le **tien**$_i$. Ce$_i$ billet me serait bien utile.
(c) J'ai perdu mon billet$_i$, peux-tu me donner le **tien**$_i$. Ce$_i$ billet me serait bien utile.
(d) J'ai perdu mon billet bleu, peux-tu me donner l'**autre**$_i$/adjectif. Ce$_i$ billet me serait bien utile

Les pronoms sont sous-catégorisés par leurs propriétés syntaxiques et sémantiques. Nous distinguons les pronoms relatifs, interrogatifs, démonstratifs, personnels, possessifs, indéfinis et cardinaux.

Le pronom relatif est un mot simple parmi : *qui, que, quoi, dont, où, lequel, lesquels, laquelle, lesquelles.)* Nous n'avons pas fait un pronom relatif complexe des prépositions suivies des formes *lequel, laquelle, lesquels et lesquelles* bien que le relatif composé occupe une fonction dans la relative en (a).

(a) Le chien avec lequel il aime jouer.
(b) Il aime jouer avec ce chien.
(c) *Il aime jouer ce chien avec.

Le relatif occupe une fonction dans la relative comme nous l'avons fait remarquer, ceci le distingue de la conjonction de subordination homographe *que* qui introduit également une phrase. L'identification de la relative, phrase épithète d'un nom[10], permet de distinguer le pronom relatif du pronom interrogatif.

10. Notons que la relative peut ne pas avoir d'antécédent ou modifier des adjectifs et adverbes. Dans les premiers cas, on peut analyser ceci comme une ellipse nominale, quand la relative modifie un adverbe, celui-ci se limite aux adverbes de lieu et de temps et est

Nous étudierons les mots relatifs et interrogatifs en 3.7.6.

Les pronoms personnels et démonstratifs sont des pronoms déictiques. Ils n'ont pas d'antécédent de 1re ou 2e personne dans la phrase mais renvoient à un sens lié à la situation d'énonciation.

Nous étudierons les mots démonstratifs en 3.7.7. Les pronoms personnels en 3.7.5.

Nous étudierons les mots possessifs dont les pronoms possessifs en 3.7.10 et les mots démonstratifs en 3.7.7.

Les pronoms indéfinis sont ceux qui marquent une quantité (ils sont toutefois distingués des pronoms cardinaux) comme *plusieurs, beaucoup, aucun* ou une «identification imprécise (*quelque chose*) ou même un refus d'identification (tel).» ([Grévisse, 1993].)

3.3.9 Conjonction

Les conjonctions de subordination et les conjonctions de coordination partagent classiquement la même catégorie. Nous nous sommes conformé à cette tradition bien que leurs propriétés syntaxiques ne soient pas les mêmes.

Cette tradition est conforme également à l'étude récente de Mireille Piot [Piot, 1993] sur une classe plus englobante de connecteurs qui comprennent, outre les conjonctions de subordination et conjonctions de coordination, les adverbes conjonctifs (*donc, pourtant, par conséquent, néanmoins*, etc.) Mireille Piot fonde son étude sur les propriétés syntaxiques de ces connecteurs que nous reprenons en partie. Les exemples sont repris en partie de [Piot, 1993].

- La conjonction de subordination introduit une phrase subordonnée.
- La conjonction de subordination autorise la permutabilité de la phrase introduite.

  (a) Pierre est venu, parce que nous étions là.
  (b) Parce que nous étions là Pierre est venu.

ainsi un «substantif» (*ici, à présent, aujourd'hui, maintenant*, etc.).

(a) Fier qu'il était d'être médecin.

(b) Maintenant que son frère est parti.

(c) Heureux qui ne boit pas.

(c) Pierre est venu, mais nous étions là.
(d) *Mais nous étions là Pierre est venu.

– La coordination de deux subordonnées permet le remplacement du second *connecteur* par que.

(a) Marie est partie quand nous étions là et (quand+que) Jacques est arrivé.
(b) Marie a demandé quand nous étions là et (quand+*que) Pierre viendrait.

Ce test souffre cependant de quelques exceptions concernant les comparatives :

(a) Marie parle comme elle lit et (comme+*qu') elle chante.

Nous ne l'avons pas retenu à cause de cette dernière raison et parce que trop complexe à mettre en œuvre pour l'annotation du corpus.

– Il est parfois possible de substituer la phrase subordonnée par le pronom *cela* ou par une nominalisation de cette phrase. Dans ce cas, le connecteur a un emploi prépositionnel ou adverbial :

(a) Jean est parti après que nous sommes arrivés.
(b) Jean est parti après (notre arrivée+cela.)

– Il est impossible d'effacer le sujet de la phrase subordonnée, alors que cela est possible pour une phrase coordonnée à l'exception de *car* et *or*.
– La conjonction de coordination a une position contrainte. Elle introduit nécessairement la phrase coordonnée.

(a) (Alors que+*Et+*Mais) Marie est sortie, Pierre est venu.
(b) Marie est sortie, (Alors que+Et+Mais) Pierre est venu.

Nous n'avons pas retenu le test souvent cité et repris de [Harris, 1968] selon lequel les termes coordonnés doivent être de même «nature» et de même «rang». Cela n'est pas toujours vrai si «nature» signifie catégorie syntaxique :

(a) Pierre est médecin et fier de l'être.
(b) Pierre sait l'âge de Marie et qu'elle ne pourra pas s'inscrire à cet compétition.

3.3.10 Interjection

Les interjections sont les «mots phrases» invariables, c'est-à-dire les mots qui construisent à eux seuls des phrases (*bonjour, bravo !, oui, non, zut !*, etc.) L'interjection n'est pas nécessairement isolée dans la phrase et peut être un substantif (comme une onomatopée) ou un discours rapporté.

(a) Son Danton, publié en 1931 ne fut, **hélas** ! jamais joué de son vivant.
(b) Ils ne sont pas si nombreux à lui dire «**non**!».
(c) **Dame oui** !

3.3.11 Verbe

Le verbe se distingue classiquement des autres parties du discours par les marques flexionnelles, en plus de la personne, de temps, d'aspect et de mode.

Les participes passés ne marquent pas la personne mais le nombre et le genre comme les adjectifs. Les infinitifs et participes présents ne marquent pas la personne.

Nous n'avons pas annoté les formes composées des verbes. Ainsi, le mode, la voix, le temps et même l'aspect des formes composées échappent à notre annotation.

De plus, nous avons ajouté les présentatifs *voici, voilà* à cette catégorie.

La conjugaison verbale permet d'identifier les formes finies. Les traits sémantiques d'aspect et de temps permettent d'identifier les formes non finies ambiguës avec les adjectifs comme nous le verrons en 3.6.1. Cependant ces traits ne sont pas nécessaires.

Le verbe est tête de phrase matrice ou enchâssée.

Le verbe peut être précédé ou suivi de clitiques comme nous l'avons vu à propos de cette dernière catégorie.

Sous-catégorisation des verbes

La sous-catégorisation verbale concerne traditionnellement le type de catégories pouvant être complément ou même sujet du verbe. La valence verbale (pouvant inclure les adverbes (*aller bien, mal tourner*, etc.), la restriction de sélection des compléments, la nature du figement des compléments,

les propriétés de montée ou de contrôle des infinitives sont autant de sous-catégorisations permettant de dresser des classes verbales comme le fait Maurice Gross dans [Gross, 1968]. Ces descriptions échappent à l'annotation morpho-syntaxique du corpus car elles impliquent une analyse syntaxique en constituants et dépendances dont il n'est pas question dans cette partie de l'annotation. La sous-catégorisation verbale n'en est pas moins un phénomène lexical.

Certains verbes non autonomes se distinguent par leur propriétés syntaxiques. *être, avoir* se combinent avec les participes passés et infinitifs, *Aller, venir* avec les infinitifs pour marquer le temps et l'aspect. *devoir, falloir, pouvoir, savoir, vouloir* se combinent avec d'autres verbes pour marquer la valeur modale.

Nous n'avons pas distingué ces verbes car cela nous conduisait à faire une analyse de l'auxiliation qui n'est pas supposée pour l'annotation du corpus. Comme le note Hava Bat-Zeev Shyldkrot dans [Shyldkrot, 1999], la liste des verbes définis comme auxiliaires n'est jamais la même et la distinction entre auxiliaire et verbe support n'est pas évidente (*Marie fait du ski, Marie prend congé. Marie fait sortir Luc.*).

Nous n'avons donc pas distingué les verbes entre eux bien que nous sommes conscients que leurs propriétés lexicales sont parfois fort différentes.

En revanche, les propriétés lexicales de sous-catégorisation des verbes nous permettent de les identifier comme tels.

Rappelons que nous avons segmenté en un seul «mot» les verbes dont les compléments figés ne forment pas des groupes nominaux mais des noms (sans déterminant, ou non actualisés) ou adverbes. Ces constructions échappent en effet à la construction générale de la complémentation d'un verbe. Cependant, nous n'avons pas annoté les figements verbaux dans toute leur généralité au niveau morpho-syntaxique.

(a) Jean **a mal**. (∗un mal)

(b) Jean **fait fi** des reproches qu'on lui a dit.

(c) Jean **a beau jeu** de rire de sa bêtise. (∗ le beau jeu)

(d) Luc **joue du piano**. (∗de ce piano)

(e) Luc **a peur**. (∗la peur + une peur bleue)

(f) Luc **a la trouille**. (∗une trouille + ∗cette trouille)

3.3.12 Mot étranger

Nous étiquetons «ET» les mots étrangers qui ne sont pas entrés dans la langue française et qui sont employés dans un contexte syntaxique peu clair. Sinon, ils ont les mêmes étiquettes que les mots français.

(a) Un:Dms match:NCms de:P football:NCms.
(b) Un:Dms building:NCms
(c) La:Dfs condition:NCfs sine_qua_non:Afs
(d) El:ET contemporaneo:ET.
(e) The:ET history:ET

3.3.13 Ponctuation

Nous présenterons les ponctuations avec les autres signes et expressions numériques dont l'annotation de corpus rend l'étude caractéristique en 3.7.12.

3.4 Comparaison avec d'autres jeux d'étiquettes

Nous allons examiner les seuls points de divergence entre les conventions qui ont été exposées ici pour l'annotation du Corpus Annoté de Paris 7 et plusieurs projets existant d'annotation morpho-syntaxique de corpus français.

Nous reprendrons les conventions exposées pour les recommandations Multext d'encodage morpho-syntaxique dans le cadre du projet GRACE par Josette Leconte ([Lecomte, 1997], revues par Jean Véronis pour le projet Multitag. Pour simplifier, nous parlerons de conventions «Grace».

Nous comparerons également nos conventions avec le guide d'annotation d'Ursula von Rekowski ([von Rekowski, 1996]) de spécification et classification lexicale issues des recommandations Eagles.

Rappelons que l'annotation du corpus de Paris 7 a été inspirée de ces conventions et recommandations bien qu'il y ait quelques points de divergence.

3.4.1 Pronoms

Nous avons distingué les pronoms clitiques des autres pronoms *personnels*. Les deux catégories font partie de la même classe dans les projets Eagles et GRACE.

En et *y* ont été classé parmi les pronoms personnels dans ces deux projets bien que le trait personnel y soit absent et qu'il s'agisse plutôt de formes dont le cas est génitif ou locatif.

Les formes fortes des pronoms personnels ont été annotés comme pronoms réfléchis dans les constructions *compter sur soi, travailler pour soi-même, se laver soi-même*, etc. dans les deux projets. Nous n'avons annoté les réfléchis que pour les formes clitiques. Il nous semble en effet que cet aspect de l'annotation des pronoms forts est propre à une analyse de la construction verbale.

(a) Je ne compte que sur **moi-même**.
(b) Elle$_i$ ne travaille que pour **elle**$_i$.

En (b), la distinction entre un pronom réfléchi et un objet du verbe pour la forme *elle* dépend d'une interprétation de la forme verbale. *Travailler pour soi* est une forme verbale figée et saturée dans le cas du réfléchi, une forme construite dans le cas d'un objet (comme pour *Luc travaille pour son patron*). Nous n'avons pas fait ce genre d'analyse à ce niveau de l'analyse du corpus.

3.4.2 Adjectifs

Les adjectifs ambigus avec les participes sont annotés grâce à une sous-classe des adjectifs qualificatifs dans le projet Grace : les *adjectifs-participes* notée Afp.

Exemples repris de [Lecomte, 1997]

(a) J'ai trouvé la porte ouverte/Afp
(b) La porte est restée ouverte/Afp

Nous nous sommes expliqué de la raison qui ne nous a pas fait choisir une telle étiquette *supra*. Ajoutons que la distinction entre cette étiquette et celle des verbes et adjectifs reste délicate dans certains cas comme nous l'illustrons en (a), (b) et (c). En (a) le verbe est clairement identifié par un complément d'agent, en (c), l'ambiguïté conduirait à assigner l'étiquette *Afp*.

Mais en (b), il est difficile de trancher entre un verbe et cette étiquette.

(a) Une voiture garée par un chauffard.
(b) Une voiture garée de travers.
(c) Le magasin est juste après la voiture garée.

Les adjectif indéfinis ont été limités à une liste fermée: *même(s), certain(e)(s), tel(le)(s) et aucun(e)(s). quelconque* n'a pas été distingué dans sa position attribut (*Cet homme est vraiment quelconque*, versus *Un quelconque emploi lui conviendrait*). Enfin les indéfinis qui participent à la détermination des noms ont été classés parmi les déterminants. Nous en avons fait des adjectifs indéfinis (*seul, différents, divers, maints*, etc.)

Les *meilleur, moindre, pire* ont été distingués par un trait *comparatif* qui n'est pertinent que pour ces seuls adjectifs. Nous n'avons pas fait de telle sous-classe dans le corpus de Paris 7.

Les adjectifs employés comme adverbes ont été recatégorisés dans le projet Multext. Nous verrons en 3.6.4 les raisons qui nous ont conduit à refuser ces recatégorisations systématiques.

Nous expliquerons également pourquoi les noms en position d'épithète ne sont pas plus recatégorisés que les adjectifs en position de substantif. (*Les petits vieux, Les nouveaux venus, les petits nouveaux* (exemples repris de [Lecomte, 1997]). Rappelons que l'encodage morpho-syntaxique se fait selon le rôle des termes en contextes dans le projet GRACE/Multext. Même s'il est incompatible avec le lexique (presque tous les adjectifs qualificatifs peuvent se trouver en position nominale par ellipse du nom, et un grand nombre de noms peuvent être épithètes d'un autre).

3.4.3 Adverbes

Le clitique négatif *ne* a été distingué comme particule des autres adverbes négatifs dans les projet Grace et Eagles. Nous n'en avons rien fait mais il n'est pas difficile d'identifier ce terme lors de l'exploitation du corpus par sa forme et sa position contrainte.

Dans le projet Grace, le trait *Degré* distingue les adverbes qui entrent dans les construction comparatives (*pis, davantage, mieux, moins* et *plus.*) L'adverbe négatif *que* dans les tours *ne ... que* est également marqué par un trait pour le distinguer. *Seul* en initiale a été catégorisé adverbe dans le projet Multext. Nous l'étudions en 3.7.2.

3.4.4 Conjonctions

Un grand nombre de coordonnants substituables par une autre conjonction de coordination ont été annotés comme tel dans le corpus de Paris 7 (*c'est-à-dire, comme, sinon, tantôt*, etc.). Un grand nombre de ces termes sont annotés subordonnants dans le projet Grace-Multext.

3.4.5 Interjections

Les interjections ont été distinguées des onomatopées dans le projet Multext. Nous n'avons pas fait cette distinction et avons regroupé tous les mots phrases et autres exclamations, susceptibles de faire une phrase indépendante ou non.

(a) Un multipartisme naissant, mais ô(Interjection) combien !

(b) Ouais(Interjection), il est complètement avachi.

3.4.6 Résidus

Cette catégorie, présente dans presque tous les jeux d'étiquettes et dans les deux projets n'a pas été retenue pour le corpus de Paris 7 puisque nous prétendons avoir distribué tous les mots dans les parties du discours proposées.

3.4.7 Articles

Dans le projet Eagles, les articles forment une classe distincte des autres déterminants. Nous avons noté les articles comme déterminants définis et indéfinis dans le corpus de Paris 7.

3.4.8 Adposition

Les prépositions sont les seules sous-types d'une classe «adposition» pour le français dans le projet Eagles.

3.5 Autres classes de mots

Nous n'avons pas retenu des classes de mots parfois admises dans des théories syntaxiques diverses.

3.5.1 Complémenteur

La position de complémenteur est admise pour les termes qui occupent la position d'introducteur (spécifieur en théorie GB) de phrase. On y trouve les classiques conjonctions de subordination mais également les «prépositions» introduisant les infinitives. Les tests permettant de mettre en évidence une telle classe sont délicats et dépendent d'une théorie précise. De plus les phrases admettent un spécifieur qui n'a pas de réalisation de surface ; ce que nous n'avons jamais noté dans le corpus Annoté de Paris 7.

3.5.2 Connecteur

Mireille Piot (citée *supra*) dresse une classe de connecteurs dont les conjonctions et certains adverbes conjonctifs. S'ajoutent à ces connecteurs les relatifs ou les introducteurs et même les verbes copules à tel point que la classe est parfois vivement critiquée car mal définie :

« Connecteurs, subsumant les anciens «mots de liaison», «conjonctions de coordination» et autres «adverbes initiaux de phrases», voire «conjonction de subordination» et «pronoms relatifs» est un de ces mots magiques par lesquels la linguistique moderne croit assurer son prestige en même temps qu'elle se débarrasse de problèmes grammaticaux gênants.» ([Wilmet, 1997] §27).

3.5.3 Auxiliaires

Nous nous sommes expliqué de la raison pour laquelle nous n'avons pas retenu cette classe pour distinguer les auxiliaires verbaux des autres verbes. Nous réservons cette annotation au niveau syntaxique comme valence verbale. Remarquons que l'auxiliation ne se limite pas aux seuls verbes selon certains linguistes (Danielle Leeman, Anne Daladier in [Leeman, 1999], [Daladier, 1999]). Danielle Leeman remarque par exemple les propriétés temporelles et aspectuelles de certaines prépositions (*dans, pendant*) qui sont à mettre en parallèle avec les temps des verbes.

3.5.4 Prédéterminants

Les déterminants semblent être tête d'un groupe spécifieur du groupe nominal. La construction de ce syntagme réserve une position de prédéterminant que nous n'avons pas retenue.

Tous les enfants.
Les trois enfants.
Aucun des trois enfants.
Ces trois enfants.
Ces quelques fleurs.
Bien de la peine.
La plupart de mes amis.

3.6 Critères de choix entre catégories

Nous présentons les critères de choix entre catégories sur la base des ambiguïtés les plus fréquentes en corpus.

3.6.1 Adjectif / Participe passé

Le verbe s'identifie par sa sous-catégorisation [11] et par les traits sémantiques qui affectent le procès ou l'énonciation d'un procès (le temps, l'aspect, le mode.)

Nous n'avons pas regroupé, au niveau d'une analyse morpho-syntaxique, l'auxiliaire verbal et le verbe au participe comme une unité verbale composée. Nous nous en expliquons supra en 3.2.2. Le participe s'identifie alors par la simple présence de l'auxiliaire de temps. On a donc un participe sans ambiguïté possible après un auxiliaire de temps, au passif, avec un complément en *par* (ou un complément d'agent en *de*).

Quand on a les mêmes compléments que le verbe conjugué on a un participe passé.

En position épithète ou attribut, on a généralement un adjectif mais sa détermination est parfois délicate. Le type de modification peut lever cette

11. Nous nous garderons de confondre la sous-catégorisation d'un prédicat qui indique ses dépendances syntaxiques de la sous-catégorisation grammaticale qui décrit plus généralement toute propriété commune à un sous-ensemble d'une même classe.

ambiguïté. Quand la forme est modifiée par *très*, c'est généralement un adjectif, quand elle est modifiée par *mal*, c'est généralement un participe[12]. En effet, ces adverbes se distinguent comme des adverbes qui sont en position de modifieurs de verbes ou d'adjectifs de façon quasi-univoque.

La morphologie dérivationnelle peut également aider à lever l'ambiguïté : les adjectifs et les verbes ne construisent pas les formes dérivées de la même façon. Les adjectifs déclinent des formes dérivées avec les préfixes *in (inconnu, invendu, insatisfait), il(illimité)*, les verbes avec les préfixes *re (redoré, reboisé), dé (déboisé, décoiffé)*. Nous pouvons utiliser ce test (qui consiste à analyser le préfixe ou à construire une nouvelle entrée par dérivation) pour savoir si le terme est un adjectif ou un verbe.

Nous voyons à partir de cet exercice de morphologie dérivationnelle que le sens contribuera largement à lever l'ambiguïté. Le sens compositionnel opéré par le préfixe n'est pas le même pour un adjectif ou un verbe. Les préfixes *re, dé* par exemple permettent de modifier le sens du verbe selon la nature du procès. ∗*Infaire* pour *défaire* par exemple, n'est pas seulement agrammatical du point de vue d'une construction morphologique mais réellement asémantique car le préfixe *in* ne peut opérer sur un procès ou un événement mais seulement sur un substantif.

Ainsi l'identification d'un événement sera déterminante, pour lever l'ambiguïté. Si l'on peut paraphraser avec une relative à l'actif par exemple, il s'agit du verbe au participe.

(a) Un:Dms jugement:NCms prononcé:VKms (qu'on a prononcé)
(b) Un:Dms recul:NCms prononcé:Ams de:P les:Dmp prix:NCmp (∗qu'on a prononcé)

(a) Il:CL3ms a:VP3s parlé:VKms
(b) Je:CL1fs me:CLR1fs suis:VP1s trompée:VKfs
(c) Elle:CL3fs est:VP3s appréciée:VKfs de:P tous:PROmp
(d) une canalisation bouchée:Afs
(e) laissé:VKms pour:P mort:Ams
(f) rester:VW bouleversé:Ams
(g) la porte est fermée:Afs (résultatif ou état permanent)
(h) la porte est fermée:VKfs tous les soirs à 8 heures (par le gardien) (sens passif)
(i) Un:Dms parking:NCms privé:Ams

12. Il existe cependant des contre-exemples comme *Elle est très déçue par ce film*

(j) Un:Dms enfant:NCms privé:VKms de:P dessert:NCms

Dans le cas où l'événement n'est pas un procès mais confère au verbe un sens statif et tous les critères permettant d'identifier le verbe de façon triviale sont absents, la chose est moins simple. Dans ce cas, l'intuition de l'annotateur est appelée pour connaître le sens et savoir par exemple si le terme est dit pour une propriété inaliénable de l'objet, auquel cas il sera adjectif, ou correspond à un état dont il est question, auquel cas le terme sera un verbe.

Dans les cas inextricables, nous pouvons penser que l'ambiguïté est bien artificielle et correspond à une classification des parties du discours qui ne tient pas compte du continuum existant à leurs frontières. Bref qu'elle n'est qu'un artefact de notre grammaire. Dans ces cas, d'autres expériences (Action GRACE par exemple) ont choisi une étiquette distinguée pour désigner ce qui est indifféremment verbe ou participe passé. Nous avons choisi dans le Corpus de Paris 7 d'annoter l'un ou l'autre indifféremment disant que le choix est tout simplement arbitraire.

3.6.2 Adjectif / Participe présent

Le participe présent est invariable contrairement à l'adjectif. Le verbe se repère, comme le participe passé grâce à ses compléments.

(a) En:P lisant:VG le:Dms journal:NCms...
(b) Les:Dfp personnes:NCfp résidant:VG en:P France:NPfs depuis:P moins:ADV de:P 3:Dmp mois:NCmp
(c) Les:Dfp erreurs:NCfp existantes:Afp
(d) Un:Dms déséquilibre:NCms persistant:Ams

3.6.3 Adjectif / Nom commun

Les noms peuvent avoir la fonction d'épithètes ou d'attributs. Nous ne les analysons donc pas comme des termes recatégorisés dans ces positions ; faisant classiquement de ces noms des adjectifs pour la seule raison que les adjectifs seuls sont susceptibles d'occuper ces fonctions.

Nous n'analysons pas non plus comme une recatégorisation ce phénomène de grammaticalisation décrit dans [Marchello-Nizia, 1999] des noms en position de préposition : *question, tendance, côté, etc.* Ils gardent la catégorie

nom dans notre corpus. En effet, nous voulons ne pas augmenter le dictionnaire de prépositions, adjectifs par la seule raison que les substantifs peuvent prendre des fonctions qui leur sont habituellement attribuées.

(a) une veste:NCfs sport:NCms
(b) Il est très famille:NCfs
(c) côté:NCms famille:NCfs il est très réservé
(d) le fichier:NCms clients:NCmp
(e) tendance:NCfs baisse:NCfs de:P les:Dfp prix:NCfp
(f) Fille d'une mère:NCfs écrivain:NCms.
(g) Un député:NCms RPR:NPms

En cas d'ellipse du nom, nous ne recatégorisons pas les adjectifs sauf changement de sens ou de morphologie. La motivation de ceci est encore de ne pas augmenter notre dictionnaire. Nous ne supposons pas pour autant qu'un élément modifié par l'adjectif est effacé à un certain niveau d'analyse ; ce qui impliquerait l'annotation de traces en surface. Nous disons simplement qu'un adjectif peut être dans cette position sans que le substantif soit présent. L'adjectif se distinguera alors nettement du substantif par le sens. En effet, l'actualisation de l'adjectif est impossible. On repérera ces adjectifs par l'invariabilité de leur déterminant par exemple comme le montrent les exemples suivant :

(a) Je mise sur le deux
(b) Je mise sur des deux toute la soirée
(c) Je veux les deux
(d) ∗Je veux des deux (sens indéfini)

(a) une:Dfs grande:Afs bleue:Afs
(b) je veux les:Dmp deux:Amp
(c) il a perdu le:Dms gauche:Ams
(d) ce:Dms dernier:Ams
(e) le:Dms premier:Ams
(f) c':CL3ms était:VI3s le:Dms soir:NCms de:P la:Dfs première:NCfs (sens théâtre)
(g) la:Dfs gauche:NCfs a:VP3s gagné:VKms
(h) il:CL3ms sera:VF3s le:Dms meilleur:Ams
(i) le:Dms rouge:NCms est une belle couleur

(j) 10:Dmp mètres:NCmp de:P haut:NCms
(k) faire:VW son:Dms possible:NCms
(l) faire:VW l':Dms impossible:NCms
(m) aller:VW à:P l':Dms essentiel:NCms
(n) le:Dms mieux:NCms serait de ... (autrement, mieux est ADV et non A)
(o) à:P la:Dfs française:Afs
(p) le:Dms français:NCms Bull:NPms (les adjectifs de nationalité ne peuvent pas être antéposés)
(q) le:Dms japonais:NCms Sony:NPms
(r) Dans:P l':Dms immobilier:NCms
(s) Le:Dms secteur:NCms immobilier:NCms
(t) Le:Dms vide:NCms
(u) Une:Dfs chaise:NCfs de:P vide:Afs

La modification du «substantif» qui porte sur la qualité (dans un tour superlatif, comparatif), affectera un adjectif et non un nom.

(a) l'homme:NCms le:Dms plus:ADV grand:Ams
(b) l'homme:NCms aussi:ADV grand:Ams que:CS lui:PROms
(c) le:Dms plus:ADV grand:Ams de:P les:Dmp hommes:NCmp

3.6.4 Adjectif / Adverbe

Nous ne recatégorisons pas les adjectifs en fonction adverbiale. *Faux, juste, fort, etc.* sont adjectifs dans les expressions *chanter faux, crier fort, tomber juste* parce qu'ils n'ont pas la même distribution qu'un adverbe comme le montrent les exemples (a) à (g). L'adjectif est invariablement au masculin singulier.

(a) tomber juste
(b) ∗tomber justement
(c) faussement aimable
(d) ∗faux aimable
(e) il parle fort
(f) il parle fortement
mais :
(g) il a fortement parlé

(h) ∗il a fort parlé

haut, bas sont adjectifs ou noms communs, jamais adverbe.

(a) en:P haut:NCms de:P
(b) vers:P le:Dms haut:NCms
(c) tout:ADV en:P haut:NCms
(d) un meuble:NCms haut:Ams
(e) rêver:VW tout:ADV haut:Ams
(f) parler:VW bas:Ams
(g) manger:VW léger:Ams
(h) rouler:VW propre:Ams
(i) Le:Dms plus:ADV vite:ADV possible:Ams

fort et **juste** sont adverbes seulement comme prémodifieurs d'un adjectif, d'un adverbe ou d'une préposition.

(a) il était:VI3s fort:ADV triste:Ams
(b) fort:ADV justement:ADV
(c) un homme fort:Ams
(d) parler:VW trop:ADV fort:Ams
(e) Ils sont venus juste:ADV à_temps:ADV
(f) Nous avons payé la juste:Afs somme:NCfs

bien, mal ne sont jamais adjectifs, mais seulement adverbes ou noms communs.

(a) Le:Dms bien:NCms
(b) bien:ADV de_les:Dfp années:NCfp plus:ADV tard:ADV
(c) tout:PROms est:VP3s bien:ADV
(d) quelqu'un:PROms de:P bien:ADV
(e) bien:ADV mûr:Ams
(f) bien:ADV évidemment:ADV
(g) bien-sûr:ADV
(h) bien-entendu:ADV
(i) bien-que:CS
(j) Il:CL3ms aime:VP3s bien:ADV qu':CS on:CL3ms le:CL3ms flatte:VS3s

loin, près sont toujours des adverbes ou font partie de prépositions composées.

(a) loin_de:P chez:P toi:PROms
(b) près_de:P 10:Dmp milliards:NCmp
(c) voir:VW loin:ADV
(d) il:CL3ms est:VP3s loin:ADV le:Dms temps:NCms où:P...
(e) il:CL3ms est:VP3s de_loin:ADV le:Dms meilleur:NCms

sauf est adjectif ou préposition, mais jamais adverbe. *Sauf à* est une préposition composée seulement devant un verbe à l'infinitif.

(a) l':Dms honneur:NCms est:VP3s sauf:Ams
(b) Il:CL3ms est:VP3s sain_et_sauf:Ams
(c) Tous:Amp les:Dmp fruits:NCmp sauf:P les:Dfp pommes:NCfp
(d) Sauf_à:P prétendre:VW y:CL3ms comprendre:VW quelque chose
(e) Sauf:P à:P Paris:NPs

3.6.5 Préposition / Adverbe

Une préposition prend toujours un complément ; sans complément elle est adverbe. Nous avons en effet défini la préposition selon les critères syntaxiques de complémentation.

(a) je vote pour:ADV
(b) je vote pour:P la:Dfs gauche:NCfs
(c) Avant:ADV il:CL3ms y:CL3ms avait:VI3s ici:ADV un:Dms restaurant:NCms
(d) tu:CL2ms partiras:VF2s avant:P la:Dfs fin:NCfs
(e) depuis:ADV il:CL3ms a:VP3s progressé:VKms

durant est toujours préposition même quand son complément est antéposé. Dans l'histoire de la langue, il était prédicat et sous-catégorisait une phrase ([Grévisse, 1993] §1011). Le terme appartient à la langue soignée, c'est peut-être ce qui explique cet emploi conservateur. Comme *excepté* ou plus anciennement *pendant* on le trouve parfois accordé car sa grammaticalisation n'est pas entièrement fixée.

(a) durant:P trois:Dfp heures:NCfp

(b) trois:Dfp heures:NCfp durant:P

Pour les autres prépositions de temps, on peut avoir un prémodifieur nominal compatible avec un complément après la préposition. En l'absence de ce dernier, l'étiquette est adverbe.

(a) trois:Dfp heures:NCfp avant:P la:Dfs fin:NCfs
(b) trois:Dfp heures:NCfp avant:ADV

Remarquons qu'un grand nombre de locutions conjonctives en *que* sont construites avec une préposition. Nous les analysons comme des composées (*depuis que, pendant que, pour que,* etc.). De même que les locutions prépositives construites avec *de*: *avant de, afin de, avant de,* etc.

Voici, voilà sont prépositions seulement comme tête de complément circonstanciel, sinon ils sont verbes. En effet, comme les verbes ils se construisent avec un pronom conjoint accusatif (*le voici, nous voici* ou un groupe nominal (voire une relative sans antécédent) objet post-posé *voici Jean voici de quoi me convaincre* ou une complétive *voici que tout s'écroule sous ses pieds.* De plus, ce prédicat — bien qu'il n'ait aucun trait aspectuel ou temporel — constitue le noyau d'une phrase, nous l'avons donc classé parmi les verbes. Comme préposition, *voici* et *voilà* peuvent se substituer à d'autres prépositions dont *il y a.*

(a) Voici:P quelques:Dmp mois:NCmp, il:CL3ms nous:CLO1p a:VP3s dit:VKms ... (Au_début_du:P mois il nous a dit, Il_y_a:P quelques mois, il nous a dit)
(b) nous:CLO1mp voici:VP3s en pleine tragédie
(c) son:Dfs arrivée:NCfs voici:P 3:Dmp ans:NCmp
(d) Voilà:VP3s sans_doute:ADV le seul sens donné à notre existence.
(e) Nous:CLO1mp voici:VP3s !
(f) l'homme:NCms que:PROR3ms voici:VP3s
(g) je pense:VP1s que:CS voici:VP3s une:Dfs occasion:NCfs unique:Afs
(h) un premier jugement voici:P trois:Dmp ans:NCms l':CL3ms avait:VI3s débotté:VKms

il y a est une locution prépositive comme tête de complément circonstanciel. Sinon, la suite — fort figée il est vrai — est analysée comme le verbe *avoir* précédé de deux clitiques.

(a) ils:CL3mp sont:VP3p partis:VKmp il_y_a:P 3:Dmp ans:NCmp

(b) il:CL3ms y:CL3ms a:VP3s 3:Dmp ans:NCmp qu':CS ils sont partis

hors est toujours préposition. Il se compose avec *de* pour faire une préposition composée.

(a) Les:Dmp prix:NCmp hors:P énergie:NCfs
(b) hors_de:P ma:Dfs vue:NCfs

3.6.6 Préfixes / Adverbe

L'étiquette «PREF» est employée pour catégoriser des mots du type : *méta-, anti-, franco-, auto-, super-*. De nombreux mots appartenant à une terminologie peuvent donner lieu à un préfixe (noms de pays, de formations politiques, d'éléments chimiques et corpusculaires, etc.). Ces préfixes s'ajoutent à une liste fermée de mots indépendants qui peuvent dans certains cas être des préfixes (*extra, ultra, super*) et à une liste fermée de morphèmes grammaticaux (*ex, pré, re*, etc.).

La frontière entre une analyse morphologique par dérivation préfixale d'un «mot» et son analyse par composition syntaxique n'est pas nette. Il est clair que l'entrée dans le lexique d'un mot construit par dérivation est un phénomène diachronique. Les mots comme *ex-URSS, pré-sélection, présélection, hypersensible* s'analysent encore aujourd'hui comme des compositions alors que *dissymétrie, (dyssymétrie ?), dysfonctionnement, restructuration* sont entièrement lexicalisés. Nous avons vu que notre notion de «mot» recouvre une analyse graphique du français écrit. Nous écarterons donc brutalement tous les phénomènes de préfixation qui ne conduisent pas à marquer par un trait d'union le préfixe. La présence du trait d'union n'est pas toujours obligatoire, nous nous bornerons alors à observer ce que les auteurs du texte du corpus ont noté.

En revanche, quand le trait d'union est présent, il s'agira de déterminer si le terme complet est un «mot composé» ou non.

Le préfixe a un ensemble de traits communs avec certains adverbes. Il a la même fonction de modifieur d'adjectifs ou de verbe, mais contrairement à l'adverbe il est conjoint à celui-ci. (*Un lot presqu'attribué, Un lot presque mieux attribué*). Comme modifieur de noms, le terme sera étiqueté préfixe et non adverbe ; dans cette position, le préfixe sera en effet assimilé à un adjectif.

l':Dfs ex-:PREF Union_soviétique:NPfs

l':Dms ex-:PREF mari:NCms de:P son:Dms client:NCms
Cet:Dms appareil:NCms est:VP3s auto-:PREF régulé:VKms
Le:Dms super-:PREF PDG:NCms et:CC son:Dms directeur_général:NCms

Quasi est préfixe devant un nom, adverbe devant un adjectif ou un adverbe. *Non* est préfixe devant un nom, adverbe devant un adverbe, une préposition ou un participe.

La:Dfs quasi-:PREF totalité:NCfs
Une:Dfs situation:NCfs quasi-:ADV parfaite:Afs
La:Dfs situation:NCfs de:P non:PREF reprise:NCfs
Les:Dmp lots:NCmp non:ADV attribués:VKmp
Non:ADV par:P méchanceté:NCfs
Non:ADV plus:ADV

Outre est toujours une préposition ou une partie d'un adverbe composé.

Outre-Manche:ADV
Il:CL3ms est:VP3s parti:VKms outre-Rhin:ADV
Outre:P les:Dfp mesures:NCfp d':P usage:NCms
Outre-mesure:ADV

3.7 Les ambiguïtés entre sous-catégories les plus fréquentes

3.7.1 Adjectif qualificatif ou cardinal

Neuf est adjectif qualificatif ou cardinal. Le qualificatif est variable en genre (*neuve*). Le numéral cardinal sert à quantifier le nom mais également, au même titre que l'ordinal, à distinguer un élément d'une énumération (*le tome deux, le deuxième tome*). *Neuf* est la seule forme ambiguë entre les sous-catégories qualificatif et cardinal. Nous distinguons l'adjectif cardinal en l'annotant «AC» et non «A».

(a) un:Dms livre:NCms neuf:Ams

(b) mes:Dmp neuf:ACmp livres:NCmp
(c) il est neuf:Ams
(d) ils sont neuf:ACmp
(e) ils sont neufs:Amp

3.7.2 Adjectifs qualificatifs ou indéfinis

L'adjectif *Seul* (*Seul*, *Seule*, *Seules*) est rarement qualificatif. Le plus souvent, il est adjectif indéfini (noté «AI»). En position de prédéterminant ou d'épithète, il est indéfini. L'adjectif épithète détaché à valeur adverbiale est également indéfini. En attribut ou en épithète post-posé il est qualificatif. Dans ce dernier cas, il ne prend pas la valeur d'un quantifiant d'unicité mais d'une propriété du substantif modifié. Cette différence est marquée dans : (*Seule Martine est allée au cinéma Martine est allée seule au cinéma.*

(a) un homme seul:Ams
(b) un seul:AIms homme
(c) seule:AIfs cette:Dfs femme:NCfs
(d) avec:P pour:P seul:AIms but:NCms de:P l':CL3ms entendre:VW
(e) seule:AIfs cette:Dfs femme:NCfs est:VP3s allée:VKfs à:P le:Dms Tibet:NPms

Divers, *diverse* et *différents* sont ambigus entre déterminant, adjectifs qualificatifs (attribut ou épithète droit) noté «A», ou indéfini (épithète gauche) notés «AI». Comme épithète gauche, il complète le déterminant avec un sens voisin de «plusieurs» pour indiquer la pluralité de personnes, de choses qui ne sont pas les mêmes ([Grévisse, 1993] §612).

(a) divers:Dmp conseils:NCmp
(b) des:Dfp aventures:NCfp diverses:Afp et:CC variées:Afp
(c) ces:Dfp diverses:AIfp péripéties:NCfp
(d) tes:Dmp différents:AImp problèmes:NCmp
(e) des:Dmp problèmes:NCmp différents:Amp
(f) ils:CL3mp sont:VP3p différents:Amp
(g) elles:CL3fp sont:VP3p diverses:Afp

Certain est adjectif qualificatif avec le sens de *sûr*, *déterminé* ; sinon il est indéfini.

Autre, *quelconque* sont adjectifs qualificatifs en position détachée et en attribut. En épithète ils sont indéfinis. (cf 3.7.9)

(a) Cet homme semble quelconque
(b) Un quelconque personnage
(c) Un autre
(d) Il aurait été autre s'il était parti

3.7.3 Conjonctions de coordination

Les conjonctions de coordination sont les suivantes: *et, ou, mais, donc, or, ni, car, soit, c'est-à-dire, voire, sinon, tantôt...tantôt, puis.* Cette liste peut surprendre car elle contient des connecteurs «adverbiaux» comme *donc, c'est-à-dire, voire*, nous avons expliqué ce point de vue en 3.3.9.

ou, et, mais, ni , c'est_à_dire ne sont pas ambigus.

car peut être NCms (un car).

donc est plus souvent adverbe (adverbe de phrase). Comme coordonnant, il peut être remplacé par *c'est-à-dire.*

(a) Il:CL3ms est:VP3s donc:ADV venu:VKms
(b) donc:ADV il:CL3ms viendra:VF3s
(c) il est honnête:Ams donc:CC pauvre:Ams

soit est verbe comme tête d'un phrase, sinon il est coordonnant.

(a) où:ADV qu:PROR3ms il:CL3ms soit:VS3s
(b) soit:CC chance:NCfs soit:CC hasard:NCms
(c) 100:Dmp dollars:NCmp soit:CC 580:Dmp francs:NCmp

3.7.4 Noms communs et noms propres

Plusieurs critères coexistent pour lever l'ambiguïté entre un nom propre et un nom commun: la graphie (une lettre en majuscule est à l'initiale des noms propres), la syntaxe (un nom commun requiert un déterminant), la sémantique (nom propre dépourvu de sens lexical mais ayant un signifié absolu).

Nous allons détailler ces critères pour mieux les exploiter mais précisons d'ores et déjà qu'aucun n'est nécessaire à la levée d'ambiguïté :

- Un nom commun peut s'écrire avec une capitale par emphase, pour les abréviations et acronymes et aussi pour désigner un produit en nommant une marque ou une société par métonymie (*Ce «Responsable» n'y connaît rien !*, *M. Dupont*, *Un Frigidaire mal lavé*).
- Un nom propre peut s'écrire avec une bas de casse[13] dans les textes de typographie peu soignée (*La pérestroika*, *l'est*).
- Un nom propre peut être précédé d'un déterminant. Ce déterminant est contraint comme le souligne Riegel ([Riegel *et al.*, 1994]) à être l'article défini. (*Le Rhin*, *Le Monde*, *Ce soir les Durand invitent.*)
- Un nom commun peut ne pas être précédé d'un déterminant en attribut ou apposition (*Jean est facteur*, *facteur, Jean ne le sera jamais*), en complément non actualisé d'expressions verbales (*faire mouche, avoir idée que*), par effacement cacophonique de *de* (*le médecin parle de difficultés chroniques*) et dans certains compléments prépositionnels (*Ça tourne par effet de Bernouilli*).

Pour expliquer le fait qu'un nom propre peut dans certains contextes ne pas désigner un référent unique, nous dirons qu'il est employé comme nom commun. Cela n'est pas circulaire ; le contexte seul ne permet pas d'identifier un nom propre. Dire qu'il est employé comme nom commun signifie que ses propriétés intrinsèques sont inchangées mais que son emploi est celui d'un nom commun.

Nous avons vu qu'une des propriétés que se partagent les noms propres est d'être dépourvu de sens lexical ([Riegel *et al.*, 1994] p.175). Cela signifie qu'ils «n'entretiennent pas de relations sémantiques (ex. de synonymie, d'hyperonymie ou d'antonymie) et ne sont pas susceptibles d'une définition au sens ordinaire du terme.» (ibid.)

Le nom propre est donc un **désignateur rigide** qui dénote un référent unique sans que cette opération de «signification» ne soit modifiable de quelque façon que ce soit. On ne peut quantifier ou qualifier un référent en employant un nom propre.

En revanche, nous admettrons que les noms propres ont un «sens» qui est modifiable en discours. Le sens d'un nom propre est précisément le mode de désignation qu'il opère. Nous n'entrons pas plus avant dans cette discussion sinon pour remarquer que certains noms propres n'ont de sens que dans une

13. Minuscule

situation d'énonciation donnée ; que *Martin*, par exemple, ne désigne rien qu'en un lieu et un instant précis. Nous admettrons qu'en synchronie, dans le cadre des travaux qui nous intéressent ici, les noms propres sont toujours des désignateurs rigides.

Le «sens» d'un nom propre contrairement à son signifié est susceptible d'être l'objet d'une opération sémantique. *Un Picasso* désigne un tableau peint par Picasso par métonymie, mais ne désigne pas l'actualisation du nom propre. Ainsi on peut renvoyer à une classe d'individus nommés par un nom propre, à des propriétés propres à l'individu nommé et à des emplois métonymiques et métaphoriques du «sens» du nom propre.

Nous dirons pour trancher qu'un nom propre n'est pas susceptible d'actualisation. Ainsi la paire minimale suivante trouve son explication par le fait que *Ferrari* désigne toute voiture sortie des usines du constructeur italien, mais que *Prost* ne désigne que lui-même bien que son «sens» soit employé par métonymie.

(a) Une Ferrari particulièrement rapide. Cette Ferrari a passé la ligne d'arrivée...
(b) ∗Un Prost particulièrement rapide. Ce Prost a passé la ligne d'arrivée...

Prénoms, Noms: NP

Les noms, prénoms sont toujours des noms propres même lorsqu'ils sont employés pour désigner une famille, ou les propriétés d'un personnage nommé. La particule des noms propres fait partie du nom, nous accordons le genre des noms propres quand le contexte le permet (par exemple quand le prénom est suivi du nom).

Quand le prénom ou le nom prend un sens tout autre et correspond à un mot entièrement lexicalisé, il s'agit évidemment d'un nom commun (*un jules, un césar.*)

(a) Jean:NPms de_Broglie:NPms
(b) Les:Dmp Dupont:NPmp
(c) tous:Amp les:Dmp Martins:NPmp de la terre
(d) E.:NPms P.:NPms Thompson :NPms
(e) Le Luther:NPs de Lucien.
(f) Faire un dessin à:P la:Dfs Picasso: NPms.
(g) l'administration Bush:NPms

(h) Mme:NCfs Dupont:NPfs
(i) Mr:NCms Dupont:NPms

Les prénoms composés sont regroupés en un seul nom propre (*Jean_Jacques:NPms, Marie_Pierre:NPfs*).

Titres

Les titres et autres distinctions sont des propriétés. Ils désignent parfois des êtres par leurs qualités mais ne sont pas des désignateurs rigides. Nous les annotons noms communs bien qu'ils aient une capitale à l'initiale le plus souvent. Remarquons qu'ils ne s'accordent pas toujours avec le nom propre (*Madame le Président de la République, Le Préfet Martine Dubâton).* Les titres ont du mal à se féminiser en français par conservatisme (de la langue mais pas seulement), mais aussi parce que les formes féminines des titres sont historiquement attribuées aux femmes des hommes possédant le titre (Une préfette est une femme de préfet, une colonelle idem.) [14]

(a) Mr:NCms le:Dms Président:NCms Chirac:NPms
(b) Soeur:NCfs Emmanuelle:NPfs (une Soeur)
(c) le Père:NCms de_Foucault:NPms (un Père)
(d) le Pape:NCms (un Pape)
(e) le Général:NCms de_Gaulle:NPms (un général)

Noms de société, d'institution unique, de parti politique

Ces noms désignent des êtres uniques, ce sont des noms propres. Leur genre n'est pas toujours renseigné.

(a) la société:NCfs Coca_Cola:NPs
(b) l' Assemblée_Nationale:NPfs
(c) le Quai_d'_Orsay: NPms
(d) Le RPR:NPms

14. Nous finissons notre remarque en disant que le féminin d'un nom commun ne se rapporte qu'en partie au sexe. Pour légitimer sans sexisme une formule comme *Madame le Préfet* qui choque assez souvent, nous dirons que le féminin de *madame* est marqué par un morphème qui désigne le sexe, alors que le masculin de *Préfet* désigne le genre grammatical porté par le nom commun. Cette distinction explique en partie l'absence d'accord.

(e) Air_France:NPs
(f) Une:Dfs machine:NCfs Honda:NPs.
(g) L'équipe Yamaha:NPs.

Sigles

les sigles sont noms communs ou noms propres selon les mêmes critères.

(a) la SNCF:NPfs
(b) un HLM:NCms
(c) les PME : NCfp

Noms de planète, de produit, nom de fête : NC

Les noms d'astres sont des noms propres à l'exception du soleil et de la lune. En effet, contrairement aux autres astres, le déterminant est requis, et le nom de l'astre désigne non pas l'astre lui-même dans la langue (sauf dans une terminologie) mais un être parmi un référentiel.

La situation est la même pour les noms de fêtes.

Les nom de produits qui sont désignés par une marque ou un fabriquant sont des noms communs, contrairement au nom de la marque et du fabriquant lui-même.

(a) le soleil:NCms (Un soleil de mars)
(b) la lune:NCfs (une lune)
(c) Saturne:NPms (∗un saturne, Un Saturne comme jamais éclipsé par la lune)
(d) un Coca_Cola:NCms
(e) un:Dms IBM:NCms
(f) une:Dfs twingo:NCfs
(g) tous:Amp les:Dm No_1:NCmp
(h) Une:Dfs 205:NCfs.
(i) l'avion F-14:NCms.
(j) le M-X:NCms
(k) le ELF-5: NCms.
(l) Le Loto-sportif:NCms.
(m) A:P Noël:NCms dernier:Ams (un noël sans neige)

Points cardinaux, noms de ville, de pays, de mer, d'île, de région

Ces noms sont des noms propres car ils désignent une entité unique. Certains de ces toponymes sont pour plusieurs lieux (par exemple *Montréal, Washington, Chambon.* Cependant ils sont nécessairement désignateurs rigides dans une situation d'énonciation. Nous n'avons pas renseigné le genre des villes sans contexte clair. Celui-ci n'est pas toujours attribué et semble même arbitraire. (dit-on «Paris libéré» ou «Paris libérée»?). En revanche, les noms d'îles semblent être au féminin.

(a) vers:P l':Dms ouest:NPms
(b) Paris:NPs, la Charente:NPfs, les États_Unis:NPmp
(c) l':Dfs île_de_Ré:NPfs
(d) la:Dfs Méditerranée:NPfs
(e) la:Dfs Mer_morte:NPfs
(f) Rio_de_Janeiro:NPs
(g) L'Europe:NPfs
(h) Berlin-ouest:NPs
(i) Malte:NPfs

Noms de langue, de nationalité ou de devises

Ces noms sont des noms communs selon nos conventions. Remarquons qu'en position d'épithète, ils sont adjectifs.

(a) Il:CL3ms est:VP3s français:Ams
(b) la production:NCfs française:Afs
(c) Il:CL3ms parle:VP3s français:NCms
(d) le:Dms français:NCms est la langue universelle
(e) Les:Dmp français:NCmp votent:VP3p à:P gauche:NCfs
(f) Les:Dfp italiennes:NCfp sont:VP3p élégantes:Afp
(g) Un franc-français:NCms
(h) Un mark-allemand:NCms

	nominatif	réflexif	accusatif	datif	génitif	locatif
1re sing	je	me	me/moi à l'impératif		–	–
2e sing	tu	te	te/toi à l'impératif		–	–
3e masc sing	il/on	se	le	lui	en	y
3e fém sing	elle		la			
1re plur	nous				–	–
2e plur	vous				–	–
3e masc plur	ils	se	les	leur	–	–
3e fém plur	elles				–	–

FIG. 3.2 – *Pronoms conjoints ou clitiques*

	réfléchi	non réfléchi	
		masculin	féminin
1re sing	moi		
2e sing	toi		
3e sing	soi	lui	elle
1re plur	nous		
2e plur	vous		
3e plur	soi	eux	elles

FIG. 3.3 – *Pronoms personnels disjoints*

3.7.5 Pronoms personnels, pronoms clitiques

Les formes ambiguës entre pronom fort (noté «PRO») et pronom faible (noté «CL») sont : *moi, toi, lui, elle, elles, nous, vous, eux.* Voir les figures 3.2 et 3.6

Nous avons vu en 3.3.7 comment sont définis les clitiques. Nous pouvons exploiter des propriétés prosodiques et morphologiques propres à chaque catégorie pour trancher quand c'est possible (la forme clitique est conjointe au verbe et n'est jamais accentuée.)

Les pronoms sujet et complément de verbe sont pronoms clitiques. La forme forte ne peut en effet porter le cas nominatif, accusatif ou datif. Le pronom apposé est toujours une forme forte.

La forme forte est un groupe nominal, elle est substituable par un nom

commun.

(a) Nous:PRO1mp, nous:CLS1mp nous:CLR1mp en:CL3ms moquons:VP1p
(b) Regardez:VY2p -nous:CLS1mp
(c) Entre:P nous:PRO1mp
(d) Lui:PRO3ms, le:Dms diplômé:NCms de:P ...
(e) Ils:CL3mp n':ADV écoutent:VP3pl qu':ADV eux:PRO3mp et:CC nous:PRO1mp
(f) Elles:CL3fp ne:ADV regardent:VP3p que:ADV nous:PRO1mp
(g) Écoute:VY2s -moi:CL1ms

Les formes *moi_même, lui_même* sont des pronoms forts composés

Pronoms Clitiques Sujet, Objet ou Réflexif

Les pronoms faibles (ou pronoms clitiques) se distinguent en morphosyntaxe par une sous-catégorie indiquant le *cas*.

Nous indiquerons le cas des pronoms clitiques (datif, accusatif, nominatif ou réfléchi) en neutralisant les différences d'ordre syntaxique (comme celles qui dépendent de la sous-catégorisation des verbes) . Ainsi nous noterons, «objet» pour datif et accusatif, «sujet» pour les clitiques nominaux et «Refl.» pour les clitiques réfléchis. Ces informations seront assimilées à des sous-catégorisations grammaticales.

Tout comme pour les autres catégories, nous ne notons la sous-catégorie des clitiques que lorsque celle-ci est ambiguë hors contexte. Les pronoms clitiques *nous*, *vous*, *te*, *me* et *toi* sont les seuls pour lesquels nous devons noter la sous-catégorie. Nous noterons «CLS», «CLO», «CLR», les clitiques Sujet, Objet et Réflexifs.

Nous et *vous* sont ambigus entre CLS, CLO et CLR:

(a) nous:CLS1mp sommes partis ensemble.
(b) nous:CLS1mp nous:CLR1mp sommes rencontrés.
(c) Jean nous:CLO1mp donne des coups de pied.
(d) Jean nous:CLO1mp frappe.

Te, me et *toi* sont ambigus entre CLO et CLR

(a) Jean te:CLO1ms donne des coups de pied

(b) Tu te:CLR1ms regardes grandir.
(c) Donne -toi:CLO2ms un peu de courage !
(d) Regarde -toi:CLR2ms dans l'eau.

Nous notons le «t» euphonique et l'éventuel tiret (*-t-il, -t-on, -il*, etc) collés à la forme du clitique. Cette graphie n'a aucune valeur syntaxique. Elle appartient à l'histoire de la langue. La forme verbale latine marquait la troisième personne par le morphème /t/ en finale. Ce morphème est conservé en moyen français, et plus tard (au XV^{e} siècle) l'affaiblissement accentuel du pronom sujet préfixé est corrélé à l'amuïssement des finales verbales. L'affixe nominal est alors analysable comme un morphème de personne. Mais la forme verbale est toujours prononcée avec le morphème de troisième personne comme en moyen français lorsque l'affixe est post-posé. Cette prononciation est marquée en français contemporain écrit par une insertion d'un «t» entre le verbe et le pronom conjoint post-posé.

De même le «l'» est collé au clitique «on». La forme *l'on* bien qu'historiquement nom désignant *homme* ([Grévisse, 1993] §724) et naturellement assortie d'un article plus tardivement est aujourd'hui entièrement grammaticalisée. Nous considérons que nous avons affaire à un clitique composé ; que «l'» ne s'analyse aucunement en français contemporain. D'ailleurs il n'existe aucune paire minimale mettant en évidence une différence entre «l'on» et «on» autre qu'euphonique.

(a) pense:VPs3 -t-il:CL3ms
(b) doit:VP3s -on:CL3ms
(c) dis:VP1s -je:CL1ms
(d) l'on:CL3ms n':ADV en:CL3ms pense:VP3s pas:ADV moins:ADV

3.7.6 Les relatifs et interrogatifs

Les mots relatifs et interrogatifs

Un pronom relatif porte la personne, le genre, le nombre de son antécédent. Un pronom interrogatif invariable (*qui, que, quoi*) ou un pronom interrogatif porte toujours la 3^{e} personne au masculin singulier comme forme non marquée du point de vue de la morphologie.

Les pronoms introduisant des «relatives» sans antécédent sont toujours étiquetés pronoms interrogatifs selon les conventions que nous avons présentés

supra.

(a) je pense à:P qui:PROIms tu penses
(b) j'aime qui:PROIms tu:CL2ms sais:VP2s
(c) je vais où:ADV tu:CL2ms vas:VP2s

que a été étudié en 3.9.10

qui, lequel, quoi sont étudiés plus loin.

Où est soit Adverbe interrogatif (noté «ADV») et dénote le lieu, soit pronom relatif (noté «PROR») et dénote le lieu ou le temps. D'après nos conventions, le pronom relatif a une fonction dans la phrase qu'il introduit. Ce qui n'est possible qu'en faisant une représentation restituant une possible *trace* qui possède cette fonction (où$_i$ vas-tu $\emptyset_i$?). Nous ne supposons jamais de telles transformations.

(a) l'époque:NCfs où:PROR3fs les:Dfp bêtes:NCfp parlaient:VI3p
(b) le lit:NCms où:PROR3ms dort:VP3s Jean:NPms
(c) où:ADV vas:VP2s tu:CL2ms?
(d) je sais:VP1s où:ADV tu:CL2ms vas:VP2s

quand est Adverbe interrogatif (noté «ADV») en interrogative directe ou indirecte, sinon conjonction de subordination («CS»).

(a) quand:CS il pleut, il mouille:VP3s
(b) je sais:VP1s quand:ADV il:CL3ms arrive:VP3s
(c) elle part:VP3s quand:CS il:CL3ms arrive:VP3s

quel est déterminant interrogatif (noté «D») ou déterminant exclamatif (noté «DE») devant un nom ou un adjectif interrogatif attribut.

(a) je sais quelle:Dfs tasse:NCfs il:CL3ms veut:VP3s
(b) quelle:DEfs audace:NCfs!
(c) Quel:Ams est:VP3s cet:Dms homme:NCms?
(d) Quel:Ams qu':PROR3ms il:CL3ms soit:VS3s
(e) quel:Ams que:PROR3ms soit:VS3s son:Dms talent:NCms

Les formes en «**n'importe**» sont analysées comme adverbe, pronom ou déterminant composés, ce sont des formes indéfinies ni relatives ni interrogatives.

n'importe_qui:PROms

n'importe_quoi:PROms
n'importe_quel:Dms
n'importe_lequel:PROms
n'importe_où:ADV
n'importe_quand:ADV
n'importe_comment:ADV

Pronoms relatifs et interrogatifs

Les pronoms relatifs introduisent des relatives, les pronoms interrogatifs introduisent une interrogative indirecte, ou ont des places variées dans une interrogative directe.

Qui et **Lequel** sont pronoms relatifs s'ils prennent la fonction sujet (rarement pour *lequel*) ou s'ils apparaissent après une préposition régissant une relative. Ils sont pronoms interrogatifs comme sujet ou complément dans une interrogative directe ou indirecte. Ils peuvent également apparaître seuls dans une interrogative réduite.

(a) moi:PRO1ms qui:PROR1ms ai:VP1s fait:VKms ..
(b) Qui:PROIms a:VP3s fait:VKms cela:PRO3ms?
(c) Qui:PROIms as:VP2s -tu:CL2ms vu:VKms?
(d) je sais:VP1s à:P qui:PROIms tu:CL2s penses:VP2s
(e) l'homme avec:P lequel:PROR3ms tu:CL2ms parles:VP2s
(f) Lequel :PROIms veux:VP2s tu:CL2ms?
(g) Lesquelles:PROIfp viendront:VF3p?

Quoi est rare comme pronom relatif et plus fréquent comme pronom interrogatif. Il ne peut être pronom relatif qu'après une préposition.

(a) À:P quoi:PROIms penses-tu?
(b) il ne sait pas quoi:PROIms faire:VW
(c) je cherche un outil:NCms avec:P quoi:PROR3ms ouvrir:VW cette boîte
(d) quoi:PROIms que:PROR3ms tu:CL2s fasses:VS2s

3.7.7 Les mots démonstratifs

Les formes pronominales *ceci, cela, celui-ci, celui-là, celle-ci*, etc. sont toujours des pronoms forts. *Ce* et *c'* ont été étudiés en 3.9.1.

Les particules *-ci* et *-là* collées à un nom sont adverbes. Faute d'une classe spécifique des affixes autre que celle des préfixes dont il a été question en 3.6.6, nous devions classer ces particules en notant leur rôle de modificateur de nom, du point de vue de l'énonciation. Cette particule ne modifie pas le sens du mot, mais lui donne une valeur anaphorique, parfois cataphorique et déictique. Nous nous éloignons de la tradition grammaticale qui veut que l'on appelle *adjectifs* les déictiques (adjectif possessifs, adjectifs démonstratifs).

(a) cette:Dfs fois:NCfs -ci:ADV
(b) Montre cet objet -ci:ADV à notre invité ; celui que tu as dans la poche.

Les mots «voici» et «voilà» ont été vus en 3.6.5

singulier		pluriel	
masculin	féminin	masculin	féminin
celui	celle	ceux	celles
ce			
celui-là	celle-là	ceux-là	celles-là
celui-ci	celle-ci	ceux-ci	celles-ci

FIG. 3.4 – *Pronoms démonstratifs*

3.7.8 Les mots négatifs

Les clitiques *n'* et *ne* sont analysés comme adverbes. Nous avons réservé la classe des clitiques aux seuls pronoms. Nous n'avons pas distingué le *ne* explétif a ce niveau d'annotation.

Le mot *non* est adverbe ou interjection (dans les mots-phrases).

Les auxiliaires négatifs comme les nomme Grévisse ([Grévisse, 1993] § 976) sont soit des adverbes (*pas, plus, guère, jamais*) soit des pronoms (*rien, personne, aucun, nul*) soit des déterminants (*nul, aucun*)

(a) je ne:ADV vois:VP1s personne:PROms
(b) Personne:PROms n':ADV est:VP3s venu:VKms
(c) Une:Dfs personne:NCfs viendra
(d) jamais:ADV Paul:NPms ne:ADV viendra:VF3s
(e) Paul:NPms n':ADV a:VP3s pas:ADV de:Dfs chance:NCfs

(f) faire:VW un:Dms pas:NCms
(g) Rien:PROms ne:ADV l':CL3ms arrêtera:VF3s
(h) on:CL3ms ne:ADV voit:VP3s rien:PROms de:P bien:ADV
(i) je n'ADV en:CL3ms vois:VP1s aucun:PROms
(j) je ne:ADV vois:VP1s aucun:Dms chien:NCms

Personne en emploi négatif est pronom, sinon c'est un nom commun (*une personne*).

rien est pronom sauf dans son très rare emploi substantivé et dans l'emploi de *rien* adverbe archaïque en français standard.

(a) il:CL3ms n':ADV a:VP3s rien:PROms vu:VKms
(b) sans:P rien:PROms dire:VW
(c) compter:VW pour:P rien:PROms
(d) les:Dmp petits:Amp riens:NCmp de:P la:Dfs vie:NCfs
(e) c'est rien bien ! (= *un peu* Emploi archaïque)

Nul peut être pronom, déterminant ou adjectif. Il est pronom ou déterminant comme auxiliaire de négation contrairement à l'adjectif. Notons que la négation dans ce cas est relative, elle ne porte que sur le groupe nominal par exemple.

(a) une partie:NCfs nulle:Afs
(b) Nul:PROms n'ADV est:VP3s censé:Ams ignorer:VW la:Dfs loi:NCfs
(c) sans:P nul:Dms doute:NCms

Aucun est déterminant ou pronom

(a) aucun:PROms ne:ADV viendra:VF3s
(b) sans:P aucun:Dms doute:NCms

3.7.9 Les mots indéfinis

Les mots *tout(e)(s)* ont été vu supra en 3.9.14

Nous avons étudié les adjectifs indéfinis *seul, divers, différents* en 3.7.2

Autre, autres, certain, certaine, certains, certaines sont ambigus entre pronom (employé comme tête de groupe nominal), adjectif indéfini

(épithète gauche) et rarement qualificatif (attribut ou épithète droit). Nous notons «AI» l'adjectif indéfini. *Certaine* est également déterminant avant un nom ou un adjectif et non précédé d'un autre déterminant.

(a) un autre:AIms homme:NCms
(b) ils sont autres:Amp
(c) d':Dmp autres:PROmp viendront:VFmp
(d) j':CL1ms en:CL3ms vois:VP1s un:Dms autre:PROms
(e) Les:Dmp uns:PROmp les:Dmp autres:PROmp
(f) Certains:PROmp pensent:VP3p que:CS ...
(g) certaines:Dfp femmes:NCfp
(h) une:Dfs certaine:AIfs femme:NCfs
(i) je:CL1fs suis:VP1s certaine:Afs que:CS ...

Quelque, quelques est ambigu entre déterminant et adjectif. Il est également adverbe invariable dans les tours concessifs.

(a) quelques:Dmp restes:NCmp
(b) ces:Dfp quelques:Afp fleurs:NCfp
(c) quelque:Dms talent:NCms qu':PROR3ms on:CL3ms lui:CL3ms trouve:VP3s
(d) quelques:Dmp uns:PROmp
(e) quelque:ADV riche:Ams qu':PROR3ms il:CL3ms soit:VP3s
(f) quelque:ADV trois:Dmp millions:NCmp de:P francs:NCmp
(g) si:CS vous:CLS2mp avez:VP2p quelque:Dms souci:NCms que:PROR3s ce:CL3ms soit:VS3s

Quelqu'un, **quelque chose** toujours pronoms composés. Ils porteront invariablement les traits morphologiques masculin singulier.

3.7.10 Les mots possessifs

Nous récapitulons les déterminants et pronoms possessifs dans les tableaux des figures 3.5 et 3.6. Les formes *mon, ton, son, mes, tes, ses, notre, votre, nos, vos* sont toujours déterminants. Les formes *mien(ne)(s), tien(ne)(s), sien(ne)(s), nôtre(s), vôtre(s)* sont toujours pronoms. Les formes *notre(s)* et *votre(s)* peuvent apparaître pour des pronoms dans les textes de typographie peu soignée. De même que pour les formes démonstratives, nous

Possesseur	Nom déterminé		
	singulier		pluriel
	masculin	féminin	
1re sing	mon	ma/mon	mes
2e sing	ton	ta	tes
3e sing	son	sa	ses
1re plur	notre		nos
2e plur	votre		vos
3e plur	leur		leurs

FIG. 3.5 – *Déterminants possessifs*

Possesseur	Pronom			
	singulier		pluriel	
	masculin	féminin	masculin	féminin
1re sing	mien	mienne	miens	miennes
2e sing	tien	tienne	tiens	tiennes
3e sing	sien	sienne	siens	siennes
1re plur	nôtre		nôtres	
2e plur	vôtre		vôtres	
3e plur	leur		leurs	

FIG. 3.6 – *Pronoms possessifs*

ne nommerons pas, comme il est d'usage, les possessifs déictiques *adjectif*. Les raisons de cette dénomination sont historiques. En ancien français, le possessif était porté par un déterminant atone ou un adjectif tonique. Dans une langue recherchée et déjà dénoncée par les remarqueurs du XVIIe siècle (Vaugelas) comme passée, on emploie encore l'adjectif possessif : *Un mien cousin* ([Grévisse, 1993]. [15]

Mien(s), tien(s), sien(s) peuvent apparaître dans des emplois archaïques comme nous l'avons déjà vu pour des adjectifs. Nous écarterons cette possibilité qui n'appartient qu'à un langage particulièrement marqué.

Les formes *leur* et *leurs* sont ambiguës entre déterminant, clitique et

15. Je ne m'explique pas qu'une tournure en état d'obsolescence depuis plusieurs siècles et jamais employée ne puisse pas être entièrement rejetée comme agrammaticale aujourd'hui.

pronom possessif. Nous les avons déjà étudiés en 3.9.6.

3.7.11 Les quantifieurs *beaucoup, trop, peu, assez, bien, tant, tellement, moins, énormément, infiniment*

Ces formes ont en commun de s'employer comme adverbes (*dormir beaucoup*), comme prédéterminants (*beaucoup de gens*) et comme pronoms (*j'en veux beaucoup*). Ils appartiennent à un paradigme d'adverbes d'intensité.

Ils sont adverbes en tête de groupe nominal indéfini (dans les tours quantitatifs), ou comme adverbe de fréquence ou d'intensité. En effet, il est difficile d'admettre qu'ils puissent apparaître dans un déterminant complexe de type *beaucoup de, peu de*, etc. comme dans *le peu de* car il est possible d'adjoindre un déterminant possessif ou démonstratif (*beaucoup de ces enfants* ∗*le peu de ces enfants*). D'ailleurs l'absence d'article en surface n'est due qu'à la règle de cacophonie qui empêche la répétition de *de* (∗*beaucoup de d'enfants*). Nous avons vu que nous refusons des suites de déterminants et que les prédéterminants sont classés parmi les adjectifs ou les adverbes. Par ailleurs, ces formes sont adverbes car elles ne modifient pas le nom mais bien le déterminant.

(a) beaucoup:ADV de:P gens:NCmp
(b) beaucoup:ADV de:P vin:NCms
(c) beaucoup:ADV d':P autres:PROmp
(d) il aime beaucoup:ADV la:Dfs musique:Nfs
(e) bien:ADV de_les:Dmp travaux:NCmp
(f) Il mange moins:ADV que:CS toi:PRO2ms
(g) moins:ADV de:P farine:NCfs
(h) ce que j'aime le:Dms moins:ADV

Ils sont pronoms en tête de groupe nominal défini (tours partitifs), comme attributs et employés seuls comme sujet ou complément (avec le génitif en). Dans ces positions, ils se substituent (avec *de*) à un groupe nominal.

(a) Des fleurs, j':CL1ms en:CL3fp veux:VP1s beaucoup:PROfp
(b) beaucoup:PROmp de:P ces:Dmp gens:NCmp
(c) beaucoup:PROmp considèrent que...

(d) beaucoup:PROmp d':P entre:P eux:PRO3mp
(e) peu:PROmp demandent:VP3p leurs:Dmp droits:NCmp
(f) Ils sont beaucoup:PROmp

Peu

Nous avons groupé *le peu de* pour en faire un déterminant complexe. *Un peu* est un adverbe que l'on peut faire commuter avec *beaucoup*.

(a) le_peu_de:Dfs vin qu'il boit ne peut lui faire du mal.
(b) Il a bu un_peu:ADV de:P biere:NCfs.
(c) Il a bu peu:ADV de:P vin.
(d) Il est un_peu:ADV malade.
(e) très:ADV peu:ADV d:P affaires:NCfp
(f) très:ADV peu:PROmp sont diffusés
(g) il viendra sous_peu:ADV
(h) L'eau coule peu_à_peu:ADV
(i) Un_peu:ADV plus:ADV vite:ADV

tant

Tant est composant de la locution conjonctive *tant que*. Il est adverbe dans le sens de *intensément, beaucoup* et comme antécédent d'une corrélative. Il est conjonction de subordination quand il introduit une concessive (tour régional) ou comme introducteur de circonstancielle.

(a) il:CL3ms travaille:VP3s tant:ADV
(b) tant:ADV de:P fleurs:NCfp
(c) tant_qu':CS on:CL3ms le:CL3ms paie:VP3s
(d) tant:CS il:CL3ms paraît:VP3s en:CL3ms avoir:VW assez:ADV
(e) il:CL1ms n':ADV est:VP3s pas:ADV tant:CS riche:Ams qu':CS il:CL3ms ne:ADV prétend:VP3s

3.7.12 Les signes de ponctuation

L'annotation d'un corpus est, dans une certaine mesure, une expérience d'analyse robuste de la langue. Pour cette raison, elle doit s'accompagner de descriptions sur des catégories moins étudiées dans une approche générale. Les grammaires parlent généralement peu des ponctuations sinon dans leur

rapport à la prosodie et au discours. Or un corpus de français écrit contemporain contient un grand nombre de ponctuations qui doivent être décrites en morpho-syntaxe.

Les signes de ponctuation ont été distingués en faible et forte. Ils sont étiquetés «PONCTS» ou «PONCTW» (pour *strong* et *weak ponctuation*) .

Les ponctuations fortes (ou PONCTS) sont :

- Le point
- les points de suspension
- le point d'interrogation et d'exclamation (sauf dans un discours rapporté à l'intérieur d'une phrase)

Les autres ponctuations sont des ponctuations faibles

- La virgule, même dans le sens de la conjonction de coordination
- le point-virgule
- le deux-points
- le tiret d'incise et de dialogue
- les guillemets
- parenthèses et crochets.

Nous ne considérons pas que les autre marques typographiques sont des ponctuations (alinéas, renvois aux notes de bas de page, folios, lettrines, lettres italiques et capitales, etc.)

Les numéros de notes et autres signes ont été étiquetés nom commun tout comme les abréviations. Les distinctions typographiques n'ont pas été étiquetées ; nous ne considérons pas qu'elles puissent être analysées au niveau morpho-syntaxique comme les mots avec facilité.

(a) C'est-à-dire:P -:PONCTW comme:P le:CL3ms pense:VP3s Mr:NCms Bliote:NPms -:PONCTW les affaires ont changées .:PONCTS
(b) L':CL3ms a:VP3s -t-il:CL3ms rencontré:VKms ?:PONCTS
(c) Ici:ADV l'on:CL3ms dit:VP3s «:PONCTW requinqué:VKms »:PONCTW .:PONCTS
(d) Il:CL3ms offrira:VF3s 120:Dmp $:NCmp +:P 21:Dmp %:NCmp de:P taxes:NCfp .:PONCS
(e) (:PONCTW 9,75:Dmp %:NCmp ,:PONCTW soit:CC +:P 1,8:Dms point:NCms) :PONCTW

3.8 Les expressions numériques

Comme les ponctuations, les expressions numériques reçoivent ici un intérêt qui leur est rarement accordé dans les grammaires. Cette étude est spécifique au travail sur corpus et à l'analyse robuste des textes écrits.

3.8.1 Les nombres

Les nombres sont écrits en lettres (vingt-sept) ou en chiffres (27). Ils ont été regroupés comme des mots composés (100_000, quatre_vingt-sept).

Million, Milliard, Billion (parfois comme anglicisme mais également pour million de million) sont des noms. On dit en effet *un million de dollars* et non $*$*un million dollars* comme si *un million* était un déterminant à lui seul.

Notons que *vingt, cent ou mille* sont parfois des noms appartenant à une terminologie. (*Un cent de boites de camembert*).

Les nombres, qu'ils soient écrits en chiffres ou en lettres sont ambigus entre déterminant, adjectif, pronom et nom.

En tête de groupe nominal, le nombre est déterminant. Dans cette position, il peut commuter avec un article ou avec un déterminant possessif ou démonstratif. En emploi épithète, prédéterminant ou attribut détaché, il est adjectif.

Employé seul (sujet ou objet ou complément de préposition), le nombre est pronom.

Dans tous ces cas, nous noterons la sous-catégorie *cardinal*.

(a) une hausse de 40:Dmp %:NCmp
(b) Trois:Dfp fois:Nfp
(c) Tu:CL2ms as:VY2s 3:Dfp gommes:NCfp et:CC moi:PROms 4:PROfp
(d) à:P zéro:Dfs heure:NCfs
(e) c':CL3ms est:VP3s un:Dms zéro:NCms
(f) Les:Dmp trois:Amp brigands:NCmp
(g) Il n'y avait que deux:Dfp solutions:NCfp, Les:Dfp deux:Afp ont échoué.
(h) Ils:CL3mp sont:VP3p tous:Amp deux:Amp venus:VKmp me:CL1ms voir:VW
(i) Ils:CLmp sont:VP3p quatre:Amp
(j) Il:CL3ms est:VP3s numéro:NCms deux:Ams

(k) Un:Dms produit:NCms de:P 1,6:Dms quintal:NCms l':Dms hectare:NCms
(l) Dans:P les:Dmp 1,4:Dms %:NCms
(m) Trois:PROmp viendront:VF3p.
(n) J':CL1ms en:CL3mp veux:VP1s trois:PROmp.
(o) Il:CL3ms mange:VP3s comme:P quatre:PROmp
(p) Tous:Amp trois:PROmp viendront
(q) 50:PROmp de:P les:Dmp 60:Amp producteurs:NCmp
(r) un:Dms litige:NCms sur:P dix:PROmp

Il est nom commun pour nommer le nombre lui-même (emploi mathématique ou pour désigner un numéro), l'emploi métonymique du nombre (pour désigner ce qui porte le numéro par exemple).

(a) Il:CL3ms y:CL3ms avait:VI3s un:Dms quatre:NCms écrit:VKms sur le mur
(b) L':Dms étage:NC 13:NC n':ADV existe:VP3 pas:ADV !
(c) Restez:VY2p sur:P le:Dms canal:NCms 16:NCms
(d) l'omelette de la 5 (Fauconnier)

3.8.2 Les dates, heures, adresses, numéros de téléphone, etc.

Avant de poursuivre la description de l'annotation des nombres, disons que les cas particuliers des dates, heures, mesures, numéros de téléphones, etc. appartiennent à un autre niveau d'analyse. Nous pensons que l'annotation morpho-syntaxique des composants de tels «syntagmes» ne peut pas être assimilée à l'annotation des autres mots.

La grammaire permettant d'analyser ces constituants est différente de celle qui permet d'organiser la structure de la phrase. Une simple expression régulière permet par exemple d'analyser une date, même contenant une coordination comme (*le jeudi 13 et le samedi 15 mai 2003*).

La structure interne d'un tel constituant ne met pas en jeu des relations de dépendance sémantique comme il est question dans un syntagme. Une date s'encode en français selon des suites qui peuvent se résumer par un automate du type proposé par le système Intex ([Silberztein, 1993]).

Un nombre désignant une année est nom commun. Le nombre désignant la date est adjectif. Cette convention est dictée par ce qui se passe pour le nombre 1. L'ordinal est utilisé pour la date.

(a) le:Dms 1er:Ams juillet:NCms 1924:NCfs
(b) le:Dms deux:Ams est:VP3s un:Dms dimanche:NCms.
(c) en:P 1924:NCfs
(d) tous:Amp les:Dmp trois:Amp juillet:NCms
(e) le:Dms deux:Ams de:P chaque:Dms mois:NCms
(f) en:P juillet:NCms 1924:NCfs
(g) l':Dfs année:NCfs 1956:NCfs
(h) l':Dms an:NCms 2000:NCms
(i) vers:P les:Dfp une:Dfs heure:NCfs
(j) les:Dfp années:NCfp 80:NCfs
(k) Samedi:NCms 3:Ams juillet:NCms
(l) courant:P décembre:NCms

Les heures

Les heures notées en toutes lettres sont décomposées. Les heures notées en chiffres sont regroupées (2h15, 3h15mn13s, 3h15mn13'20"). Les noms «demi-heure» et «quart d'heure» sont regroupés.

(a) Deux:Dfp heures:NCfp et:CC demie:NCfs plus:ADV tard:ADV il:CLS3ms revenait:VI3s
(b) Il est deux:Dfp heures:NCfp et:CC demie:NCfs
(c) Vers:P 2h30:NCfp
(d) Ça:PROms fait:VP3s une:Dfs demie-heure:NCfs que:CS je:CL1ms l':CL3ms attends:VP1s
(e) L':Dfs horloge:NCfs sonne:VP3s les:Dmp quarts:NCmp et:CC les:Dfp demies-heures:NCfp après:P 20:Dfp heures:NCfp.

Les adresses

Comme pour les dates, on se rapproche de la structure habituelle des groupes nominaux sauf que le numéro de la rue est un NC :

(a) le:Dms 10:NCms , rue:NCfs Santeuil:NPms

(b) à:P le:Dms 10:NCms de:P ta:Dfs rue:NCfs
(c) 18:NCms , rueNCfs de:P Paris:NPms 34000:NCms Montpellier:NPs
(d) rue:NCfs de:P Charenton:NPs
(e) Paris:NPms 5e:Ams (ellipse de arrondissement:NCms)
(f) dans:P le:Dms 5e:Ams

Les numéros de téléphone

Les numéros de téléphone sont des noms communs :

(a) j':CL1s ai:VP1s appelé:VKms le:Dms 01-44-23-56:NCms
(b) Tél.:NCms : (514) _234_1234:NCms
(c) Tél.:NCms : 42_12_34_56_78:NCms

Les scores

Ce sont normalement des noms communs, sans tenir compte du contexte syntaxique:

(a) 3-5:NCms est un score minable.
(b) Il a été battu 5-7:NCms, 3-0:NCms, 7-0:NCms.

Les notes et numéros de chapitres

Ils sont NC, sauf en épithète droite.

(a) voir:VW page:NCfs 5:Afs
(b) 1.2:NCms De:P la:Dfs linguistique:NCfs

Les monnaies

Nous les analysons comme des noms communs.

Livre-sterling, escudo-portugais, franc-suisse, dollars-canadien, etc. sont des noms composés, même en abrégé.

(a) 200:Dmp FF:NCmp

(b) trois:Dfp Livres-sterling:Dfp font:VP3p environ:ADV trente:Dmp francs_français:NCmp
(c) 30:Dmp K$:NCmp
(d) il:CL3ms est:VP3s marqué:VKms $:NCmp 30:NCmp
(e) Cet:Dms avion:NCms a:VP3s coûté:VKms environ:ADV 36:Dmp MFF:NCmp

Cas particuliers:

Un, une, uns, unes

Un(e)(s) est déterminant ou adjectif au singulier ou pronom au singulier et au pluriel.

Quand *un(e)(s)* est employé seul après un déterminant, il est pronom:

(a) Les:Dmp uns:PROmp ont:VP3p tort:NCms
(b) L':Dfs une:PROfs de:P mes:Dfp amies:NCfp
(c) une:Dfs belle:Afs fille:NCfs
(d) le:Dms tome:NCms 1:Ams
(e) Dieu:NCms est:VP3s un:Ams

Une peut également être nom commun: *la une, une une* (en presse).

Nombres en épithète droit

En épithète droit, les nombres sont adjectifs et s'accordent en genre et en nombre avec le nom.

(a) Louis:NPms XVI:Ams, Elizabeth:NPfs II:Afs
(b) le:Dms volume:NCms II:Ams, le:Dms tome:NCms 3:Ams

Les partitions

Quart, tiers, moitié, centième etc. sont toujours des noms communs. *Demi(e)* est nom commun en coordination, adjectif en préfixe :

(a) une demie:Afs heure:NCfs
(b) une :Dfs heure:NCfs et:CC demie:NCfs
(c) une heure:NCfs et:CC quart:NCms

3.8.3 Les mots étrangers

Les mots étrangers sont catégorisés autant qu'il est possible, c'est-à-dire lorsque ces termes sont employés dans la phrase avec la fonction qu'ils auraient en français. Les nombreux emprunts qui ne sont pas toujours lexicalisés en français et qui apparaissent en italique dans les textes de typographie soignée reçoivent une catégorie selon les mêmes critères que les autres mots, Il en est de même des emprunts lexicalisés comme les nombreuses expressions latines (*mutatis mutandi, a priori, statu quo*, etc.)

(a) les *stock-options*(Nom Commun) de la société X

(b) Le new age(Nom commun)

(c) Cette situation est *a priori*(Adverbe) acceptable

(d) Des pâtes *al dente*(Adjectif)

En revanche, tous les autres mots étrangers qui ne participent pas à la construction syntaxique de la phrase reçoivent l'étiquette *ET*. Cette distinction est faite pour les citations

(a) l'Azerbaidjan restait bien «*gospoda*(ET)»

(b) La Pravda a donc préféré «*lioudi*(ET)» qui veut dire «gens», «hommes»

(c) les mots *arbeit*(ET) *macht*(ET) *frei*(ET) à l'entrée des camps nazis

3.9 Les mots les plus difficiles

Je reprends ici en les commentant les critères qui ont été décrits dans le guide d'annotation ([Abeillé & Clément, 1997]). Ces critères sont présentés par ambiguïtés et par mots polycatégoriels. Ils ont permis de fournir à l'équipe qui était chargée de valider l'annotation morpho-syntaxique sur une durée assez longue de quoi rendre le travail cohérent. Les motivations des choix qui ont été pris sont explicitées à chaque paragraphe et viennent compléter les remarques qui ont été faites sur chaque classe grammaticale.

3.9.1 C' - CE - -CE

C' est pronom clitique (CL3ms) dans tous les contextes. En effet, le déterminant ne s'élide pas mais fait **cet** devant un son vocalique[16]). Le pronom **ce** ne s'élide jamais.

(a) c'est la vie
(b) c'était un brave
(c) cet animal, cet honneur, (∗ce animal, ∗c'animal, ∗c'honneur)
(d) ce à quoi il faudrait penser, (∗c'à quoi il faudrait penser)
(e) ce «oui» qu'il prononça avec tant de détermination (∗c'oui)

-ce (avec un tiret) est toujours un clitique postposé au verbe. On le trouve dans les expressions idiomatiques *est-ce que*, *qu'est-ce que*.

Ce est ambigu entre déterminant, pronom masculin singulier (PROms) et pronom clitique (CL3ms).

Il est déterminant devant un nom ou un adjectif et commute avec un autre article, sa forme devant voyelle est «cet» ; il est donc possible de commuter le nom ou l'adjectif pour tester.

(a) devant:P ce:Dms spectacle:NCms (devant cet aspect du spectacle)
(b) selon:P ce:Dms dernier:Ams (selon cet avant-dernier)

Il est clitique sujet devant un autre clitique (*ne* compris) ou conjoint au verbe être. Il est clitique objet devant un participe présent (*disant, faisant*) ou un infinitif :

(a) ce:CLS3ms me:CLO1ms semble:VP3s
(b) ce:CLS3ms disant:VG
(c) ce:CLO3ms faisant:VG
(d) ce:CLS3ms serait:VC3s bien:ADV
(e) est:VP3s -ce:CLS3ms fini:VKms?
(f) ce:CLS3ms ne:ADV serait:VC3s pas:ADV plus:ADV mal:ADV
(g) pour:PREP ce:CLO3ms faire:VW

16. Nous ne disons pas une voyelle pour rendre compte du «h aspiré» ou d'autres consonnes prononcées comme le /w/ de *ouistiti* ou *oui* ([Grévisse, 1993]§48, §597

Il est pronom devant une relative qui le modifie, un adverbe ou une conjonction de subordination :

(a) ce:PROms qui:PROR3ms serait:VC3s bien:ADV
(b) ce:PROms pourquoi:ADV je:CL1ms vous:CLO2mp ai:VP1s appelé:VKms
(c) il cherche:VP3s à:P ce:PROms que:CS tu:CL2ms viennes:VS2s
(d) ce:PROms que:PROR3ms l':Dms enfant:NCms fait:VW

Quand le clitique sujet *ce* est suivi d'un verbe au pluriel, on considère qu'il y a accord avec l'attribut comme dans les tours *Son vrai désespoir était ses mains aux doigts trop courts et trop larges* (J. Roy cité par [Grévisse, 1993] §897) ou *Mère Ubu, tout ceci sont des mensonges* (Jarry - idem). Dans ces cas, nous considérons que le sujet est bien au singulier.

(a) ce:CL3ms sont:VP3p les:Dmp marchés:NCmp extérieurs:Amp
(b) les:Dmp impôts:NCmp , ce:CL3ms sont:VP3p les:Dfp catégories_B:NCfp qui:PROR3fp s':CL3fp en:CL3ms chargent:VP3p

3.9.2 COMME

Le mot *comme* est largement ambigu puisqu'il peut être adverbe, adverbe exclamatif (que l'on distingue dans l'annotation manuelle comme nous l'avons vu), préposition, conjonction de subordination et de coordination.

Il est conjonction de subordination dans les interrogatives indirectes, les subordonnées causales et les comparatives non réduites :

(a) Je n'aime pas:ADV comme:CS il:CL3ms joue:VP3s
(b) Comme:CS il:CL3ms faisait:Vp3s beau:Ams, je suis sortie
(c) Il:CL3ms joue:VP3s comme:CS il:Cl3ms chante:VP3s

Dans ces cas, en effet, il est impossible de ne pas voir ce terme comme le complémenteur de la phrase subordonnée. Il est vrai que nous n'avons pas toujours identifié les complémenteurs comme des conjonctions. C'est le cas de la «préposition» *de* devant une infinitive. Mais dans ce cas précis, la préposition *de* introduit une phrase qui possède un complémenteur effacé selon l'analyse de Hélène Huot ([Huot, 1981]).

Introduisant une comparative réduite, **comme** a été annoté préposition. Dans ce cas, la réduction ne se laisse pas toujours reconnaître comme telle,

et la transformation vers une phrase non réduite procède d'un mécanisme propre à une théorie linguistique (Chomskyenne en l'occurrence). Or nous ne voulions pas adapter nos critères à une théorie mettant en avant une structure profonde de la phrase. C'est pourquoi ces phrases réduites ont été assimilées à des attributs nominaux et adjectivaux dans l'analyse et **comme** a été annoté préposition.

(a) Je:CL1ms le:CL3ms considère:VP3s comme:P un:Dms frère:NCms
(b) Il est:VP3s comme:P un:Dms père:NCms pour:P moi:PRO1ms
(c) Jean:NPms comme:P son:Dms père:NCms aime:VP3s la musique
(d) Il:CL3ms a:VP3s été:VKms engagé:VKms comme:P cuisinier:NCms
(e) Comme:P violoniste:NCms on:CL3ms ne:ADV trouve:VP3s pas mieux
(f) je le considère comme:P fou:Ams
(g) Elle aime son père comme:P toi:PRO2ms
(h) Elle chante comme:P un:Dms rossignol:NCms

Comme est préposition également quand il introduit une adverbiale

(a) Comme:P toujours:ADV il:CL3ms est:VP3s en:P retard:NCms
(b) Comme:P prévu:VKms , il:CL3ms ne:ADV viendra:VF3s qu':ADV après:P le:Dms spectacle:NCms
(c) Comme:P tous:Amp les:Dmp lundi:NCmp il:CL3ms viendra:VF3s

Dans le sens de *presque, quasi, approximativement*, **comme** est adverbe. Dans cette position, il modifie un adjectif comme les autres adverbes qui entrent dans le paradigme. Il est également adverbe quand il modifie un verbe avec ce sens d'approximation ou dans les tours (b) qui marquent l'implication du destinataire. Cet emploi — certes plutôt oral — s'apparente au datif éthique comme en (c), (d).

(a) Il était:VI3s comme:ADV mort:Ams
(b) Il est:VP3s comme:ADV parti:VKms sans:P rien:ADV dire:VW
(c) Vous allez comme:ADV me servir un pain et des brioches !
(d) Vous allez me ranger cette caisse !

comme est conjonction de coordination quand on peut lui substituer une autre conjonction de coordination. Dans ce cas, le groupe nominal coordonné est au pluriel.

(a) Jean:NPms comme:CC Pierre:NPms sont:VP3p malades:Amp
(b) Jean:NPms comme:P Pierre:NPms est:VP3s malade:Ams

3.9.3 DE - D'

De peut être préposition ou déterminant. Il peut former un déterminant complexe amalgamé ou non *de la, des, du* avec un article défini. Nous considérons que *du* et *des* sont les contractions de *de le* et *de les* même lorsqu'ils sont articles. Dans tous ces cas, nous analyserons ces formes exactement comme des mots composés.

Dans tous ces cas, *de* est ambigu hors contexte avec la préposition homonyme. Pour trancher, il faut regarder si la préposition n'est pas dans une position où elle serait nécessairement régie, ou simplement voir si elle n'est pas commutable avec une autre préposition.

Le déterminant *de* ou *d'* au singulier est indéfini ou partitif. Il faut noter cette différence de sous-catégorie car elle est ambiguë hors contexte. L'article indéfini s'emploie pour introduire quelque chose dont il n'a pas encore été question dans le discours. Il peut également avoir une valeur générale (ibid. §567). Le partitif sert à introduire une réalité non comptable, une masse, une substance diffuse.

Les formes *un, du, de la, des* ne sont pas ambiguës, il suffit de les substituer à *de* pour tester.

Nous notons DPms et DPfs les déterminants partitifs.

(a) Je ne vois pas de:Dms chien (je ne vois pas un chien - ∗du chien)
(b) je n'ai pas d':DPms argent:NCms (∗un argent, de l'argent)
(c) Je n'ai plus:ADV de:DPfs monnaie:NCfs (de la monnaie - ∗une monnaie)

Une difficulté vient pour les idiotismes. Une propriété décelable pour les locutions verbales et autres expressions figées est l'absence d'actualisation d'un élément du terme complexe. Le test ne peut alors s'appliquer dans ce cas, puisqu'il est impossible, par la nature même du figement de substituer un autre déterminant à *de*.

(a) Jean:NPms fait:VP3s du:Dms vélo:NCms (∗de ce vélo, ∗d'un vélo particulier)
(b) Jean:NPms joue:VP3s du:Dms piano:NCms (∗de ce piano, ∗d'un piano à queue moins souvent)

Notons que les déterminants complexes *de ces*, *de son* sont annotés comme des composés pour éviter de produire des suites de déterminants. Nous rejetons en effet la position de déterminant spécificateur d'un groupe nominal

déjà déterminé. Il est cependant clair que le rôle de partitif de *de* (marqueur partitif dans [Véronis, 1999]) se distingue du déterminant démonstratif ou possessif et que la détermination du groupe nominal procède des deux opérations distinctes.

Comme nous l'avons dit, nous notons *de le, de les* toutes les formes contractées *du, des* composées ou non d'une préposition. Dans le cas de l'article amalgamé, nous l'écrirons *de_le, de_les* comme nous le faisons pour les mots composés. le caractère «_» permet de distinguer l'espace qui se trouve entre deux mots de l'espace qui sépare les composants d'un mot composé.

(a) Il:CL3ms tombe:VP3s de:P le:Dms train:NCms (de ce train, préposition régie de tomber : de)
(b) Il:CL3ms boit:VP3s de_le:Dms (du) champagne:NCms (un excellent champagne)
(c) Jean:NPms vend:VP3s beaucoup:ADV de:DPms beurre:NCms (il vend du beurre)
(d) Personne:PRO3ms ne:ADV prendra:VF3s de:Dfs décision:NCfs (une décision précise)
(e) Je:CL1ms ne:ADV me:CL1ms souviens:VP1s pas:ADV de:P les:Dfp vacances:NCfp (de ces vacances)

De est déterminant indéfini ou partitif après une préposition (sauf dans le cas d'une préposition composée), après un verbe transitif direct, ou comme introducteur d'un groupe nominal sujet. C'est en effet, un ensemble de positions syntaxiques qui ne peuvent être occupées par des prépositions mais qui peuvent être occupées par le déterminant.

Il apparaît au pluriel avant un adjectif épithète antéposé, et au singulier dans des tours négatifs.

(a) Il:CL3ms a:VP2s réussi:VKms malgré:P de:Dfp sérieuses:Afp difficultés:NCfp
(b) Je:CL1ms vois:VP1s de:Dmp beaux:Amp enfants:NCmp
(c) Je:CL1ms ne:ADV veux:VP1s pas:ADV de:DPms beurre:NCms
(d) Vous:CLS2p m':CLO1ms en:CL3fp apprenez:VP2p de:Dfp belles:Afp
(e) Personne:PROms ne:ADV prendra:VP3s de:Dfs décision:NCfs avant:P lundi:NCms
(f) De:Dmp nouveaux:Amp problèmes:NCmp sont:VP3p à:P l':Dms horizon:NCms
(g) Il:CL3ms n':ADV y:CL3ms a:VP3s pas:ADV de:Dfs relève:NCfs

(h) de:Dmp trop:ADV gros:Amp cadeaux:NCmp
(i) ils:CL3mp n':ADV auront:VF3p plus:ADV d':Dms effet:NCms
(j) on:CL3ms en:CL3fp annonce:VP3s de:Dfp nouvelles:Afp

Comme composant des contractions *du* ou *des* et de *de la*, *de* fait partie d'un déterminant composé partitif avec *le* (du=de_le) ou *la* (de_la) , d'un déterminant indéfini avec *les* (des=de_les) .

(a) gagner:VW de_l':Dms argent:NCms
(b) malgré:P de_les:Dfp difficultés:NCfp énormes:Afp
(c) Je:CL1ms vois:VP1s de_les:Dmp enfants:NCmp
(d) Je:CL1ms veux:VP1s de_le:Dms beurre:NCms
(e) De_les:Dmp problèmes:NCmp sont:VP3p à:P l':Dms horizon:NCms
(f) bien:ADV de_les:Dfp erreurs:NCfp
(g) faire de_le:Dms vélo:NCms
(h) Je:CL1ms veux:VP1s de_la:Dfs farine:NCfs

De est préposition avant une infinitive. Dans ce cas, nous considérons que le verbe de la phrase matrice sous-catégorise un infinitif construit avec une préposition comme pour un groupe prépositionnel. Nous n'avons pas retenu la distinction qui est faite par [Huot, 1981] entre une préposition précédant un paradigme (groupe nominal + infinitive + phrase) du complémenteur qui précède une phrase finie ou non.

Ce complémenteur est mis en évidence par les exemples (repris de [Huot, 1981] où la forme *de* apparaît être un verbe qui ne sous-catégorise pas toujours un groupe prépositionnel.

(a) *Il promet de cela
(b) Il promet cela
(c) Il promet de venir
(d) Il rêve de cela
(e) *Il rêve cela
(f) Il rêve de venir

Cette distinction n'a pas été retenue car son identification est délicate et suppose l'analyse de l'effacement de la préposition devant le compémenteur *Il promet de de venir*. Ainsi on repérera la préposition systématiquement après un verbe non transitif direct.

Comme introducteur de complément de Nom ou d'Adjectif, comme introducteur d'un complément circonstanciel et après certaines prépositions complexes, *de* est également préposition.

(a) Je:CL1ms viens:VP1s de:P finir:VW
(b) Il:CL3ms essaie:VP3s de:P les:CL3mp encourager:VW
(c) J':CL1ms en:CL3fs ai:VP1s une:PROfs de:P cassée:VKfs
(d) Il:CL3ms travaille:VP3s de:P ses:Dfp mains:NCfp
(e) Je:CL1ms rêve:VP1s de:P beaux:Amp enfants:NCmp
(f) La:Dfs maison:NCfs de:P Paul:NPms
(g) Il:CL3ms est:VP3s fier:Ams de:P lui:PRO3ms
(h) Il vient de:P chez:P lui:PRO3ms
(i) une pomme:NCfs de:P trop:ADV
(j) j'ai un fils:NCms de :P malade:Ams
(k) un communiqué de:P les:Dmp plus:ADV étonnants:Amp
(l) je veux:VP1s plus:ADV de:P profits:NCmp
(m) jouer de:P le:Dms piano:NCms
(n) De:P le:Dms balcon:NCms on:CL3ms voit:VP3s la:Dfs lune:NCfs
(o) Je:CL1ms ne:ADV me:CL1ms souviens:VP1s pas:ADV de:P les:Dfp vacances:NCfp
(p) La:Dfs nomination:NCfs de:P le:Dms Président:NCms
(q) Il:CL3ms est:VP3s fier:Ams de:P le:Dms résultat:NCms

De est préposition composée dans les mêmes conditions:

(a) En_dépit_de:P sérieuses:Afp difficultés:Nfp
(b) Au_delà_de:P la:Dfs mer:NCfs
(c) fermer:VW le:Dms gaz:NCms avant_de:P partir:VW

Remarque : *d'entre* est une préposition composée, variante de *de*

(a) Certains:PRO3mp de:P nous:PRO1ms
(b) Certains:PRO3mp d'entre:P nous:PRO1ms

3.9.4 EN

En est ambigu entre la préposition et le pronom clitique.

Il est préposition quand il introduit un groupe prépositionnel circonstanciel, ou complément. Dans ce cas, la sous-catégorisation du nom ou du verbe, voire de l'adjectif peut déterminer son statut. Comme introducteur du gérondif, *en* est préposition. Le clitique *en* se repère facilement par substitution d'un groupe prépositionnel en *de* ou d'une relative en *dont*.

(a) en:P le:CL3ms voyant:VG
(b) une:Dfs bague:NCfs en:P or:NCms
(c) en:P en:CL3ms ajoutant:VG
(d) Paul:NPms en:CL3ms parlera:VF3s demain:ADV
(e) Il:CL3ms est:VP3s mort:Ams en:P travaillant:VG
(f) Il:CL3ms s':CL3ms est:VP3s transformé:VKms en:P arbre:NCms
(g) Paul vend des pommes:NCfp et:CC Max:NPms en:CL3fp achète:VP3s

3.9.5 LE - LA - LES - L'

Ces formes sont ambiguës entre déterminant et pronom clitique. Les clitiques sont — comme leur nom l'indique — conjoints au verbe. On les trouve donc juste avant le verbe ou un autre clitique (dont *ne*). A l'impératif, le pronom conjoint suit immédiatement le verbe.

Dans tous les autres cas, *le, la, les* sont des déterminants.

(a) J':CL1ms essaie:VP1s de:P les:CL3mp encourager:VW
(b) Je:CL1ms ne:ADV les:CL3mp lui:CL3ms donnerai:VP1s pas:ADV
(c) Pour:P ne:ADV pas:ADV les:CL3mp y:CL3ms pousser:VW
(d) En:P les:CL3mp voyant:VG
(e) Les:Dmp trois:Amp brigands:NCmp
(f) De_la:Dfs salade:Nfs

Les infinitifs substantivés sont analysés comme des noms communs. Ils sont donc précédés d'un déterminant qu'on se gardera de confondre avec le clitique.

(a) le:Dms boire:NCms et le:Dms manger:NCms
(b) le:Dms souvenir:NCms

3.9.6 LEUR - LEURS

Leur est ambigu entre déterminant, pronom clitique, et pronom possessif. Les clitiques et déterminants sont présentés comme des classes distributionnelles, il est donc possible d'appliquer des tests de substitution pour trancher. Le clitique *leur*, comme pour les formes *le, la, les*, occupera la place des pronoms conjoints en français.

(a) Leur:Dms enfant:NCms est:VP3s malade:Ams (son enfant)
(b) Il:CL3ms faut:VP3s tenir:VW compte:NCms de:P leur:Dfs performance:NCfs (ma performance)
(c) Je:CL1ms leur:CL3mp parlerai:VP1s (je lui parlerai)
(d) Parlons:VY1p -leur:CL3mp (parlons-lui)
(e) Avant_de:P leur:CL3mp en:CL3ms parler:VW (avant de lui en parler)
(f) Je:CL1ms préfère:VP1s le:Dms leur:PRO3ms (je préfère le mien)

Notons que le déterminant *leur*, en plus de marquer le nombre pluriel, marque le nombre singulier du possesseur et se distingue ainsi de *leurs*. Le genre n'est toutefois pas indiqué par la forme (comme les autres déterminants possessifs pluriels). Il convient donc d'ajouter cette marque d'accord en contexte.

De même, le pronom conjoint *leur* n'indique pas le genre de l'antécédent qu'il faut donc attribuer en contexte.

3.9.7 LUI

Ce mot est ambigu entre le pronom fort et le pronom clitique. Cette distinction porte naturellement sur les différentes propriétés qui écartent ces deux catégories sur plusieurs axes : tonique / atone, conjoint / disjoint. Nous pouvons également substituer le terme par son équivalent au pluriel qui sera non ambigu : *leur* pour le pronom conjoint, *eux* pour le pronom disjoint.

Notons que la forme est ambiguë avec le participe-passé du verbe *luire*. Cette forme est quasi-inexistante dans les textes mais ne doit pas être écartée dans un lexique qui se veut exhaustif.

(a) Je:CL1ms lui:CL3ms en:CL3ms parlerai:VP1s
(b) En:P lui:CL3ms parlant:VG
(c) Parle:VY2s -lui:CL3ms !

(d) Pour:P lui:CL3ms parler:VW
(e) Je:CL1ms travaille:VP3s pour:P lui:PRO3ms (je travaille pour *leur/eux)
(f) une:Dfs photo:NCfs de:P lui:PRO3ms (une photo *de leur/d'eux)
(g) lui:PRO3ms qui:PROR3ms parle:VP3s si:ADV bien:ADV (eux/*leur qui parlent si bien)

3.9.8 MÊME - MÊMES

Même est ambigu entre un adjectif et un adverbe. Comme adjectif, il peut être adjectif indéfini dans une position où il a la valeur de déterminant indéfini sans être autonome comme déterminant. Nous dirons qu'il occupe cette place entre le déterminant et le nom, ou lorsque le nom est effacé.

(a) le:Dms même:AIms homme:NCms
(b) ce:CL3ms sont:VP3p les:Dmp mêmes:AImp
(c) avec:P même:AIfs fortune:NCfs

Il est adverbe invariable dans un sens d'intensité proche de: (*y compris, encore*, etc.), et modifie un adjectif, un adverbe ou une phrase. Nous le trouvons également comme prédéterminant. Dans ce cas, il participe à la détermination du nom.

(a) même:ADV si:CS ...
(b) même:ADV mort:Ams
(c) même:ADV avec:P de_les:Dfp augmentations:NCfp
(d) même:ADV Jean:NPms a compris
(e) même:ADV les:Dmp chiens:NCmp sont plus propres
(f) même:ADV pas:ADV de:DPms blé:NCms

Il est adjectif en suffixe ou en épithète postposé. Les formes *lui-même, moi-même*, etc. sont des pronoms composés; on n'y distingue pas l'affixe *-même*

(a) cet:Dms homme:NCms -même:Ams qui:PROR3ms etc.
(b) c'est la prudence:NCfs même:Afs
(c) aujourd'hui:ADV même:Ams
(d) Les:Dmp romains:NCmp ne:ADV vainquirent:VJ3p les:Dmp Grecs:NCmp que:ADV par:P les:Dmp Grecs:NCmp mêmes:AImp

3.9.9 PLUS

Plus est ambigu entre adverbe négatif et adverbe positif, marque de l'addition, et participe passé pluriel des verbe *plaire* ou *pleuvoir*. Nous le noterons adverbe ou préposition.

Comme auxiliaire de la négation, il se combine avec le clitique *ne* (que nous avons étiqueté adverbe). On le trouve également avec le sens négatif sans *ne* comme adverbe substituable par *pas, nullement, aucunement, jamais* ou *guère*.

Comme adverbe de quantité, comme introducteur d'une corrélative il est adverbe.

(a) Il mange plus:ADV que:CS toi:PRO2ms
(b) Le:Dms plus:ADV grand:Ams de:P les:Dmp trois:PROmp
(c) le:Dms succès:NCms le:Dms plus:ADV grand:Ams
(d) plus:ADV gentiment:ADV
(e) Un:Dms vieux:Ams père:NCms, une:Dmf fille:NCfs plus:ADV très:ADV jeune (Sartre cité dans [Grévisse, 1993])

Il est préposition dans des contextes d'addition. De cette façon, il sera distingué de l'adverbe avec lequel il ne partage rien dans un tel contexte.

(a) deux:NCms plus:P deux:NCms égalent:VP3p quatre:NCms

Il entre souvent dans la composition d'une locution adverbiale : *d'autant_plus*, *de_plus_en_plus*, *qui_plus_est*, *plus_ou_moins*, etc.

plus de est soit une préposition composée, soit l'adverbe suivi de la préposition selon le contexte. *le plus* est un adverbe composé sauf en tour superlatif. *plus que* est adverbe composé seulement comme prémodifieur d'adjectif. *Plus de* est composée devant un déterminant quantifié (comme *près de*), sinon il est adverbe simple suivi d'une préposition.

(a) plus:ADV de:P farine:NCfs
(b) ce que j'aime le_plus:ADV
(c) une:Dfs situation:NCfs plus_que:ADV florissante:Afs
(d) avec:P plus_de:P trois-cents: personnes
(e) avec:P plus:ADV de:P facilité:NCms

3.9.10 QUE - QU'

Que est ambigu entre adverbe simple ou exclamatif (ADVE), pronom relatif ou interrogatif et conjonction de subordination.

Rappelons les propriétés communes de ces relatifs:

- Ils ont une fonction dans la phrase enchâssée.
- Ils portent les traits morphologiques genre, nombre et personne.

La conjonction *que* risque d'être confondue avec le pronom. Par exemple dans les complétives de noms ou les corrélatives introduites par *que* comme en (a) et (b).

(a) Il a plus traduit de textes que je n'aurais cru
(b) La certitude que Jean pouvait en traduire tant ne l'a pas effleuré

Les relatives n'ont pas toujours d'antécédent (par exemple: *Je choisirai qui je veux*), mais le relatif *que* en requiert un (qui peut être le pronom *ce*). La fonction du pronom *que* est objet, attribut ou complément adverbial. Sa forme élidée *qu'* peut également être sujet «réel» comme le souligne Grevisse [Grévisse, 1993] p. §689.

Nous considérons qu'il est pronom relatif dans les clivées. Nous marquons ainsi une relation de dépendance syntaxique entre le groupe nominal clivé et la phrase relative.

(a) la fille que:PROR3fs je vois
(b) l'argent qu':PROR3ms il a fallu pour restaurer la maison
(c) insensé que:PROR3ms je suis!
(d) Les dix grammes que:PROR3mp cette lettre pèse
(e) l'homme que:PROR3ms tu aimes
(f) les:Dfp idées:NCfp que:PROR3fp tu:CL2ms as:VP2s approuvées:VKfp
(g) je demande:VP1ms ce:PROms que:PROR3ms tu:CL2ms vois:VP2s
(h) j'aime:VP1ms ce:PROms que:PROR3ms je:CL1ms vois:VP1s
(i) Que:PROIms vois:VP2s -tu:CL2ms?
(j) je:CL1ms ne:ADV sais:VP1ms que:PROIms faire:VW
(k) C'est Jean:NPms que:PROR3ms je vois
(l) C'est à Jean que:PROR3ms je pense

(m) quoi:PROIms qu':PROR3ms en:CL3ms pensent:VP3p les:Dmp collègues:NCmp

Le pronom relatif *que* ne porte pas de marque flexionnelle d'accord. En revanche, comme pronom coréférent à son antécédent, nous lui apportons les traits de genre et de nombre. Le pronom interrogatif *que* occupe la fonction d'objet dans les interrogatives directes ou indirectes (subordonnées interrogatives in [Riegel *et al.*, 1994]). Il ne porte aucune marque d'accord puisqu'il est typiquement un «symbole incomplet» (ibid.) et se rapporte à un élément non réalisé.

La forme *que* apparaît dans les tours restrictifs *ne ... que*. Dans cette acception, *que* a la valeur d'un auxiliaire négatif et prend le sens de *seulement*, *uniquement*. Le tour complet se substitue à *ne ... rien/personne/nulle part, sinon* (Luc ne voit personne sinon toi). Il est évident que ce terme n'est pas seulement un connecteur grammatical comme la conjonction, nous l'avons donc étiqueté adverbe.

(a) Je:CL1ms ne:ADV vois:VP1s que:ADV toi:PRO2ms
(b) Ils:CL3mp n':ADV ont:VP3p plus:ADV que:ADV la:Dfs haine:NCfs

Dans les exclamatives, *que* prend également un sens propre ; nous l'avons donc étiqueté adverbe avec la sous-catégorie exclamatif.

(a) Que:ADVE de:Dmp bonbons:NCmp !

Que est conjonction de subordination dans une position de complémenteur. Nous le noterons ainsi quand la phrase enchâssée ne fait aucun doute, ce qui n'est pas toujours le cas lorsque le verbe est effacé dans les comparatives. Il se trouve donc après un prédicat (le plus souvent un verbe mais également un nom ou un adjectif) qui sous-catégorise une complétive, dans les comparatives ou les corrélatives (mêmes réduites) et dans les impératives.

(a) Je veux:VP1s que:CS tu:CL2ms viennes:VS2s
(b) Comme:CS il:CL3ms pleuvait:VI3s et:CC que:CS tu:CL2ms venais:VI2s ...
(c) Voilà:VP3s qu':CS il:CL3ms pleut:VP3s
(d) Il:CL3ms cherche:VP3s à:P ce:PROms que:CS vous:CLS2mp veniez:VS2p
(e) Il:CL3ms s':CL3ms étonne:VP3s de:P ce:PRO3ms que:CS vous:CLS2mp partiez:VS2p

(f) L':Dfs idée:NCfs que:CS tu:CL2ms approuves:VS2s notre:Dms projet:NCms me:CLO1ms réjouit:VP3s
(g) plus:ADV grande:Afs que:CS vous:PRO2mp
(h) plus:ADV grande:Afs que:CS vous:CLS2mp ne:ADV l':CL3ms imaginiez:VI2p
(i) Qu':CS il:CL3ms aille:VS3s se:CL3ms faire:VW voir:VW !
(j) Une capacité:NCfs plus:ADV que:CS restreinte:Afs
(k) un homme:NCms tel:Ams que:CS lui:PRO3ms
(l) la règle:NCfs est:VP3s que:CS les députés soient élus

3.9.11 S'

S' est la forme élidée du pronom clitique réfléchi *se* ou de la conjonction de subordination *si*. Il suffit de substituer la forme élidée par la forme pleine en contexte pour déterminer l'un et l'autre ; les formes non élidées n'étant pas ambiguës entre elles. La forme clitique prend des traits morphologiques en contexte qui ne sont pas marqués par la flexion.

(a) Ils:CL3mp veulent:VP3p s':CL3mp amuser:VW
(b) Je:CL1ms viens:VP1s s':CS il:CL3ms vient:VP3s

3.9.12 SI

Ce mot est ambigu entre le nom commun, l'adverbe et la conjonction de subordination.

Dans les tours comparatifs (négatifs ou interrogatifs), il est adverbe. Nous le considérons en effet comme adverbe quand il prend le sens d'intensité (*tellement, particulièrement, intensément, etc.*).

Dans les «mots-phrases» affirmatifs, nous considérons que nous devons le considérer comme un nom puisqu'il peut être déterminé (*Il m'a répondu par un «si» de la tête.*, *Il murmure timidement le oui que tout le monde attendait.* N'oublions pas qu'il peut être également le nom de la note de musique sans ambiguïté.

(a) Est-il vraiment si:ADV méchant que ça? ([Riegel *et al.*, 1994])
(b) Rien n'est si:ADV dangereux qu'un ignorant ami (La Fontaine cité dans ibid.)

(c) – Est-ce que tu le connais? – Si:ADV je le connais! (ibid.)

Si est toujours conjonction de subordination pour introduire les interrogatives et les «circonstancielles».

(a) si:ADV gentil:Ams
(b) si:ADV rarement:ADV
(c) Il:CL3ms est:VP3s toujours:ADV si:ADV en:P retard:NC:ms que:CS...
(d) Si:CS tu:CL2ms viens:VP2s, nous:CLS1p nous:CLR1p en:CL3ms irons:VP1p
(e) Je:CL1ms me:CLR1ms demande:VP1s si:CS tu:CL2ms viendras:VP2s
(f) Il m'a répondu que:CS si:NCms
(g) Il m'a répondu que:CS oui:NCms

3.9.13 TEL - TELS - TELLE - TELLES

Ces formes sont ambiguës entre déterminant, pronom et adjectif.

Tel est adjectif en prédéterminant, en épithète et en attribut (souvent avec inversion du sujet). Dans cette position, il modifie le nom lui-même comme un attribut. En fait, cette proforme dans les constructions comparatives et consécutives prend la fonction d'attribut du sujet. C'est au titre de cette fonction que nous le catégorisons adjectif.

(a) Luc n'a jamais été souple et ne le sera jamais – Il est tel qu'il a toujours été ([Riegel *et al.*, 1994] p. 234)
(b) Son retard est tel qu'il ne le rattrapera pas (ibid.)

(a) tel:Ams un:Dms lion:NCms
(b) un:Dms tel:Ams homme:NCms
(c) tel:Ams était:VI3s cet:Dms homme:NCms
(d) Il:CL3ms était:VI3s tel:Ams que:CS tu:CL2ms l':CL3ms avais:VI2s laissé:VKms

Tel est déterminant devant un nom sans déterminant. Il peut, dans cette position être substitué par un autre déterminant puisque la classe est une classe distributionnelle. Notons que ce déterminant est en désuétude, qu'il

requière un déterminant défini même lorsque celui-ci n'apporte aucune information de plus (**Tel travail ne peut être vendu !*, *Un/*le/*ce tel travail ne peut être vendu*).

(a) tel:Dms père:NCms tel:Dms fils:NCms

Tel est un pronom comme groupe nominal.

(a) tel:PROms qui:PROR3ms rit:VP3s vendredi:NCms ...
(b) avec:P tel:PROms ou:CC tel:PROms de:P vos:Dmp partenaires:NCmp

3.9.14 TOUT - TOUTE - TOUTES - TOUS

Ces formes sont ambiguës entre déterminant, pronom, adverbe et adjectif.

Tout au singulier masculin et féminin est déterminant avec le sens de l'emploi générique du déterminant indéfini. (ex. *un enfant se tient bien à table.*)

(a) Tout:Dms homme:NCms a:VP3s le:Dms droit:NCms de:P...
(b) Toute:Dfs nouvelle:Afs étudiante:NCfs se:CL3fs verra:VF3s remettre:VW...

En position de prédéterminant, bien que ses propriétés de quantifiant ou spécificateur soient proches de sa position de déterminant, nous l'étiquetons adjectif. Nous n'admettons en effet pas de «groupe déterminant» comme donnée pré requise à notre annotation. Ce groupe classerait un grand nombre de prédéterminants comme tels mais rendrait l'annotation dépendante d'une telle analyse. De plus, nous n'admettons pas de suite de déterminants dans notre analyse morpho-syntaxique.

Dans cette position, *tout*, *toute*, *toutes* et *tous* sont adjectifs et s'accordent avec le nom qu'ils modifient.

Notons par ailleurs que *Tout_le_monde* est une pronom composé.

(a) Tout_le_monde:PROms
(b) Toutes:Afp les:Dfp femmes:NCfp
(c) Tous:Amp les:Dmp deux:Amp
(d) Tout:Ams cela:PROms est:VP3s faux:Ams

Les proformes *tout*, *toute*, *toutes* et *tous* ont les fonctions d'un groupe nominal dans la phrase. Nous trouvons ces pronoms également comme quantifieur flottant d'un groupe nominal non réalisé. Dans ce dernier cas, le pronom s'apparente aux autres proformes qui peuvent être analysées comme des déterminants dans la mesure où le nom est restitué (comme les cardinaux).

(a) Tous:PROmp viendront:VF3p
(b) Elles:CL3fp viendront:VF3p toutes:PROfp
(c) Je veux tous:PROmp les:CL3mp voir:VW
(d) Je les:CL3mp verrai:VF1s tous:PROmp
(e) J'ai:VP1s tout:PROms lu:VKms
(f) Il:CL3ms pense:VP3s à:P tout:PROms
(g) Tout:PROms est:VP3s possible:Ams
(h) Tout:PROms ce:PRO3ms qui:PROR3ms évoque la mer
(i) Tous:Amp deux:Amp ont:VP3p raison:NCfs

Tout, *toute* et *toutes* sont adverbes devant un adjectif (ou un participe passé employé comme adjectif), un adverbe ou une préposition. Dans ces positions, le terme modifie le sens du procès et non de la quantification. Nous pouvons avoir une double lecture de *Les pommes sont toutes mouillées.* L'une indique que la totalité des pommes ou l'ensemble de l'entité massique est mouillée, l'autre signifie que le procès est réalisé avec intensité. Dans les deux cas, la modification par *toutes* ne porte pas sur les mêmes termes.

Il est tout à fait remarquable que l'adverbe *tout* s'accorde avec l'adjectif comme un attribut (cf. 3.3.4).

(a) Il est tout:ADV rouge:Ams
(b) J'ai vu des filles toutes:ADV rouges:Afp
(c) Une:Dfs toute:ADV nouvelle:Afs étudiante:NCfs
(d) Tout:ADV récemment:ADV
(e) Tout:ADV en:P conduisant:VG
(f) Il est tout:ADV bouleversé:VKms
(g) Tout:ADV juste:ADV.
(h) Tout:ADV à:P l':Dms honneur:NCms

Très rarement, *tout* est un nom commun précédé du déterminant défini.

(a) le:Dms tout:NCms est:VP3s de:P participer:VW
(b) le tout:NCms pour le tout:NCms

Chapitre 4

Les choix informatiques pour le corpus annoté

Nous présentons dans ce chapitre les choix qui ont été effectués à propos de l'encodage du corpus étiqueté de Paris 7 comme ressource informatique. Pour cela, nous verrons ce qu'implique l'emploi de tel ou tel format au niveau des caractères et au niveau des mots du point de vue de la théorie de l'information. Nous présentons également les outils développés qui ont permis sa construction.

Une spécification plus précise de la norme d'encodage SGML, XML sera présentée, nous verrons comment nous pourrons exploiter le corpus annoté grâce à cette norme.

La mise au point du corpus annoté s'est accompagnée du développement d'un étiqueteur automatique auquel nous avons contribué. Nous présentons cet étiqueteur relativement à l'existant dans la matière.

La construction du corpus étiqueté de Paris 7 a nécessité un grand nombre de projections d'informations venant de dictionnaires divers, pour ajouter le lemme ou les sous-catégories grammaticales par exemple. Nous présentons un outil permettant de faire ces opérations indépendamment du type de dictionnaire.

L'exploitation du corpus annoté ne peut se faire sans outils d'interrogation adaptés. C'est pourquoi nous présentons l'outil d'interrogation que nous avons développé.

4.1 Les formats de balisage de corpus

Le format d'encodage de corpus doit satisfaire un certain nombre de contraintes liées à son utilisation. Comme ressource informatique, un corpus doit être opérande de traitements automatiques reproductibles, nous verrons ce que cela implique formellement. Comme ressource linguistique, il doit permettre de vérifier ou mesurer des hypothèses sur la langue, voire d'induire des mécanismes grammaticaux. Le matériau d'une telle expérimentation doit prévenir les artefacts. Un corpus ne donnera les informations fiables seulement sur ce pour quoi il a été conçu.

Le format codifié du corpus annoté doit également faciliter les échanges et doit permettre les affichages paramétrables du corpus pour offrir des vues multiples des mêmes informations.

4.1.1 Langage de balisage

En tout état de cause, l'encodage du corpus étiqueté doit permettre de représenter de façon non ambiguë l'intégralité de son contenu. C'est-à-dire qu'un ou plusieurs documents contiendront de façon bi-univoque le corpus étiqueté dans un format informatique.

Laissons de côté pour l'instant ce qu'impose la communication de ces ressources à toute une communauté pour ne voir que les impératifs formels d'un encodage de corpus étiqueté.

Ceci peut sembler innocent, mais il suffit de traduire sous forme de grammaire formelle différents formats d'annotation inventés pour se rendre compte que la chose n'est pas toujours évidente.

Si le corpus doit représenter la graphie des mots et mots composés, leur lemme et étiquettes morpho-syntaxiques. Ces informations viennent linéairement en suivant le texte, les difficultés ne se posent guère que pour les mots composés et agglutinés. Une liste de lignes contenant la graphie suivie du lemme puis des étiquettes suffira alors. Le langage d'un tel code pourrait être décrit par la grammaire régulière suivante :

$$\text{Mot} \rightarrow \text{graphie lemme étiquette Mot}$$
$$\text{Mot} \rightarrow \epsilon$$

Un exemple de corpus encodé de «*c'est peut-être écrit dans un article publié au Monde*» pourrait être :

c'	ce	PRON
est	être	V
peut-être	peut-être	ADV
écrit	écrire	V
dans	dans	ADV
un	un	DET
article	article	NC
publié	publié	V
au	à	PREP
Monde	le_Monde	NP

Mais si le corpus étiqueté contient des unités discontinues comme *ne ... pas*, *ne ... que*, *afin ... de*, *fait ... partie* que l'on voudrait annoter comme un seul terme, un tel langage ne serait pas assez puissant. Une grammaire CFG serait requise pour annoter les termes discontinus englobant les insertions en toute généralité. On pourrait par exemple proposer la grammaire suivante :

$$\text{Mot} \rightarrow \text{graphie lemme étiquettes Mot}$$
$$\text{Mot} \rightarrow \text{Mot} \langle \text{Mot} \rangle \text{Mot}$$
$$\text{Mot} \rightarrow \epsilon$$

Un exemple de corpus encodé de «*Afin, comme le dit Michel, de lui parler*» pourrait être :

	afin	afin de	PREP	⟨
	,	,	PONCT	
	comme	comme	CONJ	
	le	le	PROCL	
	dit	dire	V	
	Michel	Michel	NP	
⟩	de	afin de	PREP	
	lui	lui	PROCL	
	parler	parler	V	

Maintenant, si l'on voulait annoter à la fois les constituances et dépendances d'un même texte, il est évident qu'une grammaire CFG ne serait pas assez puissante. Une grammaire permettant de générer un graphe et non plus simplement un arbre serait alors requise.

On voit que l'encodage du corpus doit être pensé du point de vue de la puissance des grammaires formelles générant le code. Ceci est d'autant plus vrai que nous voudrions connaître la faisabilité d'une analyse de ce code. Il est évident que si l'information à encoder se résume à notre premier exemple,

un langage régulier serait bienvenu car analysable grâce à un automate fini déterministe, c'est-à-dire dans un temps proportionnel à la longueur du corpus. La recherche d'une occurrence dans le corpus se fera alors de façon linéaire. En revanche, l'encodage du deuxième exemple est analysable dans un temps proportionnel à un polynôme de la longueur du corpus. La recherche d'une occurrence (insertions comprises) réclamera la mise à disposition de l'un des algorithmes connus pour l'analyse des langages hors contexte. Nous voyons ce qu'on risque à inventer un encodage correspondant à une grammaire de réécriture non contrainte ! Au mieux des heuristiques permettront de ramener une partie du code à un langage analysable en temps fini.

Notons que notre exemple est volontairement excessif, que l'encodage des termes discontinus peut se satisfaire d'un langage régulier. Il suffit pour cela de lister l'ensemble des discontinuités et de distinguer les composantes comme des mots simples comme ceci :

afin ⟨ de	afin de	PREP
,	,	PONCT
comme	comme	CONJ
le	le	PROCL
dit	dire	V
Michel	Michel	NP
afin ⟩ de	afin de	PREP
lui	lui	PROCL
parler	parler	V

L'encodage des mots composés dans les projets GRACE et Multitag ([Lecomte, 1997]) permet d'encoder des mots composés de plus de deux termes. Chaque composant d'un mot composé de k termes est annoté d'un numéro $1/k$,$2/k$, $\ldots k/k$. Cette dernière notation permet de traiter les mots composés grâce à des expressions régulières et permet de rendre compte de toutes les ambiguïtés comme nous l'illustrons figure (4.3). En revanche, cette notation seule ne suffit pas pour annoter les mots composés discontinus comme le montre l'exemple suivant :

afin	afin de	PREP	1/2
,	,	PONCT	
comme	comme	CONJ	
le	le	PROCL	
dit	dire	V	
Jean	Jean	NP	1/2
Michel	Michel	NP	2/2
de	afin de	PREP	2/2
lui	lui	PROCL	
parler	parler	V	

Ces ajustements se font généralement dans les encodages de corpus intuitivement sans qu'il ne soit jamais question de langages formels. Nous proposons de rendre systématique la description de la grammaire qui engendre le code pour connaître les algorithmes mis en œuvre pour la recherche d'informations dans le corpus encodé. Les formats d'annotation SGML et XML offrent le terrain d'une telle formalisation en plus de la normalisation qu'ils impliquent.

4.1.2 L'encodage des caractères

Le français écrit utilise un jeu de caractères distincts propres à représenter des graphèmes, et un certain nombre de caractères propres au seul usage typographique. Les uns sont nécessaires à distinguer les unités graphiques du français écrit, les autres appartiennent à l'art de la typographie.

Depuis près d'un quart de siècle, les typographes utilisent les moyens informatiques pour mettre en page les textes. Ces moyens ont été créé le plus souvent par une communauté anglophone qui n'a pas toujours fait la place à la diversité des écritures des langues du monde. L'usage est aujourd'hui d'éditer des textes dont la typographie est appauvrie (c'est par exemple le cas des courriers électroniques sans accents).

Mais la raison principale de cet appauvrissement était l'économie nécessaire à l'encodage des textes. Toute donnée informatique est composée de bits ou chiffres binaires. Ces chiffres binaires sont combinés en *mots*, groupes de 8, 16, 32, etc.[1] unités selon les machines. L'encodage d'une information électronique élémentaire quelconque est un élément choisi parmi 2^8, 2^{16}, 2^{32},

1. Remarquons que les puissances de deux ne sont nécessaires que pour certaines architectures (maintenant toutes) multiplexant l'information en utilisant la moindre ressource. C'est-à-dire décomposant successivement le mot 2^k bit à bit sur k bits en k instants. Il a existé des ordinateurs dont le mot était de 4 bits ou 12 bits.

etc. éléments. L'encodage d'un caractère est donc le choix d'un élément dans un ensemble de 256, 65 536, $4{,}29.10^9$, etc. éléments.

À une époque où les donnée numériques étaient extrêmement coûteuses[2], le choix s'est porté sur le minimum possible permettant d'encoder des textes et documents informatiques en anglais. Ce minimum contient :

- les lettres en *minuscules* et capitales
- les chiffres arabes : `0 1 2 3 4 5 6 7 8 9`
- les signes de ponctuation : `‘ ’ " ( ) [ ] { } , ; : . ? !`
- les signes arithmétiques et logiques : `+ - * / < > ^`
- les espaces : blanc, tabulation, nouvelle ligne, «retour chariot»
- Les signes informatiques : `& # ~ _ @`
- Les codes de communication vers des matériels d'affichage, d'impression ou de transmission (nouvelle page, inversion vidéo, fin de transmission, etc.)

Soit environ 110 éléments distincts. Or il faut $\log_2 110 \simeq 7$ bits pour encoder un élément parmi 110.

En 1963, l'ASCII (American Standard Code for Information Interchange) a été créé en donnant une correspondance entre les 128 (2^7) nombres binaires et une liste de caractères alpha-numériques et de codes. Ce standard a été normalisé sous le nom d'ISO-646 et adopté par toute la communauté de telle sorte qu'il subsiste encore de nos jours. Les ordinateurs des années 80 — années de réelle expansion de l'informatique — manipulaient des *mots* de 8 bits[3]. Le code ASCII laissait donc un bit (dit *bit de parité*) qui a été utilisé de diverses façons : ajout d'une information marquant chacun de ces caractères (octet «signé») comme souligné ou en noir au blanc (vidéo inversée), dédoublement du jeu de caractères pour tenir compte des graphies des langues du monde. Plusieurs standards ont vu le jour à la suite de ces travaux dont les plus connus sont, ASCII-DOS, MAC et ISO-8859-n, bien malheureusement incompatibles entre eux.

Avec un octet, on peut encoder $2^8 = 256$ caractères différents, c'est suffisant pour les caractères employés en typographie française par exemple, mais insuffisant pour encoder les caractères de plusieurs langues. La norme ISO-8859-n propose un encodage pour des langues rassemblées en régions

2. Quiconque a démêlé des bandes perforées connaît ce coût !

3. IBM et Apple sortaient respectivement des ordinateurs personnels configurés autour des premiers processeurs 8 bits de Intel et Motorola les 8088 et 6800

géographiques. ISO-8859-1 permet de représenter les caractères des langues de l'Ouest de l'Europe : allemand, anglais, danois, espagnol, féroïen, finnois, français, islandais, italien, néerlandais, norvégien, portugais, suédois ([Habert *et al.*, 1998]). Cependant des oublis manifestes rendent cet encodage rédhibitoire pour des textes de typographie soignée, la ligature «œ» en français par exemple. De plus, un texte français qui contiendrait des citations en catalan, croate, groenlandais, hongrois, lapon, lituanien, maltais, polonais, roumain, slovaque, tchèque ou turc ne pourrait être encodé en ISO-8859-1, mais en combinant des encodages ISO-8859-n, ce qui est naturellement impossible dans un même texte.

Le code MAC ne semble pas avoir fait l'oubli de caractères français. En revanche, comme DOS-ASCII, il ne permet pas d'encoder les caractères de nombreuses langues.

Ainsi, certains graphèmes du français ont presque disparus de beaucoup de textes (c'est le cas du *æ*), certains caractères typographiques ont totalement disparus (les élisions de *st*, *Qu*, les tirets de tailles différentes distingués pour les incises, traits d'union ou intervalles de nombres, espaces «fines» et demi-cadratins, etc.) et l'usage d'encodage appauvri fait parfois disparaître certains graphèmes pour l'écriture de courriers électroniques.

Cet appauvrissement du jeu des caractères du français devrait appartenir désormais au passé de l'informatique. Il existe aujourd'hui quelques encodages respectant la diversité des graphies du français (Unicode ou MAC par exemple).

En 1989, des constructeurs et développeurs de logiciels ont développé une norme, l'ISO-10646, censé couvrir l'ensemble des graphies des langues du monde. Pour cela, le choix s'est porté sur 32 bits d'encodage, soit une possibilité de $2^{32} = 4{,}29.10^9$, plus de quatre milliards de caractères disponibles. Cette norme offre une compatibilité avec l'ASCII (les 7 bits de poids faible) et l'ISO-8859-1 (les 8 bits de poids faible). L'encodage de 25 systèmes d'écriture est désormais terminé mais le projet est en cours. Un consortium, Unicode, exploite actuellement les 16 bits de poids faible de l'ISO-10646 comme standard d'encodage des caractères. Il devient le standard actuel pour la compatibilité qu'il offre, en amont de l'ASCII et de l'ISO-8859-1, en aval de l'ISO-10646.

En marge de ces encodages, des normes se sont développées permettant d'encoder de façon logique les graphèmes. En Latex ou SGML, il suffit de décrire le graphème selon un format standardisé en donnant la lettre et ses accents diacritiques. L'encodage se fait alors sur plusieurs *mots* («`&agrave;`» ou «`\‘{a}`» pour encoder «à» en SGML ou Latex par exemple). Ces descrip-

tions offrent de plus l'avantage de pouvoir être encodé avec un jeu restreint de caractères (sept bits suffisent), on les voit donc se multiplier comme moyen de transmission de textes en typographie riche sur sept bits. Malheureusement, il n'y a pas plus de consensus sur ces méthodes que pour encoder un caractère sur huit bits et MIME, U7, U8, HTML, uuencode, BinHex sont autant d'exemples d'encodage incompatibles deux à deux que les constructeurs proposent. Notons que la norme XML exploite l'encodage Unicode et qu'il n'est plus besoin de encoder les lettres comme en SGML.

Nous avons retenu l'ISO-8859-1 comme encodage des caractères du corpus pour le rendre le plus disponible possible à la communauté linguistique. En effet, beaucoup ne sont pas encore passés à l'Unicode et l'ISO-8859-1 est actuellement la norme la plus utilisée en France.

La ligature «œ» n'est pas réellement un problème en français car l'opposition graphique entre «oe» et «œ» suit une variation phonétique libre. «Oe» s'écrit pour les diphtongues /oe/, /oɛ/, /oã/, /oɛ̃/ ou dans des noms propres empruntés (Boeing, Monroe). «Œ» s'écrit pour les voyelles /ø/ ou /œ/.

Or il n'y a pas de distribution complémentaire entre /oe/, /oɛ/, /oã/, /oɛ̃/ d'une part et /ø/, /œ/ d'autre part en français. *Groenland* peut se prononcer indifféremment /grœnlãd/ ou /grwɛnlãd/ et *proeuropéen* peut s'entendre /prœrœpeɛ̃/ par exemple.

On peut donc en conclure que «œ» et «oe» sont deux réalisations possibles du même graphème. L'usage est aujourd'hui de remplacer les «œ» par des «oe» systématiquement dans l'édition peu soignée (courrier électronique, pages Internet). Le remplacement inverse est toujours possible en consultant des lexiques qui conservent la graphie normative.

- *cœrcitif
- *cœfficient
- *cœntreprises
- *prœuropéen
- *grœnland
- manœuvre
- manoeuvre
- bœuf
- boeuf

4.1.3 SGML

SGML (Standard Generalized Markup Manguage) est une norme ISO (ISO-8879) offrant la possibilité de définir un format d'échange de données informatiques.

Cette norme a des atouts qui font qu'elle est aujourd'hui très largement utilisée dans les domaines de l'édition, de la gestion des bases de données informatiques et de l'industrie des langues. HTML (HyperText Markup Language), le langage de définition des pages *Web*, est une application de la norme SGML. Nous ne pouvons cependant pas parler d'une seule application HTML (chaque concurrent définit régulièrement une nouvelle version en s'efforçant de la rendre incompatible avec les autres ou avec les normes SGML elle-même) et ce format de définition de documents hypertextes souffre de ne pas être exactement une application SGML qui a les avantages suivants :

- N'est pas *propriétaire*. La norme ISO-8879 n'appartient à aucune société qui pourrait revendiquer des droits d'exploitation ou qui pourrait faire évoluer la norme par stratégie commerciale.
- Proposer des documents autosuffisants, c'est-à-dire qu'un document SGML contient tout ce qui permettra de décoder son contenu.
- Rend explicite le langage formel encodant l'information.
- Est indépendant du matériel et du logiciel utilisé.
- Encoder les caractères d'un document quelconque sur 7 bits (Il est toutefois possible d'adopter l'ISO-8859-1 comme nous le verrons)

Un document SGML se compose de trois parties :

- Une déclaration SGML. Cette section est très rarement, sinon jamais présente, elle permet de modifier la syntaxe et les caractères propres au balisage. Nous ne l'utiliserons pas, et n'en parlerons pas plus.
- La DTD (Document Type Definition).
- Une instance de document.

DTD

Une DTD est, entre autres définitions, une grammaire formelle de réécriture proche des CFG (avec cependant la possibilité de décrire des langages infinis avec un nombre de dérivations fini [4]).

4. La grammaire de réécriture d'une DTD peut contenir une expression de Kleene dans une partie droite de règle.

Cette grammaire formelle permet de générer toute instance de document. Un document SGML est équivalent à une description syntagmatique (ou un arbre). Chaque structure et sous-structure du document ; un **élément**, correspond à un nœud d'un arbre qui peut être étiqueté avec du texte et avec des **attributs** marquant les propriétés de ces structures.

Élément La syntaxe de la définition d'un élément dans la DTD est la suivante :

⟨ `!ELEMENT Nom - -` (*Description de la sous-structure*) ⟩

Où la description de la sous-structure est une opération sur d'autres éléments ou un contenu textuel.

Les contenus textuels sont les suivants :

- EMPTY Contenu vide
- ANY Contenu indifférent : PCDATA ou tout élément
- RCDATA Texte pouvant contenir des entités interprétées
- PCDATA Texte ne contenant pas d'entité ou d'élément
- CDATA Texte indifférent non analysé

Les opérateurs sont les suivants :

- `A , B` séquence ordonnée
- `A | B` disjonction
- `A & B` séquence d'ordre indifférent (même chose que `((A,B) | (B,A))`
- `A*` (étoile de Kleene)
- `A?` (même chose que `(EMPTY | A)`)
- `A+` (même chose que `(A, A*)`)

En plus de cette description syntaxique de la structure de l'élément, sa définition dit comment elle devra être écrite dans l'instance de document. Les signes «−» indiquent que l'élément sera encadré par une balise ouvrante et par une balise fermante. Ces signes peuvent être remplacés par «0» rendant optionnelles ces balises.

Entités Elles permettent de décrire des objets quelconques dans l'instance de document, comme par exemple les caractères non-encodés. C'est ainsi que des caractères mathématiques, symboliques ou de l'écriture d'une langue dont le jeu de caractères n'est pas pris en charge peuvent être encodés. Il n'y a pas de limite théorique à cet aspect de la définition du type de document ; une entité peut représenter absolument tout élément atomique de la structure de document. L'usage veut que l'encodage des caractères se fasse en indiquant la lettre, ou une description de cette lettre, suivi éventuellement du nom des accents diacritiques ou d'autres indications (ligature, double accent, prononciation, etc.).

Les codes les plus souvent rencontrés (notamment pour HTML) sont[5]

acute	*accent aigu*
grave	*accent grave*
uml	*tréma*
circ	*accent circonflexe*
cedil	*cédille*
midot	*point moyen (catalan)*
lig	*ligature de deux lettres (français, allemand)*
caron	*v suscrit (tchèque, croate)*
strok	*lettre barrée (maltais, norvégien)*
dblac	*double accent aigu (hongrois)*
dot	*point (polonais)*
ogon	*cédille inverse (litualien)*
nodot	*sans point (turc)*
ring	*rond suscrit (norvégien)*
tilde	*tildé (espagnol, portugais)*

Les entités ont également le rôle de *macros*. Elles permettent de faire des substitutions dans la DTD ou dans l'instance de document pour factoriser des informations redondantes.

La syntaxe de la définition d'une élément dans la DTD est la suivante :

⟨ `!ENTITY Nom` *Description*⟩

Attributs Les attributs permettent d'indiquer les propriétés propres à chaque élément de l'instance de document. Dans la DTD, le nom et le type de chaque attribut doit être mentionné pour chaque élément.

5. Source : communication personnelle de Jérôme Vachey - Société Erli.

Les types possibles sont des données textuelles, des nombres, des éléments, des entités, etc.

Deux types particulièrement intéressants d'attributs sont les identifieurs d'éléments **ID** et les références de ces identifieurs **IDREF**. Les uns servent à identifier de façon unique un élément dans une instance de document, les autres à faire référence à ces premiers dans n'importe quel élément du document. Ces liens suffisent à représenter un graphe et non plus seulement un arbre avec un document SGML. Il reste que la structure privilégiée d'un document SGML est l'arbre bien qu'il soit toujours possible de définir des relations croisées entre les nœuds de cet arbre.

Les documents édités sont le plus souvent organisés sous cette forme. Un livre est composé de tomes, parties, chapitres, paragraphes, etc. Un dictionnaire de sections, articles, sous-articles, etc., dans lesquels s'articulent des descriptions phonétiques, graphiques, étymologiques, etc. Les documents hypertextes ne sont pas organisés autrement, les liens croisés en font une structure de graphe et sont encodés grâce aux mécanismes d'attributs d'éléments.

Instance de document

La troisième partie du document SGML est donc un élément du langage défini par la DTD. Sa structure est un arbre dont chaque nœud est décrit par une balise ouvrante et ses attributs, un contenu éventuellement vide et une balise fermante. Les pages *Web* en sont des exemples.

La structure des documents SGML est utilisée dans l'industrie éditoriale pour présenter les documents non plus suivant la typographie et la mise en page mais suivant leur structure logique. Ceci est particulièrement vrai pour l'édition lexicographique. Les séparations d'un article de dictionnaire sont associées à des parties de l'article que l'on peut aisément baliser comme autant d'éléments différents (tête de l'article, partie grammaticale, étymologie, flexion, acceptions). Les indications typographiques présentes (gras, italique, capitale, entre guillemets, etc.) sont indicatives de la nature des sous-parties (exemple, étymon, forme, etc.) mais ne sont pas informatives en soi. Le fait qu'un mot soit écrit en gras peut signifier un grand nombre de choses dans un même article de dictionnaire.

L'organisation d'un article de dictionnaire en SGML ne s'entend pas comme sa disposition et présentation graphique (on n'indique rien sur l'emploi de telle ou telle fonte), mais comme son organisation logique. Une représentation arborescente est alors bienvenue puisqu'un article se décompose

hiérarchiquement.

Ceci est vrai également de nombreux documents textuels qui composent les corpus informatisés comme les romans, articles, bibliographies, etc. SGML est donc employé actuellement comme technique éditoriale pour séparer l'organisation de la structure d'un document (travail qui peut être effectuée par un auteur, un lexicographe, un documentaliste) des travaux de mise en page effectués par l'imprimeur.

4.1.4 XML

Un document bien formé en SGML contient donc une instance de document produite par la DTD qui la précède. En revanche, il n'est pas possible de prédire la DTD minimale qui génère une instance de document car les balises ouvrantes et fermantes sont facultatives en SGML. L'analyse d'un document SGML suppose donc que la DTD ait été analysée au préalable.

XML (Extensible Markup Language) est une restriction de la norme SGML visant à rendre l'instance de document autosuffisante. Les balises XML sont systématiquement ouvertes et fermées et une balise vide s'écrit `<NOM/>`. C'est en effet lorsqu'une balise est nécessairement vide (comme une balise `<p>` indiquant la séparation de deux paragraphes) qu'il est fréquent de ne pas indiquer la balise fermante en SGML. Il est donc fort possible — l'usage l'interdit toutefois — d'écrire des DTD ambiguës en SGML et de produire des documents qui ne correspondent pas de façon bi-univoque à une analyse syntaxique.

XML est donc une restriction de SGML qui permet entre autres choses de prouver les algorithmes et de garantir une analyse sur le document encodé[6].

Les utilisateurs d'XML se gardent souvent d'écrire les DTD de leurs documents. Cela est toujours possible puisqu'un document XML suppose une DTD minimale dans la mesure où il est bien formé. Cependant, dans les contextes de réutilisabilité, une DTD permet de garantir la cohérence des documents échangés. De plus la définition de la DTD est le travail nécessaire de généralisation et systématisation de l'organisation du type de document.

Rappelons par ailleurs que l'encodage des caractères est Unicode dans un document XML. Il est toutefois possible d'indiquer une autre norme. Le Corpus de Paris 7 est encodé en ISO-8859-1 comme nous l'avons dit.

6. En toute généralité, un document XML appartient à un langage CFG augmenté d'attributs, et le graphe des attributs ne contient pas de cycle.

TEI

La TEI (Text Encoding Initiative) [Burnard & Sperberg-McQueen, 1996] est un ensemble de recommandations qui s'adressent à ceux qui souhaitent échanger, stocker, et diffuser des textes sous forme électronique.

Le principe est de rendre systématiquement explicite les caractéristiques de ces textes ainsi que leur structure.

Les caractéristiques sont :

- la description bibliographique du texte électronique
- la description de la manière dont il a été encodé
- une description libre du texte
- l'historique des révisions du texte

La description de la structure du document est très détaillée et s'appuie sur une division hiérarchique du texte en sections, chapitres, paragraphes, etc. Mais également en références diverses pour faire face aux représentations de notes, index, hyperliens, figures, etc.

La TEI est le fruit d'un long projet proposé par l'*Association for Computers and the Humanities*, l'*Association for Computational Linguistics* et l'*Association for Literary and Linguistic Computing* qui a donné lieu en mai 1994, après six ans de travail, a une imposante documentation. 400 éléments sont proposés pour structurer les documents.

Les recommandations suivent la norme SGML ISO-8879 qui s'avère particulièrement adaptée à cet encodage. La TEI a donc donné lieu à une DTD qui permet de définir sans ambiguïté un document bien formé.

Si la TEI est particulièrement riche pour encoder les types de documents, elle l'est moins pour l'encodage de corpus. ([Habert *et al.*, 1998] p.94) notent que l'élément «**w**» utilisé pour segmenter le mot est disponible dans les recommandations de la TEI, mais nullement défini du point de vue linguistique. La TEI laisse à l'utilisateur le soin d'utiliser cet élément pour marquer diverses réalités. C'est le principal défaut de la TEI: le respect des recommandations de la TEI se garantit pas un encodage reproductible. De plus, un grand nombre de réalités de l'étiquetage morpho-syntaxique sont manquants, mais rien n'empêche l'utilisateur d'ajouter ses propres éléments et attributs.

Parmi les objectifs de la TEI définis lors de la conférence préparatoire tenue au Vassar College de New-York en novembre 1987, la TEI devait «être modifiable par l'utilisateur». Il s'agit donc bien de *recommandations* que les utilisateurs suivront de plus ou moins près en les adaptant à leurs besoins.

Nous avons suivi les recommandations de la TEI pour étiqueter le Corpus de Paris 7. L'en-tête proposée (⟨*tei-header*⟩) pour annoter les caractéristiques du document nous suffisait amplement pour indiquer l'origine du corpus.

En collaboration avec Laurent Romary, nous avons fait ces quelques modifications et ajouts de la TEI pour l'encodage du corps du document :

- La balise `<s>` pour marquer la phrase typographique.
- La balise `<w>` de la TEI est utilisée pour désigner notre «mot» qu'il soit simple ou non. Nous avons utilisé la même balise pour désigner les composants des mots composés en les enchâssant dans d'autres balises `<w>`.

  Le texte marqué par cette balise (*CDATA*) est le texte source, c'est-à-dire la graphie du document telle qu'elle apparaît dans les colonnes du Monde.
- La balise `<w>` contient les attributs marquant le mots en morpho-syntaxe :
  - **lemma** Le lemme.
  - **cat** La catégorie ou «partie du discours»
  - **subcat** La sous-catégorie
  - **mph** La morphologie, sous la forme d'un code résumant le temps et mode verbal, la personne, le nombre, le genre et le nombre du possesseur pour *leur(s)*

4.1.5 XSL

Un document XSL (Extensible Stylesheet Language) est une déclaration au format XML (XSL est une application de XML) qui exprime un transcodage d'une DTD vers une autre. Autrement dit, un document XSL permet de construire un texte balisé XML (par exemple une page Internet) depuis un document au format XML (par exemple le corpus de Paris 7). Il est également possible de produire des documents non XML grâce à une spécification XSL ; du texte seul ou un document HTML.

Le rôle habituel d'une déclaration XSL est la mise en page d'un document XML. C'est-à-dire l'exploitation automatique de l'organisation logique d'un document pour sa mise en page. Les spécifications XSL sont donc souvent des «feuilles de style», œuvres de graphistes, d'éditeurs et de typographes. Nous appellerons «feuilles de style» par abus de langage tout document XSL.

Mais XSL permet également de livrer un document annoté dans n'importe quel format respectant toutefois les normes XML. C'est de cette façon que nous pouvons livrer le corpus de Paris 7 pour telle ou telle application qui ne supporte pas le format XML.

XSL permet de modifier l'encodage des balises et d'ajouter des informations (comme des attributs nouveaux sur des éléments par exemple). Mais XSL permet également de modifier la structure même du document. Le langage XSL bénéficie des spécifications *XPath* du consortium *W3*. *XPath* a pour but de désigner de façon générique un élément dans l'arborescence d'un document XML. XSL applique sur ces éléments un certain nombre de procédures comme des recherches sur le contenu ou les attributs, des tris, des boucles, des tests, des comptages, etc.

XSL offre donc un grand nombre de possibilités pour produire un nouveau document. Cependant la limite d'XSL est telle qu'il ne nous permettra pas de trouver dans le document des suites de mots correspondant à une grammaire (même régulière). En effet, XSL, contrairement à d'autres outils comme SgmlQL ([Murisasco, 1996], [Maître *et al.*, 1998]) ne fournit pas de mécanisme, à notre connaissance, pour trouver une suite d'éléments XML dans un langage.

Cela supposerait un mécanisme de manipulation des disjonctions, conjonctions et étoile de Kleene sur les éléments XML qui n'existe pas en XSL.

C'est la raison qui nous a fait développer le programme *Cluster* dont il sera question en 4.4.2.

La transcription d'un document grâce à XSL est faite grâce un compilateur[7]. Nous utilisons le logiciel du consortium *GNU* Sablotron qui possède les qualités requises dont une bonne rapidité de fonctionnement. Cet aspect est souvent négligé mais nous pensons pouvoir utiliser Sablotron dans l'avenir pour traduire le corpus en temps réel grâce à des «feuilles de style» générées automatiquement par les outils d'exploitation du Corpus. Sans cette perspective, notre choix aurait certainement été autre.

Nous avons donc privilégié un programme écrit en code C[8] (tout autre

7. Nous utilisons de terme de compilateur pour désigner un programme qui traduit un langage formel vers un autre.

8. Les outils XML sont très souvent liés à la programmation Java sans que le lien soit très clairement établi. La portabilité d'un logiciel écrit en langage Java est due au fait que le code compilé est presque toujours interprété par un émulateur de machine Java. Un tel langage doit donc être réservé à des tâches courtes puisqu'à défaut d'une machine Java, le programme sera ralenti du temps nécessaire à l'interprétation du code. Par ailleurs,

code compilé aurait fait l'affaire) pour la rapidité d'exécution qu'il offre au détriment de sa portabilité (Nous l'avons compilé sur plusieurs systèmes unix sans difficultés toutefois). Les développeurs de Sablotron ont encodé les caractères en Unicode, le logiciel accepte par ailleurs ISO-8859-1. Nous n'avions donc aucun obstacle à l'utiliser pour appliquer des «feuilles de style» sur le Corpus de Paris 7.

4.2 Les méthodes d'étiquetage automatique

Les étiqueteurs, ou désambiguïsateurs morpho-syntaxiques sont devenus des outils classiques du traitement automatique des langues. Leur fonction est d'assigner une étiquette – en général la partie du discours – unique pour chaque mot d'un texte. La segmentation en mots précède généralement cette étape, nous verrons qu'il est heureux qu'elle soit intégrée à l'étiqueteur dans certains cas, les étiquettes permettant de lever certaines ambiguïtés de segmentation.

Ces étiquettes sont simplement attribuées en consultant des lexiques qui associent une liste d'étiquettes à chaque forme, comme le DELAS ou le DELAC ([Silberztein, 1993]) ou encore par l'application de règles morphologiques. Le plus souvent, le recours aux règles morphologiques n'a lieu que pour les mots absents du lexique comme c'est le cas du Brill Tagger [Brill, 1992]. L'obstacle matériel qui empêchait de conserver un lexique robuste en mémoire d'un ordinateur faute de place appartient désormais à l'histoire de l'informatique.

Les ambiguïtés portant sur les homographes sont résolues en discours par l'étude de la distribution de chaque étiquette. Deux méthodes sont connues pour lever cette ambiguïté : l'exploitation de probabilités conditionnelles établies grâce à un corpus de référence ou l'approche par règles écrites par des linguistes.

---

l'encodage des caractères en Java est nécessairement Unicode. Cette norme a été également utilisée en XML comme nous l'avons vu. Il semble donc que Java soit souvent choisi comme langage de développement de ressources XML pour cette autre raison. Le critère qui nous fait choisir Java comme langage de développement est plutôt la portabilité des logiciels et leur interfaces.

4.2.1 Les étiqueteurs par règles

En français, certaines suites de catégories sont impossibles. Par exemple, un déterminant ne peut jamais être suivi d'un verbe, de même qu'un clitique d'un nom ou d'un déterminant.

Ainsi l'étiquette *Déterminant* proposée possiblement pour *le* dans *Jean ne le porte pas* peut être éliminée par une règle négative.

L'exhaustivité combinatoire des séquences de catégories d'une suite textuelle peut être représentée par un unique graphe dont certains parcours sont illicites. Les règles négatives permettent de marquer ces parcours pour ne laisser que leurs alternatives.

Inversement, un graphe permet de décrire les suites possibles de catégories et elles seules. En français, la suite des clitiques en position préverbale peut se décrire par un tel graphe tant leurs places sont ordonnées. Mais le plus souvent ces graphes sont établis à partir d'une forme particulière ; ils décrivent les contextes immédiats possibles pour une forme donnée. Ces règles décrivent alors une grammaire ne faisant pas intervenir de notion de constituance mais les seules suites possibles qui précédent ou qui suivent une forme. On parle alors de grammaire locale. Le système INTEX [Silberztein, 1993] permet entre autres choses de construire de telles règles sous la forme d'un automate.

Les étiqueteurs par règles exploitent généralement les deux procédés. Leur fiabilité n'est pas mauvaise au regard des étiqueteurs probabilistes. Leur simplicité de mise en œuvre et l'optimalité des algorithmes les rendent séduisants, mais ils réclament des données nombreuses qui sont généralement écrites à la main. Ces données peuvent être partiellement établies automatiquement sur la base de corpus d'apprentissage grâce aux méthodes stochastiques.

4.2.2 Les étiqueteurs stochastiques

Les étiqueteurs stochastiques exploitent un corpus d'apprentissage pour établir la probabilité de l'assignation d'une étiquette en fonction d'un contexte. Ou, ce qui revient au même, de la probabilité d'une suite d'étiquettes. On parle alors de bi-grammes pour une suite de deux étiquettes, de tri-grammes pour une suite de trois, etc.[9]

9. Les probabilités conditionnelles peuvent être également représentées sous forme d'arbre de décision.

Cette probabilité est ensuite exploitée pour décider quelle étiquette assigner à une forme ambiguë. La probabilité est très forte (voir égale à 1) pour les n-grammes qui correspondent à des règles positives, et très faible ou nulle pour les n-grammes qui correspondent à des règles négatives. Les résultats sont donc assez proches des étiqueteurs par règles pour les quelques suites de catégories qui font intervenir un contexte immédiat. Les autres différences sont dues à la nature et à la taille du corpus d'apprentissage.

Mais la qualité de l'étiquetage ne peut dépendre naturellement que de la nature du contexte morpho-syntaxique. C'est-à-dire qu'un système probabiliste d'étiquetage ne peut exploiter que les fréquentes similitudes des contextes trouvés d'une même étiquette ou d'un même mot. On a trop tendance à penser que les probabilités peuvent fournir des résultats transcendant les données qui ont permis de les établir.

Le corpus d'apprentissage doit être d'une taille suffisante pour contenir de façon statistiquement significative les contextes d'une étiquette ambiguë. Par ailleurs, la taille du corpus d'apprentissage doit être également réglée en fonction de la granularité des étiquettes. Plus l'étiquette est grossière, plus elle a de chance d'apparaître dans les mêmes contextes.

Ajoutons que le corpus d'apprentissage doit être un corpus de référence par rapport au type de texte qui sera étiqueté. Les étiquettes morpho-syntaxiques doivent être les mêmes, et le texte ne devra pas contenir de suites peu représentées. Ainsi, un texte contenant des stéréotypes et locutions non repérées en tant que telles et absentes du corpus d'apprentissage sera certainement mal étiqueté. Il en va de même pour un texte oral étiqueté grâce à un étiqueteur entraîné avec un corpus de texte écrit. Les différences de registre, de domaines entre le corpus d'apprentissage et le texte produiront les mêmes effets.

Le modèle de Markov

Le modèle de Markov, du nom du statisticien russe qui étudia en 1907 les contraintes statistique des suites de lettres, revient à calculer la probabilité conditionnelle de réalisation d'une étiquette en fonction de la probabilité conditionnelle de réalisation des étiquettes précédentes.

Un modèle de Markov est un système dont le passage à chaque état dépend d'une probabilité fonction des états qui le précèdent. Le nombre d'états significatifs qui précèdent fixe l'*ordre* du modèle.

En pratique, le modèle est représenté par un graphe ordonné dont l'en-

semble des transitions depuis un même sommet est étiqueté par une probabilité. Chaque sommet du graphe peut correspondre, par exemple, à une étiquette morpho-syntaxique de telle façon qu'un parcours du graphe corresponde à une suite d'étiquettes, on parle de chaînes de Markov.

Lors de l'étape d'apprentissage, la fréquence du parcours d'un graphe depuis un état vers un autre donne une mesure de la probabilité assignée à la transition empruntée. L'application du modèle de Markov lors de la désambiguïsation n'est pas plus compliquée ; le parcours préféré est celui qui offre la plus forte probabilité. L'état du modèle dépend alors des états qui le précédent, qui à leur tour dépendent de l'historique des étapes du système. Il vient qu'une étiquette n'est pas seulement désambiguïsée grâce aux N étiquettes qui précèdent pour un modèle d'ordre N, mais également par les étiquettes antérieures qui ont influencé l'état du modèle.

On voit que le modèle est extrêmement simple à implémenter, c'est certainement l'une des raisons de son succès. De plus, il offre la possibilité d'exploiter des algorithmes rapides et économiques.

L'étiqueteur de Brill - *Trained rule-based tagger*

La méthode proposée par Éric Brill [Brill, 1992], est une méthode probabiliste dans le sens où l'ensemble de la procédure peut être mise en œuvre grâce à un corpus d'apprentissage. Mais contrairement aux autres méthodes probabilistes, elle permet d'intégrer des connaissances linguistiques sur l'étiquetage morpho-syntaxique. Cette méthode fait l'usage de règles très proches de celles des étiqueteurs par règles, cependant une probabilité est appliquée fixant un niveau de *performance* des règles. Seules sont appliquées les règles qui maximalisent la probabilité d'améliorer l'étiquetage et qui minimisent la probabilité de le dégrader.

L'absence de visibilité de cette probabilité lors de l'application des règles et le caractère plus linguistique de leur expressivité font souvent penser que le *Brill Tagger* est un étiqueteur ne faisant pas intervenir de probabilités du tout.

Remarquons que la méthode de Brill nous montre qu'il est possible de concilier des méthodes stochastiques avec des connaissances linguistiques qui seront exploitées pour établir le type de règles contextuelles ; ce qui semblait échapper aux autres méthodes probabilistes. Un étiqueteur basé sur le modèle de Markov, par exemple, assigne une étiquette à un mot de façon à maximaliser la probabilité $Prob(mot|etiq) * Prob(etiq|Netiq.prec.)$. Les connaissances

linguistiques du mot, de l'étiquette ou du contexte ne peuvent pas être exploitées pour améliorer le modèle. Les systèmes purement stochastiques sont souvent couplés à des étiqueteurs par règles pour cette raison ; il nous semble que la méthode d'étiquetage de Brill est une bonne alternative.

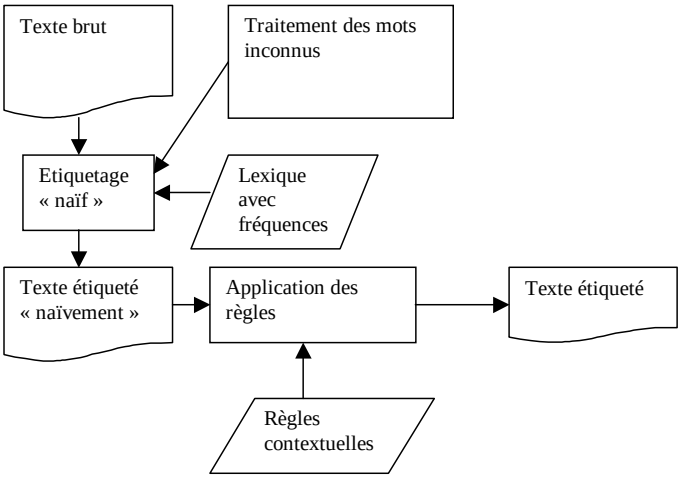


FIG. 4.1 – *Principe d'assignation des étiquettes avec l'étiqueteur de Brill*

Principe Pour ce qui suit, nous appellerons *corpus d'apprentissage* un corpus bien étiqueté représentatif des textes qui devront être étiquetés par le système. Nous appellerons *Corpus de travail* ce même corpus non étiqueté.

Le système procède en deux passes. Dans un premier temps, il assigne des étiquettes naïvement, c'est-à-dire sans considérer le contexte en attribuant l'étiquette la plus probable pour le mot selon un lexique construit à partir du corpus d'apprentissage. Les mots qui n'apparaissent pas dans ce lexique ont une étiquette calculée plus ou moins heureusement selon des critères morphologiques, en déterminant s'ils commencent par une capitale ou en fonction de leur terminaisons.

Lors d'une deuxième passe, des règles contextuelles sont successivement appliquées pour éventuellement corriger cette première étiquette.

A l'issue de ces deux étapes, le texte est étiqueté avec une qualité qui dépend directement de la pertinence des règles de correction.

Entraînement Au préalable, le système calcule pour chaque mot la probabilité qu'il soit assorti de chacune des étiquettes possibles. Ce calcul se fait simplement en consultant le corpus d'apprentissage et en comparant les fréquences d'occurrences $(etiq|mot)$. Ainsi, *la* aura une probabilité d'être respectivement déterminant, pronom ou nom d'environ 0,99, 0,01, moins de 10^{-4} selon un calcul simple sur le corpus de Paris 7.

Le système assigne les étiquettes les plus probables aux mots du corpus de travail. *la* sera alors systématiquement étiqueté déterminant dans ce corpus, quelle que soit sa position.

L'hypothèse est que la comparaison du corpus de travail avec le corpus d'apprentissage permettra de mettre en évidence les contextes communs aux mêmes différences. Ainsi, à chaque fois que *la* sera étiqueté déterminant dans le corpus de travail et pronom clitique dans le corpus d'apprentissage, on verra converger un contexte C commun à toutes les occurrences permettant de corroborer l'hypothèse d'une règle du type :

$det \rightarrow clitique/C$

On voit qu'il est crucial qu'il n'y ait pas d'interférence entre la première passe (l'étiquetage naïf) et le contexte C dictant les règles de correction. Il n'est néanmoins pas nécessaire que l'étiquetage soit complètement hors contexte, pourvu que les règles contextuelles en tiennent compte.

Dans l'étiqueteur original [Brill, 1992], les règles dont il était fait l'hypothèse étaient les suivantes :

Changer l'étiquette a en b si

1. Le mot précédent (resp. suivant) est étiqueté z
2. Le deuxième mot qui précède (resp. qui suit) est étiqueté z
3. L'un des deux mots qui précèdent (resp. qui suivent) est étiqueté z
4. L'un des trois mots qui précèdent (resp. qui suivent) est étiqueté z
5. Le mot précédent (resp. suivant) est étiqueté z et le mot suivant (resp. précédent) est étiqueté w
6. Le mot précédent (resp. suivant) est étiqueté z et le mot deuxième mot suivant (resp. précédent) est étiqueté w

Ce type de règles ne permet pas de décrire des contextes fort différents de ceux qui sont captés par les étiqueteurs stochastiques. Ainsi, les résultats de l'étiqueteur de Brill 92 ne différent que très peu des résultats d'un modèle de Markov ([Brill, 1993]).

Par la suite, Éric Brill [Brill, 1993] a étendu le modèle pour qu'il puisse tenir compte d'informations lexicales comme contexte d'application d'une règle d'étiquetage. Cela permettait de capter le contexte permettant d'étiqueter *as* adverbe puis préposition dans les tournures *as ... as* ; ce qui échouait tant avec les étiqueteurs stochastiques qu'avec le Brill tagger.

Changer l'étiquette a en b si

1. Le mot précédent (resp. suivant) est w
2. Le deuxième mot qui précède (resp. qui suit) est w
3. L'un des deux mots qui précèdent (resp. qui suivent) est w
4. Le mot courant est w et le mot précédent (resp. suivant) est x
5. Le mot courant est w et le mot précédent (resp. suivant) est étiqueté z

Lors de la phase d'apprentissage, l'exhaustion combinatoire des règles supposées apporter une amélioration à l'étiquetage naïf est produite automatiquement. Ces règles sont tour à tour appliquées au corpus de travail et le résultat est comparé au corpus d'apprentissage. Chacune des règles corrige un ensemble d'étiquettes à tort ou à raison et la différence entre le nombre d'erreurs corrigées et le nombre de nouvelles erreurs permet d'évaluer la performance de la règle. Une probabilité est appliquée pour l'ensemble des règles, et seules les règles qui ont une forte probabilité d'amélioré le corpus de travail (et une faible probabilité de le dégrader) sont conservées.

Pour reprendre notre exemple, la règle *changer Det en Clitique si le mot suivant est un verbe* aura certainement une forte probabilité d'améliorer le corpus et une faible probabilité de le dégrader, elle sera donc conservée contrairement à la règle *changer Det en Clitique si le mot suivant est un nom*.

Le *tagger* de Brill souffre cependant de quelques limites et son implémentation sans adaptation ne donne pas les résultats que l'on attend d'un étiqueteur robuste et fiable.

L'étiqueteur de Brill ne contient pas de segmenteur en phrases et mots composés. Or cette annotation est importante en soi mais également pour limiter les ambiguïtés artificielles des composants de formes composés.

Le lexique de corpus (du corpus d'apprentissage) est le seul dictionnaire qui est exploité par l'étiqueteur de Brill. Les mots inconnus se trouvent en nombre lors de l'étiquetage *naïf* et les mécanismes de calcul de l'étiquette grâce à la morphologie flexionnelle sont insuffisants.

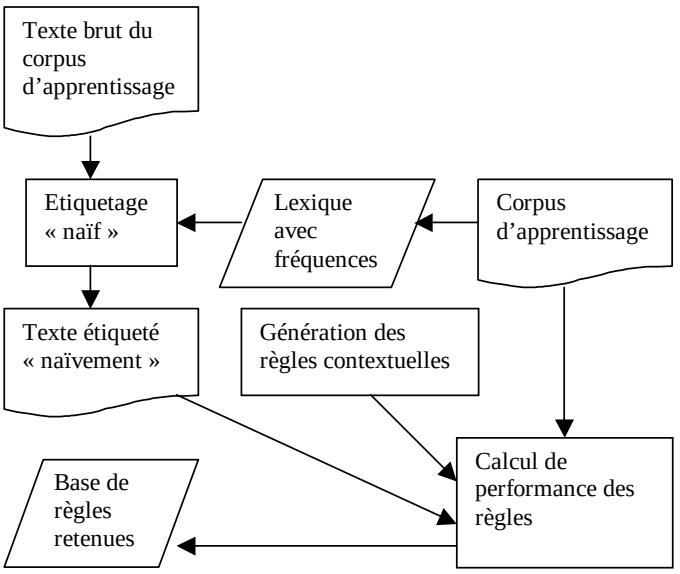


FIG. 4.2 – *Principe d'entraînement de l'étiqueteur de Brill*

Éric Brill lui-même avait admis que les règles étaient trop peu expressives dans leur version initiale ([Brill, 1993]). Les règles contextuelles devraient pouvoir exprimer des généralisations linguistiques ne se limitant pas à l'identification de quelques mots à droite et à gauche. Les phénomènes d'accord par exemple devraient pouvoir s'exprimer grâce aux règles contextuelles.

L'implémentation initiale du *Brill tagger* n'est pas optimale, Emmanuel Roche et Yves Schabes montrent par exemple qu'il est possible de rapporter un jeu de règles contextuelles à un transducteur fini ([Roche & Schabes, 1995].)

Nous ne connaissons pas d'implémentation du *tagger* de Brill qui prenne en compte ces critiques. Il semble cependant que tous les points que nous avons soulevé ne rendent pas rédhibitoire l'algorithme lui-même et qu'il serait possible d'implémenter un transducteur fini comme le proposent Emmanuel Roche et Yves Schabes, pondéré de probabilités calculées en fonction d'un corpus d'apprentissage. La projection d'un lexique se voulant exhaustif et la segmentation en mots et en phrases seraient également implémentables lors de la première phase. Nous discutons maintenant comment la phase *naïve* peut être relativement sensible au contexte dans le but d'étendre l'expressivité des règles contextuelles.

Contexte étendu Le principe du *Brill Tagger* repose sur l'exploitation automatique des différences entre le corpus de travail (le corpus étiqueté «naïvement») et le corpus d'apprentissage. Dans son état actuel, la première passe ne tient pas compte du contexte des termes à étiqueter pour que les différences d'étiquetage entre le corpus d'apprentissage et le corpus étiqueté «naïvement» soient conditionnées par ces contextes.

Nous pouvons nous interroger sur ce qu'apporterait un étiquetage «naïf» qui prenne en compte le contexte. Nous nous trouverions dans la situation d'un étiqueteur classique pour le première passe (étiqueteur par n-gramm, par chaînes de Markov, par arbre de décision, etc.). Puis nous pourrions induire automatiquement des règles qui intéresseraient un contexte plus étendu que les mots immédiatement adjacents au terme à étiqueter.

Cette technique offre une solution pour exploiter un contexte large et maîtrisé dans le cadre des méthodes stochastiques, ce qui fait largement défaut : les systèmes existant voient leurs résultats se dégrader rapidement lorsque le contexte est étendu au delà de trois ou quatre mots, et n'offrent généralement pas la possibilité d'exploiter une autre donnée qu'une suite de mots précédents le mot à désambiguïser. Nous la retenons également pour l'intérêt qu'elle offre de réduire la combinatoire des règles générées. Toutes les règles redondantes avec le mécanisme stochastique de la première passe sont en effet inutiles, il est donc possible d'augmenter leur portée ou leur finesse avec la même puissance de calcul. Nous pensons qu'une telle technique pourrait être avantageusement appliquée à d'autres systèmes qui réclament un contexte étendu comme l'analyse syntaxique probabiliste utilisant des grammaires lexicales (comme le *supertagging* en grammaire TAG).

4.2.3 L'étiqueteur développé à Paris 7

L'étiqueteur de Paris 7 a été implémenté par Rodrigo Reyes en 1997 [Reyes, 1997] lors d'un stage. Nous l'avons par la suite augmenté, nous lui avons ajouté quelques fonctionnalités et nous avons implanté un certain nombre de règles avec Louis-Gabriel Pouillot en 1999. Cet étiqueteur fonctionnant avec un dictionnaire, nous avons corrigé et augmenté ce dictionnaire de façon significative de sorte que l'ensemble des termes du corpus que nous devions étiqueter exhaustivement soient reconnus.

Le principe de cet étiqueteur consiste à appliquer des données stochastiques dans une première passe puis des règles contextuelles dans une seconde. Il s'agit donc d'un étiqueteur mixte.

La première passe utilise les étiquettes qui précèdent ou qui suivent immédiatement le mot à désambiguïser pour décider quelle étiquette attribuer selon une méthode probabiliste. C'est la méthode connue sous le nom de «trigrammes» que nous avons présenté supra, nous reviendrons sur ce qu'elle a d'original dans l'étiqueteur développé à Paris 7.

format des règles La syntaxe des règles permet de décrire une transformation du type

$A \rightarrow B/C$

où

- A et B sont des étiquettes dont on distingue la morphologie pour régler les phénomènes d'accord.
- C est le contexte décrit par une conjonction $C_1\&C_2\&C_3\&\ldots\&C_k$ où C_i désigne l'une de ces possibilités :
  1. Le mot précédent (resp. suivant) est étiqueté z
  2. Le deuxième mot qui précède (resp. qui suit) est étiqueté z
  3. Les deux mots qui précèdent (resp. qui suivent) sont étiquetés z et x
  4. L'un des deux mots qui précèdent (resp. qui suivent) est étiqueté z
  5. L'un des trois mots qui précèdent (resp. qui suivent) est étiqueté z
  6. Le mot précédent (resp. suivant) est étiqueté z et le mot suivant (resp. précédent) est étiqueté w
  7. Le mot précédent (resp. suivant) est étiqueté z et le mot deuxième mot suivant (resp. précédent) est étiqueté w
  8. Le mot précédent (resp. suivant) est w
  9. Le deuxième mot qui précède (resp. qui suit) est w
  10. Les deux mots qui précèdent (resp. qui suivent) sont w et x
  11. L'un des deux mots qui précèdent (resp. qui suivent) est w
  12. Le mot courant est w et le mot précédent (resp. suivant) est x
  13. Le mot courant est w et le mot précédent (resp. suivant) est étiqueté z
  14. Le mot précédent (resp. suivant) est reconnu par l'expression régulière r
  15. Le deuxième mot qui précède (resp. qui suit) est reconnu par l'expression régulière r

16. Les deux mots qui précèdent (resp. qui suivent) sont reconnus par les expressions régulières r et s
17. L'un des deux mots qui précèdent (resp. qui suivent) est reconnu par l'expression régulière r

Lexiques L'étiqueteur de Paris 7 utilise un lexique complet.

Nous avons étendu le lexique en étiquetant l'intégralité du corpus grâce au programme INTEX ([Silberztein, 1993]) qui contient le dictionnaire DELACF. Ce dictionnaire développé au cours des années 90 dans le laboratoire du LADL contient plus de 150 000 formes composées [Silberztein, 1996]. Nous avons croisé cette liste de mots composés avec celle que nous avions déjà pour ne valider que les composés identifiés par le système INTEX. Cette liste a été validée manuellement pour ne retenir que les mots composés qui répondaient à nos critères. En particulier, nous avons exclu de nombreux mots composés qui nous semblent être des collocations ou expressions figées. Nous avons ajouté de nombreux mots composés absents comme des noms propres, des adjectifs, des prépositions ou des verbes.

Les mots inconnus c'est-à-dire absents du lexique de l'étiqueteur, sont analysés grâce à un transducteur. Plus pratiquement, une liste d'expressions régulières mises en correspondance avec une classe syntaxique sont testées. Si le mot inconnu appartient au langage défini par l'expression régulière, la classe syntaxique lui est assigné lors de la première passe. Cette méthode était sensiblement celle du *Brill tagger*. Ces expressions régulières décrivent des caractéristiques du mot comme sa terminaison ou le fait qu'il commence par une capitale, qu'il contienne un caractère spécial comme un chiffre arabe ou un point. Nous avons privilégié la flexion des mots pour prédire l'étiquette morpho-syntaxique à apposer aux mots inconnus, la marque morphologique propres à certaines classes est donnée par la flexion du mot en français.

Pour prédire une classe syntaxique et la morphologie à partir d'une forme, nous avons calculé les fréquences des couples (*terminaisons*,*classe+morphologie*) automatiquement avec le corpus d'apprentissage et quelques manipulation de fichiers. Ensuite, nous avons établi arbitrairement un seuil de représentativité de ces fréquences pour ne garder que 200 terminaisons marquant possiblement une étiquette morpho-syntaxique.

Étiquetage stochastique de la première passe Nous avons dit que la première passe de l'étiqueteur de Paris 7 utilisait des trigrammes. Mais au lieu d'utiliser les trigrammes d'étiquettes, elle utilise des trigrammes de para-

digmes d'étiquettes. Ces paradigmes ont curieusement été appelé «génotypes» par [Tzoukermann *et al.*, 1995] et repris par Rodrigo Reyes [Reyes, 1997]. Cette métaphore ne dit rien de plus, elle suit une habitude de transposer la terminologie de la biologie et de la biochimie à notre science (nous pensons à la notion de valence de Tesnière). En l'occurrence, elle pourrait conduire à croire que les mécanismes de passage du génotype au phénotype sont transposables au passage du paradigme à l'étiquette ; ce qui n'est certainement pas vrai. Le principe est donc de mesurer la probabilité d'une étiquette sachant le mot et les n paradigmes d'étiquettes qui précèdent. Les mesures faites par Rodrigo Reyes montrent que cette probabilité est plus fine.

En outre, la probabilité n'est pas dégradée par le mauvais étiquetage des n mots qui précèdent comme c'est la cas des chaînes de Markov et autres méthodes par n-grammes. Les paradigmes sont des données qui sont issues du lexique et qui ne dépendent aucunement de l'étiqueteur. Il n'y a pas d'effet de dégradation successive (effet «boule de neige») de l'étiquetage.

Résultats Pour conclure cette présentation de l'étiqueteur de Paris 7, nous présentons ses défauts et qualités au regard de l'utilité qu'il peut rendre. Tout d'abord, rappelons qu'aucun étiqueteur permettant d'assigner les étiquettes définies pour l'étiquetage du corpus de Paris 7 n'était disponible lors de son développement. Les principaux défauts de cet étiqueteur sont

- la perte d'automaticité de l'ensemble de la procédure permettant d'obtenir les règles.
- la faible qualité de l'étiquetage que l'on doit relativiser en fonction de la finesse de l'étiquetage. Cette faible qualité s'explique par le fait que tout le travail conduisant à écrire les règles n'est pas encore terminé et que les règles ne sont pas déclaratives. En effet, l'application d'une règle peut être en conflit avec une autre ou peut être suivi de l'application d'une autre règle. La définition des règles par le linguiste est ainsi un travail délicat qui nécessite d'anticiper l'ensemble des opérations.

Ces défauts sont compensés comme nous l'avons vu par une finesse d'étiquetage assez bonne (122 étiquettes). Par ailleurs, l'étiqueteur segmente en phrases et en mots en tenant compte des mots composés.

Découpage en mots Les mots sont définis pour cette application comme des suites de caractères séparés par des blancs, l'apostrophe ou le tiret. Les

mots qui commencent par une capitale et qui suivent une ponctuation forte (point, trois points, point d'exclamation ou d'interrogation) sont recherchés tels dans le lexique. S'ils n'y appartiennent pas, ils sont donc potentiellement des mots communs marquant le début d'une phrase avec une capitale. Il sont alors recherchés après que la capitale ait été remplacée par une bas de casse[10]. La procédure est simple et permet de repérer les noms propres.

Les mots sont ensuite rassemblés en groupe de dix à deux mots pour vérifier si leur concaténation correspond à une forme du lexique. Ainsi le mot *aujourd'hui* a été décomposé en deux graphies *aujourd'* et *hui* puis la concaténation de ces deux parties a été identifiée comme correspondant à un mot. Cette méthode permet de repérer les mots composés mais donne une priorité systématique au premier mot le plus long dans la chaîne. Les mots qui sont ambigus entre formes simples et formes composés sont à tort toujours identifiés comme des mots composés. Notons que la structure de donnée permettant de rendre compte de toutes les ambiguïtés des mots composés est un graphe comme celui que nous présentons figure 4.3. Dans le premier schéma, il y a autant de façons de décomposer *fer à cheval* qu'il y a de chemins, dans le second schéma, chaque transition correspond soit aux mots simples *fer*, *à*, *cheval*, soit à un composant des mots composés *fer à cheval* et *à cheval*. Cette dernière représentation est celle recommandée par le consortium Genelex[Genelex 93, 1993].

4.3 Projection du lexique

Implémentation de l'arbre à lettres - Lexed

Le logiciel Lexed est une implémentation de la structure de donnée connue sous le nom de l'*arbre à lettres* ([Wehrli, 1997] p.156).

Cette structure permet de conserver un dictionnaire de façon suffisamment compacte pour qu'elle puisse tenir en mémoire d'un ordinateur, et est arrangée de manière à ce que la recherche d'une forme soit très rapide. Ce point est crucial, les dictionnaires en français peuvent contenir plus de 650 000 formes fléchies et le temps d'accès aux entrées correspondant à ces formes ne doit pas rendre rédhibitoire l'analyse de plusieurs millions de mots.

Le temps d'accès à une forme doit être proportionnel à une fonction de la longueur du mot à analyser et non à une fonction de la longueur du dic-

10. Minuscule d'imprimerie.

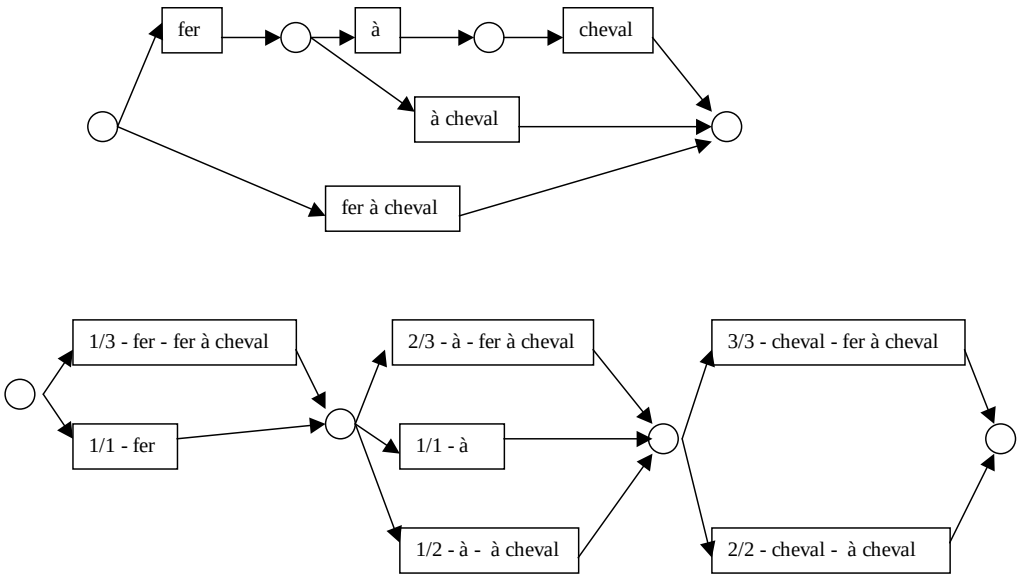


FIG. 4.3 – *Deux représentations possibles pour la suite "fer à cheval,,*

tionnaire.

Dans l'*arbre à lettres*, l'ensemble des formes d'un lexique est représenté par les chemins d'un automate fini déterministe dont chaque transition correspond à un caractère unique. Chaque sommet terminal de l'automate correspond à une forme et pointe sur une liste chaînée d'entrées. Nous illustrons fig. 4.4 un tel automate pour une liste très courte de mots.

Cette représentation a le double avantage de faire l'économie des nombreux préfixes communs aux formes d'un même lexique, et d'accéder à l'entrée d'un dictionnaire dans un temps proportionnel à la longueur du mot quel que soit la grosseur du lexique. Au pire des cas, la complexité en temps du programme est en $O(nm)$ où n est la longueur du mot et m est le nombre maximum de transitions partant d'un même sommet.

Pour optimiser le programme, il est possible de modifier l'ordre des transitions partant d'un même sommet en fonction de la probabilité conditionnelle de cooccurrence des caractères. Nous nous sommes contenté pour l'heure de construire l'arbre en ordonnant les formes selon leur fréquence dans un corpus de référence. L'ordonnancement des transitions n'est pas optimal mais

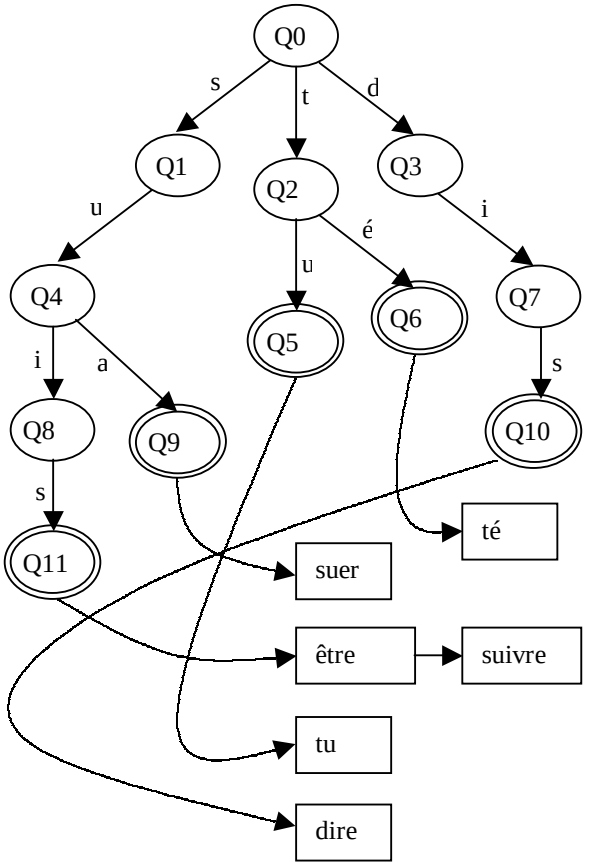


FIG. 4.4 – *Extrait d'Automate Fini Déterministe d'un «arbre à lettres»*

le nombre de tests de comparaison entre le caractère courant et l'ensemble des transitions à chaque sommet est ainsi réduit dans la grande majorité des cas. Et le logiciel donne des résultats largement suffisants dans la pratique.

Les formes qui ont en commun les mêmes préfixes partagent les mêmes chemins, il est également aisé d'économiser les suffixes communs à de nombreuses formes du français en retournant chaque mot et en les parcourant de droite à gauche.

Fonctionnement du logiciel

Construction de l'arbre Le logiciel lit un fichier qui contient sur chaque ligne une vignette (la forme fléchie) suivie d'une entrée de dictionnaire. Les formes homographes sont écrites sur plusieurs lignes.

L'arbre, initialement réduit à un seul sommet est construit en parcourant la vignette caractère par caractère. Chaque sommet donne lieu à une nouvelle transition étiqueté par le caractère courant si celle-ci n'existait pas déjà. La transition correspondant au dernier caractère de la vignette pointe sur un sommet marqué comme terminal.

Chacun des sommets terminaux pointe sur une liste chaînée dont les membres pointent à son tour sur un index des entrées du dictionnaire. Les entrées redondantes dans le dictionnaire (comme les étiquettes morpho-syntaxiques) ne sont évidemment pas dupliquées.

Le dictionnaire est conservé dans deux fichiers. L'un encode l'arbre, l'autre les entrées.

Parcours de l'arbre Dans un premier temps, les deux fichiers sont chargés en mémoire. Lorsque les entrées sont très volumineuses, seul le fichier correspondant à l'arbre est chargé en mémoire, l'accès aux entrées se fait alors en accédant directement aux disques durs.

Chaque forme est lue caractère par caractère pendant que l'arbre est parcouru depuis la racine vers les feuilles. Si le dernier sommet atteint est terminal, le logiciel parcourt la liste chaînée correspondante et les entrées sont directement pointées dans le fichier ou dans une table en mémoire.

Lexed fournit en résultat la ou les entrées de la vignette.

Communication TCP/IP Lexed permet de fournir les entrées d'un dictionnaire à un autre logiciel grâce à une connexion TCP/IP. Lexed est configuré dans ce cas comme un «serveur» qui attend la requête d'un «client» pour lui transmettre les entrées demandées. Le client peut être un logiciel quelconque sur une machine quelconque du réseau. Nous utilisons actuellement cette possibilité pour différentes raisons :

- Le chargement en mémoire de l'arbre correspondant à un dictionnaire volumineux prend quelques secondes. Cela est trop long dans certains cas. La mise à disposition d'un serveur permet de ne pas procéder régulièrement à ce chargement.
- Le serveur peut être lancé depuis n'importe quelle machine d'un réseau Intranet ou du réseau Internet, Cela permet de mettre à disposition

d'une communauté un dictionnaire électronique. Nous l'utilisons actuellement comme serveur de lexique de l'analyseur XLFG[11].
- La machine fournissant les entrées d'un dictionnaire très volumineux est très affaiblie en capacité de calcul et en capacité mémoire. Il est pratique d'avoir accès à une machine cliente non utilisée par ailleurs pour exploiter ces entrées.

Utilisation de Lexed dans le projet Lexed fourni le gros avantage de n'être pas un logiciel dédié à une tâche particulière du traitement automatique des langues. Nous l'utilisons pour trouver rapidement l'entrée d'un dictionnaire dont les vignettes et entrées sont indifférentes.

Enrichissement des étiquettes Lexed est utilisé à chaque fois que nous avons besoin de projeter un lexique sur un corpus étiqueté. C'est de cette façon que nous enrichissons les étiquettes morpho-syntaxiques avec toutes les informations non ambiguës en lexique.

Un dictionnaire a été constitué associant chaque couple (forme fléchie, étiquette morpho-syntaxique) avec :

- le lemme
- la sous-catégorie grammaticale

Il a suffit de croiser le corpus étiqueté avec ce dictionnaire grâce à Lexed pour obtenir ce même corpus lemmatisé et enrichi des informations de sous-catégorisation.

Lexique des nombres L'étiquetage automatique ainsi que l'enrichissement automatique des étiquettes doit prévoir un traitement morphologique distingué pour les nombres cardinaux et ordinaux écrits en chiffres ou en lettres. Il est en effet assez simple de construire un analyseur reconnaissant ces formes qui se déclinent à l'infini mais qui ne peuvent apparaître, pour cause, dans un dictionnaire fini.

L'étiqueteur de Paris 7 ne contient pas un tel module, il s'est donc agi de construire un lexique de ces formes en consultant celles qui sont effectivement présentes dans le corpus.

11. XLFG est un analyseur syntaxique pour grammaires LFG que nous avons développé. Nous n'en parlerons pas ici.

Nous avons écrit un petit programme qui permet de générer exhaustivement cardinaux et ordinaux du français en lettres et en chiffres (jusqu'aux billions). Nous avons naturellement croisé ces formes générées automatiquement avec celles qui sont effectivement présentes dans la corpus.

Les formes million, milliard, billion, trillion, etc., sont analysées comme des noms. Les mots composés cardinaux et ordinaux sont donc limités de 1 à un million selon nos conventions. Nous avons donc construit un lexique de ces formes cardinales et ordinales en faisant varier le genre et le nombre.

Ce gros lexique (6 millions de formes de plus de 40 caractères en moyenne) a été croisé grâce à Lexed avec le corpus pour n'en conserver que les formes présentes.

4.4 Les outils d'interrogation

Les feuilles de style XSL dont il a été question au chapitre précédent ne sont certainement pas suffisantes pour extraire les informations correspondant à des requêtes sophistiquées que peuvent réclamer les études de corpus. Il est en effet impossible d'exprimer un langage rationnel en toute généralité grâce à une feuille de style XSL bien que celles-ci puissent être utilisées pour filtrer un document XML. Or l'exploitation du corpus suppose d'extraire des syntagmes et des figures diverses qui ne sont pas toujours exprimables pas des simples suites de mots mais bien par des grammaires rationnelles. L'extraction des formes clivées en *qui* par exemple comme (*C'est Jean qui parle maintenant*) suppose que l'on identifie une forme *c'est* suivie d'un groupe nominal dont la longueur est indéterminée, modifié par une relative en *qui*. En l'absence d'annotation syntaxique du corpus, il est impossible d'identifier de telles constructions sans utiliser une grammaire rationnelle qui pourrait être représentée par cette expression :

`c'` ⟨ `verbe être` ⟩ `.*` ⟨ `Nom commun` ⟩ `.*` ⟨ `pronom relatif qui` ⟩

Où:

- ⟨ `être` ⟩ désigne tous les mots dont le lemme est *être* et la catégorie *verbe*.
- `.*` désigne toute suite de mots[12] (y compris le mot vide ∅.) C'est l'opération d'étoile de Kleene sur tout mot (désigné par "." comme

12. *Mot* est ambigu entre le terme utilisé pour désigner les éléments d'un monoïde muni de l'opération de concaténation et le terme linguistique de segmentation. Comme l'objet est le même dans ce qui suit, nous ne chercherons pas à expliciter l'acception.

le veut l'usage.)

– ⟨ `nom commun` ⟩ et ⟨ `pronom relatif qui` ⟩ désignent respectivement tout mot dont la catégorie est *nom commun* et *pronom relatif*. De plus, la forme du relatif doit être *qui*.

Le langage de requête devra permettre d'extraire des mots ou des figures pour en étudier le paradigme ou le contexte. Notons que l'unité de segmentation que nous avons privilégiés pour le corpus de Paris 7 est le mot. La première tâche que doit accomplir l'outil d'interrogation est donc le filtrage de ces mots à partir d'expressions régulières pour décrire les figures recherchées. L'exploitation des résultats pourra mettre en œuvre d'autres mécanismes comme le tri des contextes, le tri des occurrences trouvées ou le calcul des fréquences des éléments trouvés en fonction des éléments apparaissant dans les mêmes contextes.

Un tel outil pour les données XML n'étant pas disponible à notre connaissance, nous avons développé un langage de requêtes (*Cluster*) et un interpréteur de ce langage permettant de dresser l'automate fini déterministe correspondant, c'est-à-dire un *reconnaisseur* pour un tel langage.

L'interpréteur développé permet de baliser les suites textuelles correspondant aux expressions régulières sur les mots. Par souci d'optimisation du programme, nous avons restreint l'annotation SGML du document analysé par *Cluster*. Ces restrictions sont les suivantes :

– Tout élément SGML correspondant à un "mot" de l'expression régulière se trouve sur une seule ligne.
– Les attributs d'un même élément sont ordonnés.

Conscient du manque de robustesse de notre programme, nous avons développé un petit outil exploitant les bibliothèques d'analyseurs XML permettant de mettre un corpus annoté en XML dans ce format. Il va sans dire que les versions futures du programme *Cluster* exploiteront ces même bibliothèques pour accepter tout document XML bien formé.

4.4.1 Les outils d'interrogation existants

Certains sont utilisables pour plusieurs types de corpus, d'autres sont spécialisés pour un corpus donné.

Le corpus *Penn Treebank* est accompagné par exemple, d'un outil (*tgrep*) permettant de faire de simples recherches dans la structure arborescente des

descriptions syntagmatiques. Cet outil est un filtre permettant de trouver l'ensemble des descriptions syntagmatiques correspondant à une configuration syntaxique dans le *Penn Treebank*. Citons par ailleurs les programmes SgmlQL ([Maître *et al.*, 1998]) et XML-QL ([Deutsh *et al.*, 1999]) qui permettent de faire des recherches dans les documents en SGML et XML de nombreux types de corpus annotés.

INTEX

Le système INTEX ([Silberztein, 1993], [Silberztein, 1996]) permet d'interroger un corpus en fonction d'expressions régulières. Pour cela, l'utilisateur produit la grammaire de Kleene en dessinant un réseau d'automates non récursif à l'aide d'un éditeur de graphes. Intex permet également d'exploiter les dictionnaires DELA pour spécifier chaque transition. Plus précisément, l'utilisateur construit des transducteurs, c'est-à-dire qu'il définit quelle étiquette assigner à chaque transition et cela permet d'annoter un corpus à l'aide d'étiquettes correspondant à des grammaires de Kleene. Il est bien entendu d'usage de réutiliser les automates déjà dessinés comme simples éléments de transition.

Il est aussi possible d'extraire d'un corpus (étiqueté ou non) les configurations recherchées. INTEX fournit en résultat des concordances ou des listes de séquences reconnues.

XKWIC

Dans [Christ, 1994], Oliver Christ présente un outil d'interrogation de corpus annoté (*xkwic*). Cet outil permet d'extraire des passages du corpus en fonction de requêtes exprimées grâce à un langage exprimant des expressions régulières sur les formes et les catégories. Par exemple, la requête

[pos="JJ.*"] [pos="N.*"] "and|or" [pos="N.*"] [pos="IN" & word ! = "that"]

décrit une configuration correspondant à la séquence d'un adjectif (JJ, JJR, JJS), d'un nom (NN, NNS), d'une conjonction *and* ou or, d'un nom et enfin d'une conjonction de subordination (IN) qui ne doit pas être *that*.

Remarquons que dans ce langage, il est possible de décrire une configuration par une expression régulière qui ne doit pas correspondre au mot extrait. Il s'agit donc d'un ensemble de règles plutôt qu'une expression régulière complexe.

Le langage d'interrogation permet l'usage de variables (pour désigner l'accord flexionnel entre deux mots par exemple) et permet de décrire des configurations où apparaissent des enchâssements syntagmatiques.

Pour optimiser l'algorithme de recherche, le corpus est représenté de façon interne avec une structure de données adaptée à l'algorithme de recherche. Le corpus est indexé en plusieurs fichiers encodant différentes propriétés de chaque mot comme la position dans le texte, le constituant qui le contient et son enchâssement, etc. Cette organisation permet de trouver très rapidement les extraits recherchés et donne une certaine souplesse à l'outil d'interrogation (comme par exemple la possibilité de chercher des extraits de corpus parallèles, de chercher des mots qui ne sont pas filtrés par des expressions régulières, etc.)

Les résultats sont présentés par deux outils. Le premier construit les concordances de la séquence trouvée, le second présente deux séquences parallèle d'une même portion d'un corpus aligné.

L'outil d'interrogation développé par Laura Kallmeyer

Laura Kallmeyer ([Kallmeyer, 2000]) présente un outil de requête des corpus syntaxiquement annotés qui permet de tenir compte des relations de dominance et précédence entre éléments. De plus, cet outil permet d'utiliser plusieurs types de corpus, annotant les relations de coindexation (Penn Treebank) ou les constituants croisés (Negra Corpus).

Laura Kallmeyer utilise une base de donnée relationnelle pour encoder l'ensemble des informations qui seront exploitées par une requête SQL. L'outil se compose d'un module de construction de la base de donnée utilisé par l'administrateur, et d'un module de production de requêtes SQL à partir des interrogations.

Pour ce dernier module, un langage spécifique à été développé pour exprimer les relations de précédence (opérateur `..`), les relations de dominances (opérateurs `>` et `>>`) et les formes (`token()`), catégories (`cat()`) et fonctions (`fct()`) des différents termes. Le module d'interrogation, écrit en java, est un compilateur qui produit automatiquement une requête SQL à partir d'une expression bien formée dans le langage défini par l'auteur. Cette méthode permet de tirer parti de la qualité des outils de gestion de base de données relationnelles, de leur maintenance, de leur fiabilité tout en proposant à l'utilisateur final un langage propre à exprimer les requêtes sur corpus.

4.4.2 Le programme «Cluster»

Le fonctionnement de Cluster est très proche du mécanisme de construction et de parcours d'automates d'INTEX. La différence essentielle est, outre les choix informatiques et les algorithmes employés, ergonomique : les grammaires de Kleene sont explicitées par un langage avec *Cluster* et dessinées grâce à un éditeur de graphes avec *Intex*. *Cluster* souffre des même limitations, à savoir qu'il ne permet que l'analyse des grammaires de Kleene et ne permet pas de faire des recherches sur des mots qui ne sont pas reconnus par des expressions régulières comme c'est le cas de l'outil d'Oliver Christ *XKWIC*. Nous réservons cependant ce programme à être une maquette destinée à l'étude de l'ergonomie du langage. L'étude de l'optimisation de *Cluster* passera par une étude sur les structures de données. Pour l'instant, Cluster exploite sans modification majeure des documents en XML.

Le programme Cluster prend en entrée un fichier où ont été décrits les syntagmes réguliers avec le langage décrit ci-après[13].

Syntaxe Ce langage permet d'écrire des expressions régulières où chaque "**mot**" correspond à une balise XML et possède la syntaxe suivante (Nous notons entre parenthèses les éléments optionnels) :

$B(= RegExp)([arg_1 = val_1, arg_2 = val_2, \ldots, arg_n = val_n])$

où

- B désigne le nom d'une balise XML.
- $RegExp$ désigne le contenu de la balise. C'est une expression régulière sur les lettres.
- arg_i désigne le i^{e} attribut de la balise XML.
- val_i désigne la valeur du i^{e} attribut de la balise XML. C'est une expression régulière sur les lettres.

L'ensemble des expressions régulières est le plus petit ensemble tel que :

- Si a est un "**mot**", a est une expression régulière
- Si A et B sont des expressions régulières, alors (AB) est une expression régulière.
- Si A et B sont des expressions régulières, alors $(A|B)$ est une expression régulière.

13. Nous appellerons du même nom le compilateur et le langage comme c'est l'usage en compilation.

– Si A est une expression régulière, alors $A*$ est une expression régulière.

Sémantique

– a désigne le langage $\{a\}$.
– (AB) désigne le langage construit par concaténation des mots du langage A avec les mots du langage B.
– $(A|B)$ désigne l'union du langage A et du langage B.
– $A*$ désigne le langage $\bigcup_{i\in[0,\infty]} A^i$ (où $A^i = \{\overbrace{AAA\ldots A}^{i}\}$, et $A^0 = \{\emptyset\}$.)

Nous avons ajouté d'autres opérateurs permettant de simplifier les expressions régulières comme c'est désormais l'usage.[14]

– $A+$ pour $A(A*)$
– $A\{j,k\}$ pour $(A^j|A^{j+1}|A^{j+2}|\ldots|A^k)$ $(j \leq k)$
– $A?$ pour $(A|\emptyset)$

L'expression régulière qui correspondra à notre exemple des formes clivées sera en langage *Cluster* :

```
w="(c|c')" w[lemma="être", cat="V"] w* w[cat="N",
subcat="com"] w* w="qui"[cat="PRO", subcat="rel"]
```

Les expressions régulières sont rassemblées et séparées par des points-virgules comme autant de disjonctions d'expressions régulières qui concernent le même type d'annotation.

Fonctionnement Nous savons que les expressions régulières correspondent à des automates finis déterministes. Nous n'avons pas réécrit les fonctions permettant de générer ces automates car cela existe déjà dans les ressources informatiques que nous utilisons. Mais la construction d'un automate déterministe peut-être très coûteuse puisqu'elle fait intervenir un algorithme dont la complexité en temps est exponentielle. En effet, pour créer un automate fini déterministe à partir d'un automate fini non déterministe (obtenu presque trivialement à partir d'une expression régulière) ou à partir d'expressions régulières, il faut construire l'ensemble des parties d'un ensemble. Opération nécessairement exponentielle.

14. Nous avons également donné priorité de l'opération de disjonction sur l'opération de concaténation, le parenthèsage n'est donc pas toujours nécessaire. De plus nous avons $(A_1A_2\ldots A_k)$ pour $(A(A_2(\ldots A_k)\ldots)$ et $(A_1|A_2|\ldots|A_k)$ pour $(A|(A_2|(\ldots|A_k)\ldots)$.

Ainsi, même en utilisant les ressources existantes (le programme *Lex*, les bibliothèques d'expressions régulières disponibles pour différents langages, etc.) la compilation des règles peut mettre plusieurs heures, voire plusieurs jours sur une machine puissante [15] sans optimiser les expressions régulières.

Nous n'avons pas choisi de restreindre les expressions régulières pour ne pas limiter la puissance d'expressivité du langage *Cluster*. Alors, pour rendre opérante la construction de l'automate, nous avons également choisi de construire un automate fini non déterministe comme le proposent certaines fonctionnalités de la bibliothèque standard *regexp* du langage C quand les expressions régulières sont très grosses.

Automate fini déterministe Nous avons utilisé le programme *Flex* de GNU distribué sur tout système Unix. Ce programme est conçu pour générer automatiquement un analyseur sur des expressions régulières. Le programme *Cluster* fonctionne alors comme un *compilateur de compilateur* : il traduit un fichier *Cluster* en un fichier *Flex* puis appelle les compilateurs *Flex* et *CC* pour produire un fichier exécutable. Cette méthode est séduisante puisqu'elle est très simple et produit un programme qui bénéficie des optimisations du logiciel *Flex*.

Automate fini non déterministe Dans ce cas, nous avons utilisé la bibliothèque *Regexp* pour ne pas réécrire l'algorithme de construction d'un automate non déterministe à partir d'une expression régulière. L'automate est alors représenté par une structure de donnée distribuée avec le projet *Regexp* et permet d'analyser un document avec des expressions régulières extrêmement complexes. Dans ce dernier cas, l'analyse est plus longue puisqu'elle est issue du parcours en profondeur d'un automate fini non déterministe, mais la construction des sous-ensembles n'est pas effectuée. La compilation est donc immédiate et l'analyse relativement longue.

Dans les deux cas, le programme, écrit en C, construit une représentation interne des expressions régulières pour vérifier leur bonne formation et pour générer l'expression régulière dans le format *Flex* ou *Regexp*.

Dans les prochaines versions de *Clusters*, nous utiliserons la bibliothèque *Expat* qui permet l'analyse syntaxique des documents XML avec toutes les garanties que nous attendons : portabilité, respect des conventions XML (en-

15. Plus de 48 heures ont été nécessaires pour compiler un fichier *Flex* issu d'un fichier *Cluster* sur une machine Ultra Sparc. Le programme C résultant a été compilé sans aucune optimisation en plusieurs heures sur la même machine !

codage des caractères en unicode, ISO-8859, etc.), fiabilité, robustesse et rapidité.

Pour l'heure, le programme *Cluster* permet de faire une analyse d'expressions régulières sur des documents XML adaptés avec une rapidité et une fiabilité suffisante.

Exemples d'interrogation du corpus grâce à *Cluster*

Cet exemple est destiné à extraire du corpus tous les clitiques qui se trouvent entre le verbe *faire* et un verbe à l'infinitif.

La requête est la suivante :

```
id {
w=[lemma="faire", cat="V"] w[cat="CL"] w[cat="V", mph="W"]
}
```

Cette requête définit une séquence de trois mots, le premier est un verbe (de catégorie V) et a comme lemme *faire*, le second est un clitique et le troisième est un verbe dont la morphologie est W, c'est-à-dire un infinitif.

La compilation de la requête met 1,0 seconde (tous les essais ont été effectués sur une UltraSpark 30 et sur 500 000 mots), l'exécution du code (c'est-à-dire l'application de l'automate déterministe) sur 500 000 mots prend 29,40 secondes, et l'interprétation (le parcours de l'automate non déterministe) 59,95 secondes.

Le résultat est le marquage par une balise `<id>` des extraits du corpus qui correspondent à cette configuration. En voici un extrait :

```
<w lemma="cependant" cat="ADV">cependant</w> <w lemma="," cat="PONCT" subcat="W">,</w> <w lemma="le" cat="D" subcat="def" mph="fs">la</w> <w lemma="surenchère" cat="N" subcat="C" mph="fs">surenchère</w> <w lemma="vert" cat="A" subcat="qual" mph="fs">verte</w> <w lemma="aller" cat="V" subcat="" mph="P3s">va</w> <id ><w lemma="faire" cat="V" subcat="" mph="W">faire</w> <w lemma="se" cat="CL" subcat="refl" mph="3mp">s'</w> <w lemma="affronter" cat="V" subcat="" mph="W">affronter</w> </id><w lemma="un" cat="D" subcat="ind" mph="mp">des </w> <w lemma="groupement" cat="N" subcat="C" mph="mp">groupements</w> <w lemma="professionnel" cat="A" subcat="qual" mph="mp">professionnels</w> <w lemma="concurrent" cat="A" subcat="qual" mph="mp">concurrents</w> <w lemma="." cat="PONCT" subcat="S">.</w>
<w lemma="de" cat="P">de</w> <w lemma="le" cat="D" subcat="def" mph="ms">l'</w> <w lemma="Est" cat="N" subcat="P" mph="ms">est</w> <w lemma="de" cat="P">de</w> <w lemma="le" cat="D" subcat="def" mph="fs">l'</w> <w lemma="Europe" cat="N" subcat="P" mph="fs">Europe</w>
<id ><w lemma="faire" cat="V" subcat="" mph="P3s">fait</w> <w lemma="se" cat="CL" subcat="refl" mph="3mp">se</w> <w lemma="tourner" cat="V" subcat="" mph="W">tourner</w> </id> <w lemma="tant" cat="ADV">tant</w> <w lemma="de" cat="P">de</w> <w lemma="regard" cat="N" subcat="C" mph="mp">regards</w> <w lemma="vers" cat="P">vers</w> <w lemma="celui" cat="PRO" subcat="dem" mph="3mp">ceux</w> <w lemma="qui" cat="PRO" subcat="rel" mph="3mp">qui</w>
<w lemma="son" cat="D" subcat="poss" mph="3fps">Ses</w> <w lemma="onomatopée" cat="N" subcat="C" mph="fp">onomatopées</w>
<id ><w lemma="faire" cat="V" subcat="" mph="P3p">font</w> <w lemma="se" cat="CL" subcat="refl" mph="3ms">se</w> <w lemma="retourner" cat="V" subcat="" mph="W">retourner</w> </id>
<w lemma="le" cat="D" subcat="def" mph="ms">l'</w> <w lemma="oiseau" cat="N" subcat="C" mph="ms">oiseau</w> <w lemma="." cat="PONCT" subcat="S">.</w>
```

Bien évidemment, un tel format est inexploitable pour un utilisateur final, nous verrons *infra* un outil permettant d'interfacer Cluster pour l'utilisateur.

Les résultats de cette requête font apparaître les clitiques réfléchis *se* (*s'affronter*, *se tourner*, *se retourner*, *s'interroger*, etc.) et dans une moindre proportion des clitiques sujet inversés, *faites-vous*, *faisait-on*, etc.

Le deuxième exemple d'interrogation que nous proposons permet d'extraire toutes les suites de plus de deux verbes consécutifs.

Nous pouvons écrire cette requête de plusieurs façons. Un élément `X` répété plus d'une fois peut s'écrire : `X X X*` ou `X X+` ou encore `X{2,}`.

La requête suivante est compilée en 2 secondes (51,59s), l'exécution a lieu en 30,4 secondes et l'interprétation en 33 minutes 47 secondes !

```
   id {
w=[cat="V"]{2,}
}
```

Cette requête permet de dresser quelques statistiques intéressantes sur les suites de verbes. 87,2% de ces suites sont formés de deux mots, 12,2% de trois, 0.4% de quatre et il existe même 6 suites de 5 verbes.

Deux de ces suites sont des noyaux verbaux (*ont affirmé avoir été torturées*, *a dû démentir avoir ordonné*), les autres font intervenir une relative immédiatement suivi du noyau verbal (*les fonctionnaires dont les postes avaient été supprimés ont trouvé*, *qui paie peut prétendre être indemnisé*).

4.4.3 Interface du concordancier

Nous venons de voir que l'usage de *Cluster* était réservé à un utilisateur habitué à la syntaxe des expressions régulières et que les résultats qu'il fournit ne peuvent pas être exploités immédiatement. L'outil doit donc s'accompagner d'une interface pour être exploitable.

L'usage du langage *Cluster* ou d'un langage équivalent ne correspond pas aux requêtes sur un corpus telles qu'on pourrait les exprimer. Nous verrons dans le prochain chapitre ce qu'il peut en revanche apporter pour l'enrichissement des annotations. Il suppose que l'on connaisse les expressions régulières et demande une formalisation de la requête. Les linguistes et autres usagers peuvent réclamer par exemple "*tous les noms communs qui suivent immédiatement un autre nom*" ou encore "*toutes les formes qui apparaissent entre un déterminant et un nom*" sans savoir ou vouloir exprimer ces requêtes par les expressions correspondantes. Bref, nous avons dû développer une *interface* entre le programme *Cluster* et l'usager. De plus,

l'exploitation du corpus ne se satisfait pas du marquage des suites reconnues par les expressions régulières mais suppose également le tri et le filtrage des résultats : soit des occurrences trouvées, soit de leurs contextes.

Nous avons développé un **concordancier** mis à la disposition des linguistes exploitant le corpus de Paris 7 sur Internet (`http://talana.linguist.jussieu.fr/~lionel/demo-concordancier.hlml` [Clément & Kinyon, 2000]). Nous le présentons comme une maquette permettant d'étudier l'ergonomie d'un tel logiciel et le génie logiciel s'y rapportant. Il est donc appelé à évoluer et à être un modèle de ses successeurs, notamment d'un outils d'exploitation du corpus annoté pour les constituances et dépendances syntaxiques. L'outil permet néanmoins d'obtenir aujourd'hui des résultats qui intéressent des études linguistiques comme celles que nous présentons à la fin de ce chapitre.

Il permet d'interroger le corpus étiqueté de Paris 7 pour obtenir deux sortes de résultats : les concordances d'un terme et la fréquence d'un terme. Par terme, nous entendons un mot, une suite de mots ou un paradigme décrit par les propriétés de la forme graphique mais également du lemme, de la catégorie, de la sous-catégorie et de la morphologie.

Il est également possible de faire des recherches affinées non plus sur l'ensemble du corpus mais sur les résultats déjà recueillis lors d'une précédente requête. L'utilisateur peut également faire des recherches sur une partie du corpus et non sur sa totalité.

Le concordancier est actuellement une page Internet qui appelle un script *CGI* générant automatiquement un document en langage *Cluster*.

Le script *CGI* a été écrit grâce à un langage interprété de commandes Unix (*Shell*) et des programmes écrits en C. Ils comprennent un ensemble de procédures liées à la sécurisation de la distribution du corpus (vérification de l'origine des requêtes, de l'emplacement de la page Internet, envoi de messages électroniques à la personne chargée de maintenir le corpus, mise à jour de fichiers, etc.)

Le programme *CGI* produit un fichier *Cluster* correspondant à la requête de l'utilisateur et le compile (ou l'interprète) pour annoter dans la partie du corpus sélectionnée les suites textuelles recherchées.

Un deuxième fichier *Cluster* est alors compilé pour annoter les contextes de ces suites textuelles. Le contexte s'exprime par un certain nombre de mots à gauche et à droite de la suite textuelle correspondant à la requête. On voit qu'il sera possible d'affiner la nature du contexte dans la suite.

Ensuite, si cela à été demandé, le programme charge le résultat dans un *panier* unique pour chaque utilisateur, lequel *panier* pourra être exploité pour affiner une recherche ultérieure.

Les concordances sont présentées comme traditionnellement en trois colonnes contenant respectivement le contexte gauche, l'occurrence trouvée et le contexte droit. Différents tris sont proposés pour focaliser l'étude sur les contextes ou sur les occurrences. Les résultats affichés peuvent être indifféremment la forme, le lemme, ou la catégorie.

Les résultats sont affichés avec ces différents éléments d'information :

- affichage des lemmes et/ou catégories et/ou formes,
- tri de droite à gauche du contexte gauche,
- tri de gauche à droite du contexte droit,
- tri des occurrences trouvées,
- affichage de la fréquence de l'occurrence,
- affichage du pourcentage de l'occurrence relativement à toutes les occurrences correspondant à la même requête.

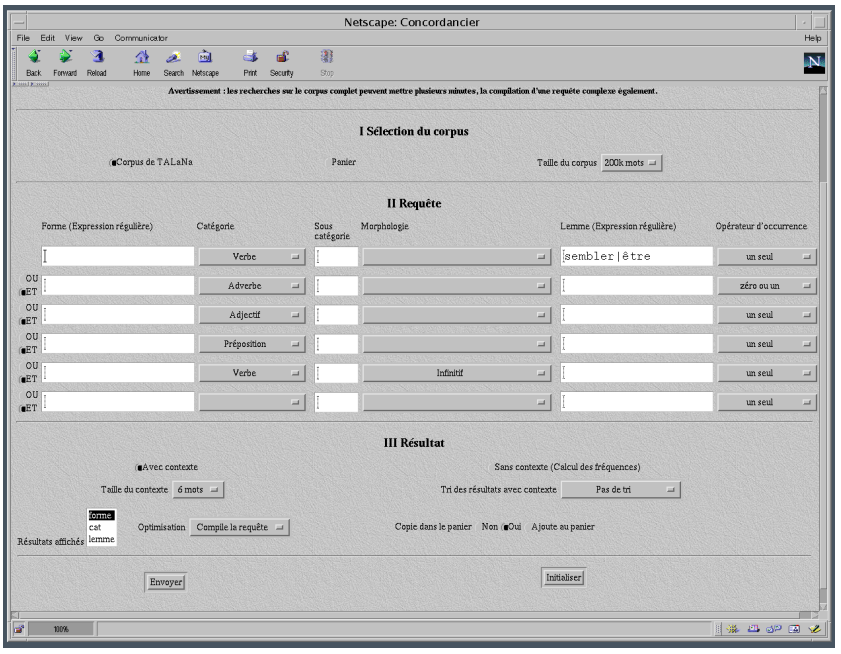


FIG. 4.5 – *Capture d'écran de la page Internet du concordancier — Requête*

Nous présentons figure 4.5 une impression d'écran du concordancier tel qu'il se présente à l'utilisateur avec une requête simple visant à dresser la

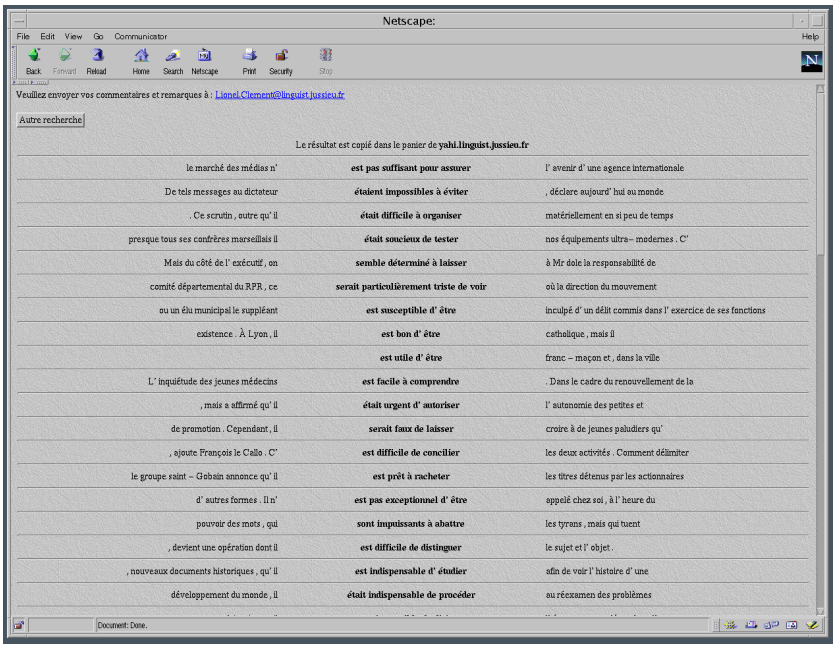


FIG. 4.6 – *Capture d'écran de la page Internet du concordancier — Résultat de la requête*

liste des adjectifs qui apparaissent dans les figures du type *Jean est pénible à entendre*, *Ce résultat semble très difficile à croire*. Nous présentons en 4.6 une impression d'écran des résultats recueillis pour cette même requête. Dans cet exemple, nous avons donc cherché dans un premier temps les suites `Verbe être/sembler/trouver + adjectif + préposition + Verbe à l'infinitif`. Puis, nous avons interrogé ces résultats pour n'extraire que les adjectifs de ces figures afin de les trier par fréquence. Nous avons trouvé une courte liste d'adjectifs (*prêt, difficile, facile, impossible, etc.*) très représentés dans cette position comme illustré figure 4.7. Les dix adjectifs les plus fréquents dans cette position occupent à eux seuls 47% des usages :

1	prêt	14%
2	difficile	11%
3	impossible	7%
4	facile	5%
5	capable	4%
6	indispensable	4%
7	incapable	2%

Par ailleurs, hors ce contexte grammatical, les adjectifs mentionnés n'oc-

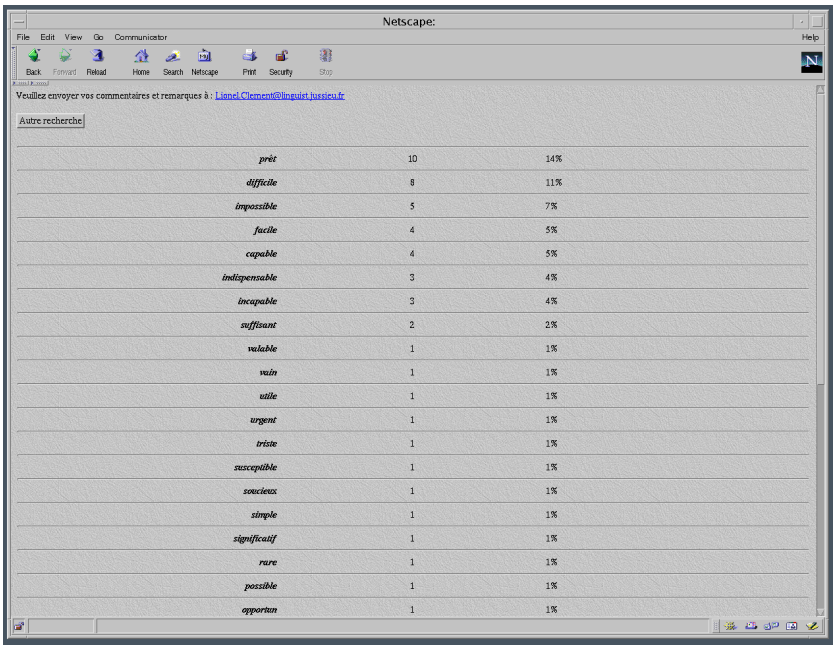


FIG. 4.7 – *Capture d'écran de la page Internet du concordancier — Résultat d'une recherche d'occurrences*

cupent pas la majorité des emplois. Il semble donc que la figure grammaticale s'accompagne d'une préférence lexicale pour quelques adjectifs, sans que l'on puisse parler de total figement lexical. Claire Blanche-Benveniste (in [Blanche-Benveniste, 1996]) parle de "degrés de grammaticalisation" pour exprimer le *continuum* entre une forme totalement figée et une forme acceptant des choix lexicaux non restreints. Nous n'étudierons pas davantage la restriction lexicale des adjectifs qui apparaissent dans cette position, cette démonstration a pour but de montrer que l'usage d'un concordancier permet d'obtenir des résultats que l'on ne peut aisément déduire de la seule intuition sur la langue. Ceci n'est possible que parce que le corpus est suffisamment volumineux, lemmatisé (nous avons cherché les lemmes des adjectifs et non leurs formes), et annoté pour la morpho-syntaxe.

Chapitre 5

Interrogation du corpus

Un corpus annoté doit permettre des réponses plus rapides et plus fiables à des questions linguistiques classiques (quels adjectifs se trouvent avant le nom en français? quel type de verbes trouve-t-on dans les relatives en *dont*? etc.). Il doit permettre également de poser des questions qu'on n'oserait pas aborder directement sur corpus brut (le schéma ADJ/N est-il plus fréquent que le schéma N/ADJ, dépend-il du N?, etc.) tant les dépouillements manuels seraient longs et laborieux. Le Corpus Annoté de Paris 7, associé à des outils tels que le concordancier que nous avons présenté en 4.4.3 nous permet de répondre simplement à ces questions.

5.1 Interrogations de type linguistique

Le nombre et la variété des interrogations sur le corpus annoté est bien entendu infini. Dans un premier temps, nous avons cherché à reproduire des types d'interrogation qui avaient été réalisé par des moyens artisanaux sur d'autres types de corpus (littérature ou français parlé) pour voir si nous obtenons le même type de résultat sur notre corpus de français contemporain journalistique. Les résultats sont bien sûr à prendre avec précaution puisque rien ne dit que notre corpus soit de taille suffisante pour être représentatif de l'étude, ni que le genre journalistique n'influence pas les résultats.

Dans un second temps nous avons mené quelques interrogations de type nouveau, plus quantitatif, susceptible d'intéresser les psycholinguistes comme les linguistes informaticiens.

5.1.1 Place de l'adjectif épithète

Nous appuierons notre étude sur l'article de Marc Wilmet [Wilmet, 1981]. Dans ce texte, Marc Wilmet entend décrire la place des adjectifs épithètes. Contrairement aux travaux de ses prédécesseurs, son étude sera validée par l'analyse d'un gros corpus. Pour cela, il a demandé à 80 étudiants de dépouiller systématiquement chacun les 50 premières pages d'une œuvre contemporaine laissée à leur discrétion et d'y recenser les adjectifs qualificatifs antéposés et postposés. Les données reposent donc sur un corpus de 4 000 pages. En estimant qu'une page de livre de poche convient environ 400 mots, le corpus étudié par Marc Wilmet devait contenir environ 1,6 million de mots.

Les résultats montrent, outre une préférence pour les postpositions (33,56% des formes sont préposées, 66,44% sont postposées), une préférence pour les épithètes les mieux représentées en position antéposée. Sur 183 formes dont la fréquence est la plus élevée, 53% sont antéposés et 47% sont postposés. Marc Wilmet remarque que cette observation est encore plus tranchée pour les six adjectifs les plus fréquents : *grand, petit, bon, jeune, beau* et *vieux*. Dans ce cas, 96% des formes sont antéposées. Inversement, les adjectifs les moins fréquents sont moins antéposés (11%. contre 89% de postpositions).

La conclusion provisoire que propose Marc Wilmet au vu de cet élément est que l'antéposition de l'épithète qualificative représente un ordre marqué réservé aux formes les plus représentées qui ont les deux positions.

Il semble que cette spécialisation d'une liste restreinte d'adjectifs l'emporte sur les propriétés propres à l'antéposition et à la postposition. Marc Wilmet le remarque pour les propriétés sémantiques et phoniques des adjectifs épithètes.

Le travail de Marc Wilmet a réclamé le concours important de 80 personnes et le dépouillement méthodique de 29 016 exemples. Rappelons que ce travail a été fait à la fin des années 70 et que l'outil informatique n'était pas disponible pour de tels travaux linguistiques. Voyons si l'exploitation informatisée du corpus de Paris 7 peut fournir un matériau d'étude comparable avec une économie de moyens évidente.

Nous avons interrogé le corpus de Paris 7 avec le concordancier pour obtenir les lemmes des adjectifs antéposés et postposés. Nous avons trouvé une plus grande proportion de qualificatifs dans les textes journalistiques.

Le matériau d'étude de Marc Wilmet est le texte littéraire (les étudiants ont très majoritairement choisi des romans), il est fort probable que le genre du texte ait une influence sur les résultats recueillis.

Sur 59 054 occurrences trouvées, 30 652 sont postposés et 12 556 sont antéposés. Nous retrouvons donc les chiffres de Wilmet : environ 52% de postpositions contre 21%.

La répartition des adjectifs postposés est assez plate, c'est-à-dire qu'elle ne contraste pas des formes très fréquentes avec des formes rarement représentées. Les formes les plus représentées en positions postposées ne représentent que 2% de leur total.

Voici la liste des épithètes qualificatives les plus fréquentes :

lemme	fréquence
français	776
économique	498
américain	459
européen	458
national	451
politique	300
social	299
allemand	294
public	278
financier	247
international	240
mondial	204
général	184
industriel	178
actuel	176
étranger	165
soviétique	163
commercial	163
japonais	161

La liste des adjectifs postposés ne provoque pas de commentaires tant la répartition des occurrences est plate. En revanche la liste des qualificatifs antéposés dont nous présentons les 8 plus fréquents correspond aux résultats de Marc Wilmet.

Adjectif	**fréquence**
grand	508
petit	227
bon	211
véritable	131
jeune	108
futur	79
long	72
mauvais	64

Nous voyons que l'exploitation du corpus de Paris 7 offre donc sensiblement le même matériau d'étude de l'antéposition des épithètes qualificatives malgré la différence de genre. Les résultats obtenus n'ont demandé que quelques minutes d'interrogation grâce au concordancier. Cela illustre l'intérêt de l'informatisation de l'exploitation des ressources linguistiques.

On s'attend également à ce que le concordancier puisse fournir avec plus de facilité des résultats sur quelques points plus précis. Si l'on s'intéresse par exemple à la liste des adjectifs épithètes les moins représentés en position antéposée et des noms qu'ils modifient et que l'on interroge le corpus avec cette seule requête, on obtient une liste dont voici un extrait :

- arrogante suffisance
- ardent passionné.
- alternant apprentissage.
- aléatoire procédure.
- affectueuse ironie.
- adroite comparaison.
- affreux mot.
- agréable pensée.
- actif allié.

Nous voyons que les adjectifs antéposés peu représentés n'ont pas le même comportement que ceux qui le sont largement. Les tournures exemplifiées correspondent à un registre de langue élevé et il serait intéressant d'interroger un corpus oral de conversation pour remarquer leur très probable absence alors que les adjectifs *grand, petit, bon, véritable, jeune* s'antéposent sans difficulté à l'oral courant.

Par ailleurs, certains des points caractérisant les épithètes qualificatives antéposées remarqués par Marc Wilmet portent sur ces occurrences sans toutefois intéresser les adjectifs très représentés qui occupent cette position.

Reprenons quelques points qu'a mis en évidence Marc Wilmet pour caractériser la place de l'épithète dans [Wilmet, 1981].

- "Affinité de l'antéposition avec la quantification".
  Ce qui du point de vue sémantique explique la "fusion intime du déterminant et du déterminé" (*un savant amoureux*=un expert en amour, *un amoureux savant*=un savant qui est amoureux), "la spécialisation de l'épithète" (métaphores, litotes, hyperboles, etc. : *un mortel ennui, un grand homme*) ou sa "neutralisation" (*une blanche colombe*).
  Cette propriété de l'antéposition s'opposerait avec le caractère de "complément déterminatif" de l'épithète postposée : Le substantif et son complément sont autonomes du point de vue sémantique et la suite Substantif Adjectif marque une relation "intellectuelle".
  Remarquons que ces propriétés semblent particulièrement exemplifiées par les épithètes antéposées peu représentées mais le sont beaucoup moins pour les adjectifs *grand, petit, bon, véritable* ou *jeune. une adroite comparaison* ne se compare pas avec *une grande comparaison* ; dans le deuxième exemple, le sens premier de l'adjectif est conservé et la sémantique du groupe nominal est compositionnel.
- Comportement "anormal" de l'antéposition. "Débarrassée de sa fonction sémantique primaire, l'antéposition laisse cours aux effets "par évocation" : ton poétique, calques stylistiques, emprunts, parodies (y compris les archaïsmes et les régionalismes délibérés" (Marc Wilmet [Wilmet, 1981] p. 60).
  Là encore, la courte liste des antépositions les plus fréquentes semble échapper à cette propriété stylistique évoquée par Mac Wilmet. L'emploi antéposé de ces adjectifs ne marque aucun comportement "anormal".

On voit que les données présentent un double comportement des épithètes qualificatives antéposées selon qu'elles appartiennent ou non à la liste restreinte des adjectifs les plus représentés dans cette position. On est amené à distinguer les antépositions des épithètes qualificatives marquant un comportement sémantique et stylistique des antépositions liées aux propriétés lexicales de quelques adjectifs. Quelques adjectifs sont marqués lexicalement ou non par la propriété d'antéposition.

Cette étude montre que l'emploi d'un concordancier permet de confronter une étude linguistique à une étude empirique sur corpus avec une économie de moyen évidente. Un concordancier permet également le dépouillement de données plus rares (comme ici les épithètes qualificatives antéposées peu

fréquentes) dont les propriétés sont déterminantes pour l'étude d'un phénomène grammatical.

5.1.2 Les pronoms relatifs

La hiérarchie des fonctions grammaticales

Il peut sembler ambitieux d'étudier la répartition des fonctions grammaticales dans un corpus seulement étiqueté au niveau des mots. Mais nous nous intéressons ici aux formes pronominales (pronoms relatifs et clitiques) qui sont morphologiquement différenciés selon la fonction.

Sur un corpus écrit de l'anglais, Keenan [Keenan & Hawkins, 1987] a étudié la fréquence des pronoms relatifs. Cette étude a eu pour but de mettre en correspondance la hiérarchie d'accessibilité universel, dont font l'hypothèse Keenan & Comrie entre les fonctions sujet, objet, objet indirect, oblique et la fréquence d'emploi des pronoms. Les pronoms relatifs sujets devraient être plus fréquemment employés que les pronoms objet directs et indirects eux même plus représentés dans les textes que les pronoms obliques.

Cette hypothèse repose sur le coût supposé plus important pour les réalisations non canoniques des fonctions obliques que des fonctions objet et sujet.

L'étude de Keenan porte sur l'anglais et les résultats qu'il affiche montre que la fréquence d'usage des pronoms relatifs suit l'ordre proposé :

suj ⟨ **obj** ⟨ **obj-ind** ⟨ **obl** :

Pronom relatif sujet	46%
Pronom relatif objet	24%
Pronom relatif objet indirect	15%
Pronom relatif génitif	5%
Divers	10%

Nous avons utilisé le concordancier pour faire une telle étude sur le corpus de Paris 7. Le corpus de Paris 7 n'ayant pas encore été annoté pour les fonctions grammaticales, nous devions expliciter le contexte des pronoms relatifs ambigus sur la fonction.

Les pronoms relatifs du français sont *qui, que, quoi, où, dont* et les formes composées avec *lequel* (*auquel, duquel, à laquelle*, etc.).

– Le pronom relatif sujet est marqué par la forme *qui* non précédée d'une

préposition. Le pronom relatif objet est *que*

- Le pronom relatif objet indirect est *qui* pour les animés et *quoi* pour les inanimés précédé d'une préposition, la forme *dont* ou l'une des formes composées avec *lequel* (*auquel*, *duquel*, etc.).
- Le pronom relatif *dont* marque le génitif.
- Enfin la forme *où* marque le locatif.

Sur 10 304 pronoms relatifs recensés, nous avons extrait cette répartition :

Fonction	description	effectif	fréquence
Sujet	*qui* non précédé de préposition	6 291	61%
Objet direct	*que*	1 565	15,2%
Génitif	*dont*	1 076	10,4%
Locatif	*où*	782	7,6%
Objet indirect	*qui, quoi ou lequel* précédé d'une préposition	539	5,2%
divers			0,3%

Nous retrouvons en partie les résultats de Keenan sur l'anglais, en particulier, la fonction sujet est très majoritairement représentée puisque le relatif sujet dépasse 60% des emplois des relatifs. Comme l'étude portant sur l'anglais, vient ensuite le relatif objet. Le génitif, le locatif et les relatifs objet indirect sont moins représentés et ne suivent pas l'ordre de préférence observé par Keenan.

Dont est peut-être surestimé car on ne peut faire la différence à ce stade entre le complément de verbe (i.e. (d)), le complément de nom (i.e. (c)) ou le *dont* parenthétique (i.e. (a), (b)).

(a) ... les anciens cadres du PC - dont d'anciens vainqueurs de la bataille de Saigon en 1975 - qui ont dénoncé ...

(b) une soixantaine de rebelles présumés, dont certains étaient armés, ont été transférés à Abidjan

(c) une opération dont il est difficile de distinguer le sujet et l'objet.

(d) ...chercheurs dont l' université aura massivement besoin.

Le génitif est très fréquent relativement aux autres formes obliques. Cela est certainement dû à la nature même des textes étudiés. Nous verrons dans la section suivante que les différents genres littéraires et de corpus oraux peuvent influencer la fréquence d'emploi d'une forme grammaticale.

Emplois stéréotypés du relatif dont

[Blanche-Benveniste, 1996] s'intéresse à l'emploi supposé stéréotypé des relatifs en français oral de conversation. Cet emploi contraste les " genres " : l'emploi du relatif *dont* s'accompagne d'une " collocation " avec une courte liste de verbes à l'oral de conversation alors que la liste semble plus ouverte pour les autres genres.

D'après un comptage sur le corpus de l'oral retranscrit de français constitué par l'équipe du GARS (Groupe Aixois de Recherche Syntaxique), près de la moitié des emplois de *dont* sont occupés par le verbe *parler*. Huit verbes occupent plus de 90% des emplois :

parler	48,4%
avoir besoin	12,9%
faire partie	8,1%
prendre conscience	4,8%
sortir	3,2%
dépendre	3,2%
être question	3,2%
être convaincu	3,2%

Ce résultat est comparé avec le texte écrit extrait d'un exemplaire du journal *Le Monde*. Dans cette dernière ressource, Claire Blanche-Benveniste note un moins grand figement, on trouve une majorité d'emplois avec le verbe *parler* mais "La liberté grammaticale de *dont* est indéniablement plus grande dans la presse écrite que dans les conversations." (ibid.).

parler	8,9%
faire parler de	8,9%
avoir besoin	6,3%
dire	5,1%
disposer	5,1%
rêver	3,8%
souffrir	3,8%
sortir	2,5%
dépendre	1,3%

Observons si cette étude qui porte sur un numéro du journal *le Monde* est confirmé par le Corpus Annoté de Paris 7. Nous avons extrait tous les verbes qui suivent un relatif en *dont*. Il est à remarquer que les verbes supports, l'emploi figé des passifs (comme *être convaincu*) et autres compositions verbales n'interviennent pas dans les résultats et peuvent même produire du *silence*. De plus les verbes *être* et *avoir* apparaissent majoritairement alors

qu'ils sont auxiliaires des temps composés. Enfin les verbes *pouvoir, devoir, aller, vouloir, venir* sont également très fréquents alors qu'ils sont auxiliaires de temps ou de mode.

Les résultats sont donc faussés par le manque de précision du Corpus Annoté de Paris 7 sur les sous-catégories verbales. Nous avons éliminé les verbes *être, avoir, faire, devoir*, et *pouvoir* pour éviter un trop fort bruit. De plus, comme pour la précédente étude, les relatifs *dont* n'ont pas été distingués.

Les résultats montrent néanmoins que la répartition est plus "plate" que celle qui a été obtenue par Claire Blanche-Benveniste sur l'oral. Le verbe qui apparaît le plus fréquemment (*bénéficier*) représente près de 3% des occurrences.

Les verbes les plus fréquents dans cette construction sont les suivants :

lemme	**nombre d'occurrences**	**fréquence**
bénéficier	12	3,2%
détenir	10	2,7%
disposer	10	2,7%
aller	9	2,4%
venir	8	2,1%
savoir	8	2,1%
parler	8	2,1%
vouloir	7	1,8%
avoir besoin	7	1,8%
jouir	6	1,6%
souffrir	5	1,3%
étayer	4	1%
autres		75,2%

La répartition des verbes en dehors de cette construction est sensiblement la même et montre que le figement lexical de cette construction est un phénomène que l'on ne peut retenir pour ce type de textes de français écrit contemporain.

5.2 Études de fréquences

5.2.1 Fréquences lexicales

Les fréquences lexicales du français ont souvent été calculées à partir de données brutes ([Catach, 1984]). Comme le montre [Silberztein, 1993], de tels calculs sont nécessairement erronés du fait de la proportion élevée de formes ambiguës.

Voyons comment la désambiguation des parties du discours opérée sur notre corpus améliore ces calculs.

Si nous trions ces formes par ordre de fréquence, nous obtenons la liste de la seconde colonne (tableau 5.1) comme les formes les plus courantes, qui comprennent seulement des mots fonctionnels (prépositions, déterminants, conjonctions) et qui sont comparables avec ce que d'autres auteurs trouvent dans différents corpus français. Mais la plupart de ces formes sont en fait ambiguës : *de* peut être une préposition ou un déterminant, *le* peut être un déterminant ou un pronom, *en* peut être une préposition ou un pronom clitique. Si on s'intéresse aux mots les plus courants dans le corpus, il est donc nécessaire, d'une part de distinguer ces formes ambiguës et, d'autre part de rassembler différentes flexions du même mot (*d'* et *de* pour la préposition ***de***, *le, la, les, l'* pour le déterminant ***le***, etc.).

En faisant cela et en triant les formes par lemmes (désambiguïsés), nous obtenons la liste dans la troisième colonne qui est tout à fait différente. Le mot le plus courant est le déterminant ***le*** et quelques verbes (être, avoir) se trouvent parmi les 10 mots les plus fréquents.

Si maintenant nous trions les catégories elles-mêmes (quatrième colonne du tableau), la répartition des catégories montre que la catégorie la plus représentée est le nom (24,5% soit près du quart des occurrences), les verbes sont moitié moins nombreux et les adjectifs eux-mêmes moitié moins nombreux que les verbes. Contrairement à ce que laissent croire les calculs sur les fréquences lexicales (où *de* et *le* arrivent toujours en tête), les mots grammaticaux ne sont pas les plus fréquents du corpus.

5.2.2 La préférence lexicale pour les mots composés

Lors de la phase de segmentation en mots, nous avons d'abord vérifié les préférences pour les mots composés ([Gibs, 1985]). Nous avons pris les

Ordre de fréquence	Par forme	Par lemme		Par partie du discours	
1er	de (Prép ou Dét)	le (le, la, les, l')	Det	Noms	24,5% (20% NC, 4,5% NP)
2e	le (Dét or CL)	de (de, d')	Prep	Déterminants	16,8%
3e	les (Dét or CL)	à	Prep	Prépositions	14,6%
4e	la (Dét or CL)	un (un, une, des, de, d')	Det	Ponctuations	13%
5e	à	être (suis, est etc)	V	Verbes	11,4%
6e	l' (Dét or CL)	et	CC	Adjectifs	6,5%
7e	et	avoir (ai, a etc)	V	Adverbes	4%
8e	en (Prép or CL)	il (il, ils, elle, elles)	CL	Conjonctions	3,3% (2,3% CC, 1% CS)
9e	un	en	Prep	Clitiques	2,8%
10e				Autres pronoms	1,8%

TAB. 5.1 – *Fréquences lexicales par forme, par lemme et par catégorie (partie du discours)*

séquences qui sont éventuellement ambiguës entre les mots composés et les séquences en mots simples et avons calculé leur nombre d'occurrences respectifs. Voici des exemples de telles paires :

en fait : adverbe composé ou pronom clitique *en* suivi du verbe *faire*
d'ailleurs : adverbe composé ou préposition *d'* suivie du nom *ailleurs*

Quelques résultats sont montrés dans le tableau 5.2.

Forme	Nombre d'occurrences en mots composés	Nombre d'occurrences en mots simples
d'abord	154 (97 %) Adv	5 (3%) : Prep NC
alors que	231 (96%): CS	8 (4%): Adv CS
plus de	305 (60%) Prep	(40%) Adv Prep or Det
il y a	221 (57%) : Prep	(43%) : CL CL V
le plus	123 (39%) Adv	(61%) Det Adv
sur ce	0 (Adv)	65 (100%) Prep Det

TAB. 5.2 – *Proportion relative des catégories des mots simples et composés*

La préférence est attestée (plus de 93% des occurrences sont un mot composé en moyenne) mais dépend des catégories impliquées. Pour les mots composés nominaux et verbaux (comprenant généralement des noms, des verbes et des adjectifs), l'interprétation en mots composés concerne presque 100% des occurrences. Pour les mots composés adverbiaux, la préférence est moindre, et il y a des exceptions comme "sur ce" ou "le plus" dans le

tableau 5.2. Cette différence peut être expliquée par une préférence pour les catégories grammaticales (clitique, déterminant, préposition, etc.) associées avec les mots impliqués dans l'interprétation décomposée.

Nous vérifions que la préférence pour l'interprétation en mot composé est une préférence lexicale parce que le nombre total d'occurrences de mots composés dans le corpus est bien plus bas que celui des mots simples (50 614=6,2% versus 765 953=93,8%, sans compter la ponctuation).

5.2.3 La préférence lexicale pour les catégories grammaticales

Quand on considère les formes ambiguës syntaxiquement, les probabilités des différentes parties du discours sont généralement très inégales (cf [Church, 1988]).

Indépendamment des préférences syntaxiques qui peuvent être associées à telle ou telle unité lexicale (par exemple *ferme* est plutôt verbe que nom ou adjectif), nous avons cherché des principes de préférence plus généraux.

Regardons s'il y a une préférence globale des mots grammaticaux sur les mots lexicaux. Les fréquences de chaque catégorie sont présentées dans le tableau suivant :

Catégorie	**fréquence**	**proportion**
Nom	226 879	24,5% (20% noms communs, 4,5% noms propres)
Déterminant	156 008	16,8%
Préposition	134 753	14,6%
Ponctuation	122 448	13%
Verbe	105 901	11,4%
Adjectif	60 310	6,5%
Adverbe	45 204	4%
Conjonction	30 623	3,3% (2,3% coordonnants, 1% subordonnants)
Clitique	26 055	2,8%
Pronom	17 172	1,8% (1% pronoms relatifs, 0,8% autres pronoms%)

Les noms, verbes, adjectifs et adverbes couvrent 46,4% des mots en comptant les ponctuations et 54,6% sans les compter. Il n'y a pas de préférence globale pour les mots grammaticaux ou lexicaux si l'on considère l'ensemble des mots du corpus.

En revanche, nous avons trouvé une forte préférence pour les catégories grammaticales par rapport aux catégories lexicales pour les mots ambigus entre ces deux classes. Nous entendons par catégories grammaticales les listes fermées de mots fonctionnels (déterminants, prépositions, pronoms clitiques et autres pronoms, conjonctions de subordination et de coordination), et par catégories lexicales les verbes, noms, adverbes et adjectifs. Nous avons pris l'ensemble des formes ambiguës entre ces deux classes et avons étudié les fréquences respectives de leurs occurrences. Voici quelques exemples :

Formes ambiguës	**Nombre total d'occ.**	**Nombre d'occurrences**	
		lexicale	**grammaticale**
car	235	5 (2,1%) Nom	230 (97,8%) Conjonction de coordination
cela	284	1 (0,3%) Verbe	283 (99,7%) Pronom
dans	5341	0 (0%) Nom	5341 (100%) Préposition
devant	285	33 (11,5%) Verbe	252 (88,4%) Préposition
entre	1195	23 (1,9%) Verbe	1172 (98%) Préposition
envers	25	3 (12%) Nom	22 (88%) Préposition
la	24471	1 (0.0%) Nom	24470 (100.0%)
lui	763	0 (0%) participe de luire	763 (100%) pronom clitique et pronom fort.
or	189	30 (15.9%) Nom masc	159 (84.1%) Conjonction de coordination
si	989	0 (0%) Nom masc	989 (100%) Conjonction de subordination, Adverbe
son	2427	10 (0.4%) Nom masc	2417 (99.6%) Déterminant
sous	359	25 (6.9%) nom masc pluriel	334 (93.1%) préposition
ton	31	22 (70.9) nom masc sing	9 (29.1%) déterminant

TAB. 5.3 – *Fréquence relative par catégorie de quelques formes ambiguës*

En tout, nous avons trouvé une proportion écrasante d'emplois de catégories grammaticales (plus que 95% en moyenne, parfois 100%). Quelques-uns font exception ; leur forme lexicalisée est particulièrement rare ou réservée à une langue de spécialité (*dans* pluriel de *dan*, *par* nom commun – terme de golf, *sur* adjectif synonyme d'acide, etc.). En revanche d'autres apparaissent plus couramment (noms communs *or, envers*, verbe *lui*, etc.).

Cette préférence est confirmée par des interrogations d'autres corpus, en particulier de corpus d'oral transcrits. Dans la mesure où elle est stable, elle peut être considérée comme représentative des stratégies de désambiguïsation humaine.

Chapitre 6

Enrichissements syntaxiques du corpus annoté

Les annotations (catégories, mots composés, lemmes) présentées précédemment sont utiles pour toutes sortes d'interrogations, mais pas forcément suffisantes. Du point de vue linguistique, si l'on s'intéresse aux constructions passives, aux relatives ou aux sujets inversés, un niveau supplémentaire d'annotation est nécessaire. De même, du point de vue de l'informatique linguistique, si l'on veut entraîner ou évaluer des parsers et non des taggers. C'est ce que nous allons maintenant présenter en distinguant des sous-syntagmes assez figés (*clusters*) et le découpage en constituants proprement dit. D'autres enrichissements sont bien sur envisageables comme celui des fonctions grammaticales ou des valences verbales.

6.1 Annotation de *clusters*

Notre travail sur les *clusters* s'inspire des travaux du LADL sur les grammaires locales ([Maurel, 1989], [Maurel, 1991], [Gross, 1995]).

Un grand nombre de termes tels que les *dates, adresses, mesures, nombres, titres*, etc., échappent à la segmentation en mots composés car ils ne font pas l'objet d'un figement morpho-syntaxique complet. Néanmoins ces séquences ont la propriété de contenir des paradigmes plus ou moins fermés de mots comme les noms de jours, les noms de lieux, etc. ; il convient donc d'en faire une analyse distinguée des autres niveaux syntaxiques. Nous prendrons la liberté de les appeler *Clusters* pour nous garder de les confondre avec d'autres

types d'éléments du langage. Maurice Gross appelle cela des grammaires locales ou des micro-grammaires en envisageant toutefois ces éléments du langage à un niveau syntaxique plus avancé [Gross, 1995].

En voici quelques exemples :

- Du 7 au 9 juin 1997
- Entre 1987 et 1990
- L'hiver 1944
- 8, rue du chat qui pêche
- le Président Valéry Giscard d'Estaing
- Madame le ministre de l'intérieur
- 76 km/h
- \$56
- £78
- 78FF

Bien qu'ils soient composés pour partie de séquences de termes constants, ces syntagmes ne peuvent pas non plus être confondus avec des expressions figées. Le figement de ces dernières porte sur tout ou partie des éléments qui les constituent mais l'expression figée ne requière aucunement une description syntaxique qui lui est propre. Ainsi les expressions idiomatiques comme *casser du sucre sur le dos de quelqu'un, compter sur ses doigts, avoir le vent en poupe, casser sa pipe*, etc. ne présentent aucune irrégularité syntaxique sinon des restrictions lexicales portant sur leurs éléments. Ces expressions figées ne se réduisent pas non plus à une unité morphologique, elles ne peuvent pas se réduire par exemple à un verbe unique comme le montre cet exemple :

(a) *Jean ne casse sa pipe pas. ([Gross, 1996])
(b) Jean ne casse pas sa pipe.

La syntaxe des *clusters*, en revanche, met en œuvre des séquences qui n'apparaissent pas à d'autres niveaux comme celles des titres (exemple : *madame le premier ministre* - Nom-déterminant-adjectif-nom), et l'on peut y reconnaître des unités morphologiques en distribution avec des syntagmes ou des mots libres.

La structure fonctionnelle des *clusters* n'est cependant pas celle des autres syntagmes. Les noms propres sont épithètes de noms de lieu dans les adresses (*place Colonel Fabien*), les noms de mois sont conjoints à des nombres indiquant le jour et l'année dans les dates (*Le 8 juin 1944*), sans que les fonctions

grammaticales de ces termes soient clairement définies, de longues séquences de noms se trouvent dans les titres *M. Jean Dupont député RPR*. De telles structures ne se retrouvent guère dans la langue en dehors de ces expressions.

Les *clusters* constituent des unités morphologiques. Les insertions à l'intérieur de ces éléments sont impossibles et chaque *cluster* est une unité syntagmatique au niveau de la syntaxe. Les dates sont en distribution avec les adverbes de temps, les titres avec des noms propres, etc. Ces syntagmes s'analysent comme d'autres constituants du point de vue de leur contexte. Les suites *du 7 au 9 juin 1997, entre 1987 et 1990, l'hiver 1944* peuvent être substituées à *hier* et apparaître comme ajout phrastique comme dans :

(*Du 7 au 9 juin 1997*+*hier*), je suis allé au conservatoire de peinture.

Le figement des *clusters* peut se réduire à la description d'une grammaire de Kleene[1]. Ils s'articulent autour d'un mot puisé dans une liste fermée et d'un contexte qui ne s'exprime pas en termes de syntagmes mais de catégories. C'est-à-dire qu'ils sont *reconnaissables* par le type de séquence qui les constitue et non par leur structure. Les grammaires de Kleene sont fortement équivalentes à des réseaux d'automates finis non récursifs[2]. C'est sous forme de ces graphes qu'on trouve les descriptions des expressions rationnelles dans le système INTEX ([Silberztein, 1993]).

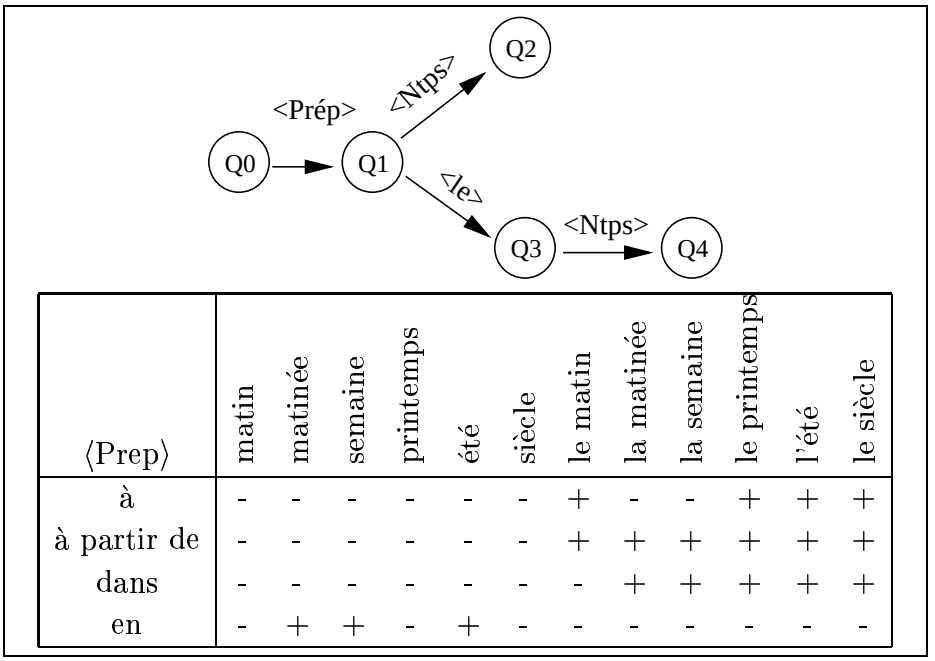
Le système INTEX permet de décrire des grammaires de Kleene pour des mots composés et des grammaires locales comme celle des pronoms clitiques du français. [Silberztein, 1993] propose une grammaire des nombres en français. [Maurel, 1991] a fait une telle description pour les adverbes de date en français, mais remarque qu'une représentation sous forme d'automate conduit à construire un graphe extrêmement complexe. En multipliant les prépositions, les déterminants et les noms de temps dans les adverbes de date, un réseau non récursif devient simplement impossible à manipuler tant le nombre de nœuds et de transitions est grand.

[Maurel, 1991] propose de décrire ces graphes complexes sous forme de réseau de classes associées à des matrices booléennes (cf [Gross, 1975]).

Les classes sont utilisées pour factoriser des informations redondantes dans un graphe classique. Les tables d'acceptabilités représentent les n-uples de classes conduisant à des séquences agrammaticales, c'est-à-dire des chemins dans le graphe qui conduiraient à des séquences agrammaticales. La représentation factorisée utilisée par Denis Maurel est donc strictement équivalente à un réseau non récursif classique mais plus compact.

1. Ou grammaire rationnelle.
2. Un réseau d'automates finis récursif est une grammaire non contextuelle.

Nous reprenons un graphe et sa table d'acceptabilité en figure 6.1. Les séquences comme *à partir du matin*, *dans la matinée*, *en matinée* sont reconnues mais non **à partir de matin* ou **dans le matin*. Or la séquence est simplement décrite comme la succession d'une préposition et d'un nom de temps avec ou sans l'article *le*.



⟨Prep⟩	matin	matinée	semaine	printemps	été	siècle	le matin	la matinée	la semaine	le printemps	l'été	le siècle
à	-	-	-	-	-	-	+	-	-	+	+	+
à partir de	-	-	-	-	-	-	+	+	+	+	+	+
dans	-	-	-	-	-	-	-	+	+	+	+	+
en	-	+	+	-	+	-	-	-	-	-	-	-

FIG. 6.1 – *Graphe et table d'acceptabilité associée pour des adverbes de date du français (repris de [Maurel, 1991])*

Nous n'avons pas retenu cette méthode pour décrire les *clusters* car notre but est d'articuler l'analyse de ces séquences à une analyse d'un niveau supérieur, non de décrire la syntaxe de la langue par des grammaires locales. Or les redondances dont la représentation sous forme d'automates de classes fait l'économie sont dues à des régularités qui n'intéressent pas seulement les *clusters* mais la syntaxe de la langue en général. Dans l'exemple que nous avons repris de [Maurel, 1991], les suites **à la matin*, **en le matin* sont agrammaticales non pas parce qu'elles n'appartiennent pas aux adverbes de date comme *au matin* ou *en matinée*, mais parce que les règles d'accord ou d'emploi de la préposition *en* ne sont pas respectées.

C'est pourquoi il nous paraissait nécessaire de développer un module d'analyse propre à repérer les expressions régulières, sans mettre en cause un module d'analyse morphologique en amont et un module d'analyse syn-

taxique en aval. Nous le dissocierons donc des autres procédures d'analyse.

Ces syntagmes ont la particularité de pouvoir être décrits par des grammaires régulières dans la mesure où ils ne contiennent aucune description syntagmatique et ils s'articulent autour d'un mot appartenant à une liste fermée.

Nous avons retenu une partie des adverbes et noms de dates, des mesures, des nombres et certains titres et certaines adresses. Nous n'avons pas retenu toutes les séquences que nous aurions nommées *cluster* car celles-ci ne se distinguent pas toujours par une grammaire de Kleene qui leur est propre. En effet, les noms de lieux qui apparaissent nécessairement dans les adresses par exemple comme *cité*, *villa*, *rue*, etc., ne sont pas rares dans d'autres contextes et les séquences des termes comme *allée du petit bois*, ou *place du colonel Fabien* ne sont pas propres aux adresses. Nous nous bornons donc à reconnaître des *clusters* de façon fiable, sans les reconnaître tous au risque de produire du *bruit* dans nos données. Nous favorisons le *rappel* au détriment de la *précision* lors de cette phase d'annotation car il nous semble que le *silence* produit lors de cette phase ne nuit pas aux autres niveaux d'annotation comme pourrait nuire le *bruit* induit par des règles moins restrictives.

Nous avons privilégié une description linguistique de ces expressions régulières «par règles» et non en utilisant un éditeur de graphes comme le proposent les systèmes comme Intex ([Silberztein, 1993]). Il nous semble en effet, qu'il est plus simple de décrire les régularités par des expressions régulières en les nommant et en notant directement les syntagmes et paradigmes qu'en les dessinant lorsqu'elles sont complexes.

Le programme *Cluster* que nous avons présenté au chapitre précédent permet d'annoter des séquences reconnues par une grammaire de Kleene. Les dates, nombres, mesures, adresses et titres sont décrits par autant de fichiers distincts qui sont traités par le programme.

6.1.1 Nombres

Les nombres en français, qu'ils soient écrits en lettres ou en chiffres, respectent une grammaire de Kleene. Nous ne cherchons pas cependant à reconnaître ici la bonne formation de ces suites, mais toutes et seulement les suites qui sont des nombres. Pour rendre la tâche plus simple, nous avons pris le parti de décrire un langage moins précis. L'automate que nous avons décrit ne correspond pas à l'automate générant le langage des nombres en français, mais à un automate qui reconnaît toutes et seulement les suites

```
#define cardinaux   [mM]ille | ([sS]oixante[_-]et | [qQ]uatre[_-]vingt)[_-]une |
                    ([qQ]uarante | [cC]inquante | [tT]rente | [vV]ingt)[_-]et_une |
                    [qQ]uatre | ([qQ]uarante | [cC]inquante | [sS]oixante | [tT]rente
                    | [vV]ingt | [qQ]uatre[_-]vingt)[_-]quatre | ([nN]on | [qQ]uar |
                    [oO]ct | [sS]ept | [cC]inqu | [sS]oix)ante | [tT]rente | ([sS] |
                    [tT]r | [sS]oixante[_-]tr | [qQ]uatre[_-]vingt[_-]tr | [sS]oixante[_-]s
                    | [qQ]uatre[_-]vingt[_-]s)eize | [oO]nze | [qQ]uinze | ([sS]oixante |
                    [qQ]uatre[_-]vingt)[_-]quinze | ([sS]oixante[_-]et | [qQ]uatre[_-]vingt
                    | [sS]oixante_et)[_-]onze | [qQ]uatorze | ([sS]oixante | [qQ]uatre[_-
                    ]vingt)[_-]quatorze | [dD]ouze | ([sS]oixante | [qQ]uatre[_-]vingt)[_-
                    ]douze | [nN]euf | ([qQ]uarante | [cC]inquante | [sS]oixante |
                    [tT]rente | [vV]ingt | [qQ]uatre[_-]vingt | [dD]ix | [sS]oixante[_-
                    ]dix | [qQ]uatre[_-]vingt[_-]dix)[_-]neuf | [mM]il | ([qQ]uarante |
                    [cC]inquante | [sS]oixante | [tT]rente)[_-]et[_-]un | ([qQ]uatre |
                    [qQ]uatre)[_-]vingt[_-]un | ([qQ]uarante | [qQ]uarante | [cC]inquante
                    | [sS]oixante | [tT]rente | [vV]ingt)_et_un | [zZ]éro | [cC]inq |
                    ([qQ]uarante | [cC]inquante | [sS]oixante | [tT]rente | [vV]ingt |
                    [qQ]uatre[_-]vingt)[_-]cinq | [mM]illiards | [tT]rois | etc.
#define numeraux   ([mM]ille | [mM]illier | [mM]illion | [mM]illiard | [bB]illion |
                    [bB]illiard | [dD]izaine | [dD]ouzaine | [vV]ingtaine | [tT]rentaine
                    | [qQ]uarantaine | [cC]inquantaine | [sS]oixantaine | [cC]entaine
                    | [mM]illes | [mM]illiers | [mM]illions | [mM]illiards | [bB]illions
                    | [bB]illiards | [dD]izaines | [dD]ouzaines | [vV]ingtaines |
                    [tT]rentaines | [qQ]uarantaines | [cC]inquantaines | [sS]oixantaines
                    | [cC]entaines)
#define digit      [0-9][0-9,.]*

num                {
                   w=digit w="([-,.])"? w=digit* ;
                   w=digit* (w=digit | w=cardinaux)+ w=numeraux? (w=de
                   w=cardinaux)?
                   }
```

FIG. 6.2 – *Expressions régulières des nombres en langage Cluster (extrait)*

textuelles qui sont des nombres et qui apparaissent effectivement dans les textes. L'ensemble des chemins de cet automate produit également des suites comme **trente-six huit* ou **vingt et quatre-vingt.*

Nous présentons fig 6.2 le fichier cluster de la grammaire des nombres.

Les éléments définis grâce à `#define` sont des *macros*, ils permettent simplement d'alléger l'écriture des règles en abrégeant une expression complexe reprise plusieurs fois par un simple identificateur.

6.1.2 Dates

Les dates s'articulent autour de noms d'heures, de jours, de mois, de saisons et de périodes (*soir, matin, midi, hier, aujourd'hui, week-end*, etc.). Nous nous limiterons aux expressions temporelles qui contiennent un tel terme en excluant toutes les autres comme *en semaine*, *pendant que Luc dormait*, *après la demie*, *avant le pont*.

Elles s'emploient éventuellement avec des prépositions temporelles (*depuis, pendant, durant, avant, après, à partir de, dans, vers, en*, etc.). Et certaines figures comme celles qui expriment un intervalle de temps ou une conjonction de plusieurs périodes se reconnaissent par des séquences propres (*du 3 au 6 mars*, *les 4 et 5 juillet*).

Les dates s'emploient parfois avec des nombres et il nous faut donc reconnaître ceux-ci avant. Le programme *Cluster* permet de décrire un réseau d'automates non récursifs, il nous était donc possible d'inclure la grammaire des nombres dans celle des dates. Nous avons cependant préféré faire précéder l'analyse des dates par l'analyse des nombres, ce qui revient au même. La procédure de reconnaissance des différents *clusters* retenus est donc une succession d'analyses de grammaires locales ordonnées.

Nous présentons en figure 6.3 un exemple de fichier *cluster* encodant l'automate des dates.

Dans cet exemple, certaines suite de mots sont simplement identifiées par `<num>`. Ces suites correspondent au balisage assuré lors de la reconnaissance des nombres.

6.1.3 Mesures

Les mesures comprennent des noms d'unités qui leur sont propres (*minutes, kilo, centimètres, km/h, barils/jour*, etc.) employés comme nom avec des cardinaux comme déterminant.

La figure 6.4 illustre un exemple de grammaire des mesures.

6.1.4 Titres

Les titres sont construits autour d'un nom propre et d'un ensemble de noms désignant un individu ou une personnalité (M., Président, Pr.). La figure 6.5 illustre un exemple de grammaire des titres.

```
#define jours          ([lL]undi | [mM]ardi | [mM]ercredi | [jJ]eudi | [vV]endredi | [sS]amedi | [dD]imanche)
#define mois           ([jJ]anvier | [fF]évrier | [mM]ars | [aA]vril | [mM]ai | [jJ]uin | [jJ]uillet | [aA]oût |
                       [sS]eptembre | [oO]ctobre | [nN]ovembre | [dD]écembre)
#define saisons        ([éÉE]té | [hH]iver | [pP]rintemps | [aA]utomne | [sS]aison)
#define milieu         ([mM]ilieu | [dD]ébut | [fF]in)
#define depuis         ([dD]epuis | [pP]endant | [dD]urant)
#define heure          ([hH]eure | [mM]inute | [sS]econde)
#define prep           ([dD]epuis | [pP]endant | [dD]urant | [aA]vant | [aA]près | [àÀ]_partir_de | [dD]ans |
                       [vV]ers | [eE]n)
date        {
            w=de w=le num w=à w=le num w=mois? num?;
            w=depuis w[cat="D"]? ( w=saisons | w=mois ) num? w=prochain?;
            w[cat="D"] ( w=saisons | w=mois ) num? w=prochain?;
            ( w[cat="D"] | w="en" ) w=milieu w[lemma=de]? ( w=saisons | w=mois ) num? w=prochain?;
            prep ( w=saisons | w=mois ) num? w=prochain?;
            prep w=mois? num;
            w[cat=D, mph=ms] w=jours w=soir? num? w=mois? num?;
            w[cat=D] num? prochain ( w=jour w=soir? | w=saisons | w=mois ) num?;
            w="le"? w=jours w=soir? num w=mois? num?;
            w="le" num w=mois? num?
            }
```

FIG. 6.3 – *Expressions régulières des dates en langage Cluster*

```
#define unite          (% | [kK]ilo(s)? | [hH]eure(s)? | [mM]inute(s)? | [sS]econde(s)? | [fF]ranc(s)? |
                       [cC]entime(s)? | [lL]ivre(s)? | [dD]ollar(s?) | [kK]m | £ | $)
#define temps          ([hH]eure | [hH] | [mM]inute | [mM]n | [sS]econde | [sS]ec | [jJ]our | [jJ] | [mM]ois |
                       [aA]nnée)
mesure      {
            num w=unite (w='/' w=temps)?;
            w[lemma="de"]? num (w=unite (w='/' w=temps)? )? w="à" num w=unite (w='/'
            w=temps)?;
            w="[0-9_]+h[0-9_]+"
            }
```

FIG. 6.4 – *Expressions régulières des noms de mesures en langage Cluster*

6.1.5 Adresses

La figure 6.6 illustre un exemple de grammaire des adresses. Un grand nombre d'adresses ne seront pas reconnues par cette simple grammaire, en revanche elle produira peu de bruit comme nous l'avons expliqué *supra*.

(a) Le 6 de la rue Censier
(b) 3, place de la République, Paris 15^{e}

Ces exemples seraient délicats à analyser par les règles de la syntaxe générale. En (b), *3* est analysé comme nom commun (par analogie avec

```
#define monsieur        ([mM]onsieur | [mM]adame | [mM]ademoiselle | M. | Mr | Mme | Mlle)
#define titre           ([dD]octeur | [pP]rofesseur | [mM]aître | [mM]e | [dD]r. | [pP]r. | [cC]olonel | [aA]miral
                        | [gG]énéral | [lL]ieutenant | [bB]rigadier | [sS]ergent | [sS]ergent-chef | [lL]ady | [lL]ord
                        | [sS]ir | [mM]onseigneur)
nom             {
                // Madame le Pr. Dupont
                // Dupont de Bretelle
                // le Pr. Dupont
                // Pr. Dupont
                (w=monsieur)? w[cat=D]? w=titre? w[cat=N, subcat=P] ( w[cat=P, lemma='de'] w[cat=N,
                subcat=P] )*;
                }
```

FIG. 6.5 – *Expressions régulières des noms de titres en langage Cluster*

```
#define lieux           ([pP]lace | [rR]ue | [rR]uelle | [cC]hemin | [vV]oie | [aA]llée | [aA]venue | [bB]oulevard
                        | [sS]quare | [rR]oute | [iI]mpasse | [rR]ésidence | [bB]loc | [pP]avillon | [rR]ond_point
                        | [bB]d.? | [bB]d. | [aA]v.?)
adresse
                {
                num w[cat="PONCT"]? w=lieux[cat=N, subcat=common] w[cat=P]? name (name)?;
                num w[cat="PONCT"]? w=lieux[cat=N, subcat=common] w[cat=P] w[cat=D] w[cat=N, sub-
                cat=common];
                w=lieux[cat=N, subcat=common] name;
                w=lieux[cat=N, subcat=common] w[cat=P] w[cat=D] w[cat=N, subcat=common]
                }
```

FIG. 6.6 – *Expressions régulières des noms de lieux en langage Cluster*

l'exemple (a)). Si *3* est la tête du syntagme nominal, quelle est la fonction de *place de la République* ?

6.2 Annotation en syntagmes

Nous présentons ici le travail d'équipe auquel nous participons depuis 1999. L'annotation en constituants consiste à regrouper certaines séquences de catégories et à leurs assigner une étiquette. Cette procédure consiste à reconnaître des syntagmes et l'organisation structurelle de ces syntagmes dans le texte.

De même que pour l'annotation morpho-syntaxique, nous avons procédé en deux étapes successives pour annoter syntaxiquement le corpus. Un marquage automatique est suivi d'une validation systématique par une équipe d'étudiants linguistes.

Le découpage en syntagmes et l'assignation de catégories syntaxiques à ces

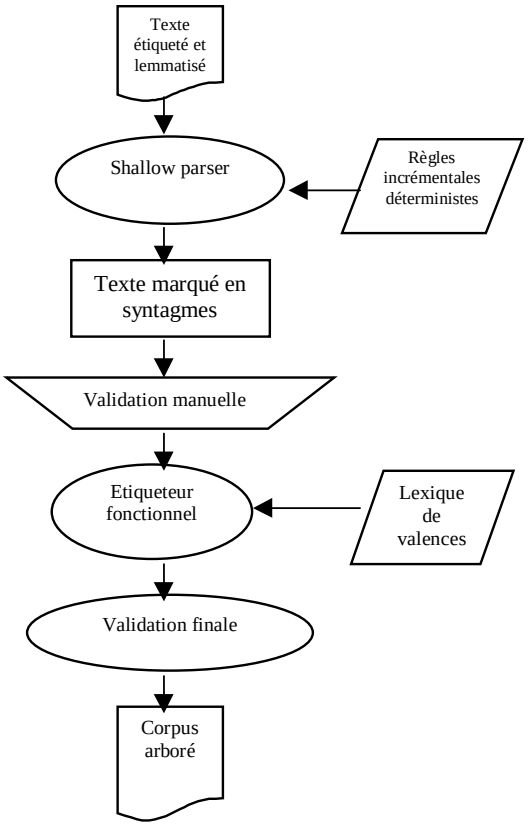


FIG. 6.7 – *Annotation en constituants du corpus*

syntagmes sont effectués par un analyseur de surface développé par Alexandra Kinyon [Kinyon, 2001]. Nous présenterons cet outil et discuterons du type d'algorithme utilisé et de ses alternatives.

Contrairement à d'autres projets d'annotation syntaxique de corpus (Prague Dependency Treebank, Penn Treebank), il ne s'agit pas d'appliquer une théorie syntaxique particulière, mais de contribuer à l'émergence d'un standard d'annotation syntaxique. Les choix devront ainsi être traduisibles dans plusieurs cadres théoriques et être le plus neutre possible au regard de ces théories.

Le niveau d'annotation syntaxique ne suppose pas un niveau de surface calculé par mouvements ou transformations à partir de celui-ci. Nous procédons à une analyse surfaciste de la structure de la phrase et les dépendances fonctionnelles permettront le marquage des éléments qui ne se trouvent

pas dans leur position canonique. Faire autrement aurait conduit à rendre inutilisable le corpus par des linguistes qui n'entendent pas utiliser un modèle générativiste.

Le marquage des constituants est minimal dans le sens où certains syntagmes souvent contestés comme le groupe verbal ou le syntagme déterminant n'ont pas été retenus ici.

Nous allons présenter l'analyseur de surface qui a permis d'annoter automatiquement les syntagmes puis les choix linguistiques et enfin les syntagmes retenus.

6.2.1 L'analyseur de surface

Alexandra Kinyon a développé un analyseur de surface[3] ([Clément & Kinyon, 2000], [Kinyon, 2001]) dont le but est le marquage des frontières de constituants et leur étiquetage. Bien que permettant le marquage de constituants enchâssés, cet outil ne reconnaît qu'une sous-classe des grammaires hors contexte, et a fortiori des grammaires contextuelles. Cette particularité lui offre une très grande robustesse et rapidité comme nous allons le montrer.

Les premières mesures sur les résultats de l'analyseur de surface ont été calculées avec 1 000 phrases validées manuellement. Les mesures ont été faites en comptant l'écart entre les bornes correctement placées et étiquetées sur l'échantillon annoté automatiquement et le même échantillon annoté manuellement. L'analyseur fonctionne avec une cinquantaine de règles, le résultat de ces mesures donne plus de 80% de précision sur les frontières ouvrantes et plus de 60% sur les frontières fermantes[4]. Les bornes fermantes sont logiquement moins bien analysées car plus difficiles à repérer.

Le programme fonctionne sur une base de règles qui déterminent les frontières gauches et droites des constituants en fonction du constituant courant et de diverses propriétés des mots lus (morphologie, partie du discours, etc.). Il prend en entrée un texte *taggé* (étiqueté et désambigué) et se sert des catégories fonctionnelles comme frontières ouvrantes d'un syntagme (déterminant pour les groupes nominaux, préposition pour les groupes prépositionnels, etc.). Généralement il ferme un syntagme quand il en ouvre un autre, sauf dans les cas où l'enchâssement est permis (infinitives, relatives, etc.). Il permet par exemple de marquer le début d'une relative dans

3. Ou *shallow-parser*.

4. 80% si on se contente de syntagmes non enchâssés, 60% si on compare avec les enchâssements réalisés par les humains.

un groupe nominal à l'endroit d'une préposition suivie d'un relatif. Nous reviendrons sur l'intérêt de marquer indépendamment les frontières gauches et droites des constituants mais intéressons-nous à présent du type d'algorithme utilisé.

Le programme d'Alexandra Kinyon garantit le bon équilibre des frontières de constituants et l'absence de croisement. L'algorithme est donc basé sur un automate fini à pile (*push-down storage automation PDS*) qui comme le prouvent [Chomsky & Miller, 1968] permet de reconnaître les langages hors contexte.

Pour marquer les constituances, les automates à pile sont particulièrement adaptés puisqu'ils permettent de reconnaître à la fois les symétries (ou le balancement des frontières de syntagmes) et l'auto-enchâssement. Cependant plusieurs difficultés apparaissent à utiliser cette famille d'automates. Les connaissances linguistiques de la structure des syntagmes doit se rapporter à une grammaire hors contexte. Nous ne développerons pas ici les différents arguments qui ont conduit à abandonner ce type de grammaire par Chomsky lui-même dans [Chomsky, 1957]. Par ailleurs, la complexité en temps pour reconnaître une suite comme appartenant à une grammaire hors contexte est polynomiale et satisfait peu les exigences de grande rapidité de l'outil. Bref, s'en tenir à un automate fini à pile ne permettait pas de développer un analyseur de surface à la fois robuste et rapide. Pour résoudre ce problème, Alexandra Kinyon a contraint les règles à être déterministes. Or nous savons qu'il n'existe pas d'équivalence déterministe des automates à pile contrairement aux automates finis non déterministes.

Le langage engendré par l'automate est donc beaucoup plus contraint, il se situe entre les langages hors-contextes et les langages réguliers : il est de type LR(k). Un automate à pile déterministe ne permet pas de reconnaître toutes les symétries. En particulier, le langage symétrique $\{xy/x,c,y \in V_t^*\}$ (avec y image de x) auquel appartient les chaînes (*abccba, abba*, etc.). ne peut être engendré par un automate à pile déterministe[5].

Les frontières gauches et droites des syntagmes sont annotées indépendamment dans l'analyseur de surface. Nous venons de voir qu'un automate à pile déterministe permettait d'équilibrer ces frontières dans la mesure où l'automate se trouve dans une configuration différente lors de la mémorisation des syntagmes ouverts de celle qui permet de les fermer. C'est-à-dire que les règles doivent déterminer les conditions d'ouverture ou de fermeture d'un

5. La symétrie d'une chaîne est engendrée par un dépilement systématique de tous ces premiers éléments. Cette procédure est faite à partir d'un état $Q2$. Or cet état est atteint de façon non déterministe depuis un état $Q1$ avec les même éléments qui sont empilés.

syntagme sans conflit. Ceci convient bien à un analyseur de surface dont les conditions de marquage de frontières de syntagmes sont conditionnées par des mots fonctionnels.

Un automate déterministe, qu'il soit à pile ou non, permet de reconnaître un élément du langage dans un temps proportionnel à la longueur de cet élément. On voit l'intérêt d'utiliser de tels automates pour développer un analyseur de surface.

D'autres tentatives on été faites pour reconnaître les frontières de constituants par des grammaires peu précises. Jacques Vergne (exposé au séminaire TALaNa du 4 décembre 2000 - non publié), dans le cadre d'une étude sur des unités prosodiques, a utilisé une grammaire régulière pour reconnaître les frontières de *chunk*[6]. Le bon balancement des syntagmes ne peut cependant être assuré par le langage régulier engendré mais par des heuristiques (notamment l'utilisation de variables jouant le rôle d'une pile finie). De même, les enchâssements ne sauraient être respectés à l'infini sans automate à pile.

Parmi différents essais, nous avons utilisé le programme *Cluster* pour reconnaître les principaux constituants (groupes nominaux, verbaux, prépositionnels et adjectivaux) du corpus, et non leurs frontières, avec cependant une impossibilité à pouvoir analyser l'auto-enchâssement. Ceci revient à programmer un réseau d'automates non récursifs. Cette procédure est souvent explorée pour développer des *chunkers* ou analyseurs de surface non récursifs dans le but de reconnaître par exemple des suites pour un extracteur automatique de terminologie ([Gaussier *et al.*, 2000]).

Nous avons également développé un analyseur syntaxique (XLFG) pour une sous-classe des grammaires contextuelles, les grammaires lexicales fonctionnelles[7] (lexical-functionnal grammars) [Clément & Kinyon, 2001], [Clément, 2000], mais également pour une sous-classes des grammaires attribuées (programme YAB). Nous ne présentons pas ces analyseurs dans ce mémoire, mais nous notons que cette expérience nous a montré combien l'analyse de surface est une procédure incontournable pour qui veut analyser de façon robuste du texte *tout-venant*. En effet, par leur nombre et par la complexité algorithmique nécessaire à les reconnaître (et analyser), les ambiguïtés de structure rendent rédhibitoire une analyse robuste d'un corpus qui devrait se passer d'un analyseur de surface comme celui que nous venons de présenter.

6. Nous reprenons l'emprunt du terme *chunk* par Jacques Vergne lui-même pour désigner les unités prosodiques qu'il a défini.

7. LFG est une théorie avant d'être un formalisme linguistique. Nous employons ici le terme de façon réductrice pour n'envisager que la grammaire formelle.

Il nous semble que l'usage d'un automate à pile déterministe est la meilleure alternative pour reconnaître des constituances enchâssées sans passer par un reconnaisseur de grammaires hors contexte ou contextuelle. La reconnaissance de syntagmes pour l'annotation syntaxique d'un corpus représentatif par sa nature et par sa taille peut être envisagée de façon très robuste et efficace par un tel système.

L'expérience de construction de corpus syntaxiquement annoté nous renseigne sur l'analyse syntaxique telle qu'elle doit être envisagée dans les applications du Traitement Automatique de la Langue. Nous pensons que la reconnaissance de *chunks* ou de syntagmes par des automates déterministes et par des automates déterministes à pile dans les textes *tout-venant* doit être envisagée en complément de l'analyse de langages hors contexte et contextuels.

6.2.2 Quelques choix linguistiques

Nous présentons quelques choix linguistiques portant sur les structures de constituants. Une documentation [Abeillé *et al.*, 2000b] accompagne le corpus annoté ; elle sert à la fois de guide d'annotation et de document de référence pour qui veut exploiter le corpus annoté. Les choix ont été motivés d'une part par la volonté d'être le plus neutre qu'on puisse être au regard des théories syntaxiques dans un corpus arboré, d'autre part par un souci de simplicité et de reproductibilité (les consignes d'annotation doivent être simples à comprendre et à appliquer pour les annotateurs humains). La neutralité a un prix : celle de ne pas avoir une description syntaxique aussi fine que le proposent certains modèles syntaxiques.

Discontinuités

La forme négative verbale en français se construit par la collocation du clitique *ne* et d'une forme pronominale, interrogative ou adverbiale (*pas, point, gère, nullement, jamais, personne, quiconque*, etc). On pourrait être tenté de faire de ces deux formes un syntagme discontinu unique et le marquer comme tel puisqu'il apparaît sous forme continue dans les tours infinitifs (c) et s'y trouve en distribution avec une forme vide dans les tours non négatifs comme le montrent les exemples (a) à (d).

(a) Le ministre répond aux questions.
(b) Le ministre **ne** répond **pas** aux questions.

(c) Le ministre semble **ne pas** répondre aux questions.
(d) Le ministre semble répondre aux questions.

Cependant avec d'autres mots négatifs on trouve des constructions analogues : *Jean ne veut voir personne*, *Jean ne répond à aucune question*, sans que l'on puisse faire un constituant *ne personne*, *ne aucune* (**Jean préfère ne personne voir. *Jean demande de n'aucune question répondre*).

Certaines discontinuités semblent également apparaître quand plusieurs termes participent de la détermination d'un même nom.

(a) Les enfants ont tous été au cinéma.
(b) Ils iront tous au cinéma.

En (a), le quantifieur *tous* et l'article *les* déterminent *enfants* de telle sorte qu'on serait tenté de regrouper les trois termes en un seul constituant comme dans *tous les enfants* ou encore *tous les* en un groupe déterminant. Cependant, l'ordre est ici inverse et ceci conduirait non seulement à créer des constituants discontinus mais également d'ordre indifférent. De plus un syntagme *ils tous* ou *tous ils* serait mal formé en opérant de même en (b).

L'information contenue dans l'annotation de constituants doit permettre d'étudier la structure et l'organisation des syntagmes. Il est donc crucial que cette annotation marque l'ordre linéaire, l'ordre de dominance et la nature de ces constituants. Pour ces mêmes raisons, nous nous sommes interdit l'annotation de syntagmes croisés.

Nous n'avons donc jamais noté de constituants discontinus. Nous avons cependant conservé la possibilité d'annoter des dépendances à longue distance entre des éléments indépendamment des dominances et précédences syntagmatiques grâce à l'adoption du mécanisme de références croisées de la norme XML. Ceci permettra de conserver les rares discontinuités morpho-syntaxiques que nous avons signalé en 3.2.3 (**afin**, justement **de**, **compte tenu** notamment **de**, etc.).

L'annotation des fonctions grammaticales permettra également de spécifier des liens de dépendance syntaxiques indépendamment de l'arborescence des constituants.

(a) Paul en veut trois (*Pronom*).
On notera la fonction **objet** pour *trois* et rien pour *en*.
(b) Combien Paul veut-il de pommes?
On notera **sujet** pour *Paul*, **objet** pour *de pommes* et **modifieur** pour *combien* (avec un lien de dépendance avec *de pommes*)

L'étiquetage fonctionnel ne doit pas proposer deux fois la même fonction à différents éléments d'une même phrase comme si on disait que *trois* et *en* sont tous deux objet en (a) ou *combien* et *de pommes* en (b). Le principe de non-redondance fonctionnelle, exposé par [Milner, 1982] est violé dans les phrases suivantes :

(exemples repris de [Milner, 1982])

(a) *Un garçon que j'ai prévenu Paul que je punirai demain.
(b) *Un garçon à qui j'ai promis à Paul que je donnerai une pomme
(c) *Qui avez-vous prévenu Paul que vous puniriez demain ?
(d) *À qui avez-vous promis à Paul que vous donneriez une pomme ?

Il semble que ces exemples sont agrammaticaux pour cette seule raison. Les phrases *Un garçon à qui j'ai prévenu Paul que je parlerai demain*, *Un garçon que j'ai promis à Paul que je punirai demain* sont contruites sur la même structure avec les verbes ponts *prévenir* et *promettre* et sont grammaticales.

Catégories vides

Certaines fonctions comme les sujets des infinitives, les sujets et objets extraits des relatives et interrogatives n'ont pas de réalisation en position canonique. Certaines théories et certains corpus annotés (i.e. Penn Treebank) insèrent des catégories vides co-indicées au syntagme extrait :

Dans ces cas, nous n'avons pas tenu à marquer une catégorie vide qui devrait porter la fonction réalisée pour plusieurs raisons :

1. La catégorie vide qui porte une fonction suppose l'existence d'une *trace* syntaxique et donc d'une réalisation de surface calculée à partir d'une réalisation canonique de l'élément comme le propose la théorie transformationnelle (Chomsky 75). Or nous ne faisons pas l'hypothèse de telles transformations. Nous supposons que les réalisations de surface suffisent pour les marques syntaxiques de constituance et dépendance. Nous discuterons *infra* de quelques cas d'ellipses qui posent le même type de problème.

2. Toutes les formes infinitives, toutes les formes réfléchies n'assignent pas nécessairement une fonction à des éléments qui sont traditionnellement marqués comme *trace* syntaxique.

Dans les exemples suivants, le fait de marquer des catégories vides est davantage lié à une théorie qui en fait état dans ces constructions qu'à l'assignation d'une hypothétique fonction.

(a) Ce devoir est très difficile à $_{\emptyset}$ terminer.
(b) Les chemises se lavent $_{\emptyset}$ à l'eau froide.
(c) L'idée de $_{\emptyset}$ partir en vacances traverse tous les esprits.
(d) Une licorne a été vue $_{\emptyset}$.
(e) Il a été vu une licorne.

De plus, si l'on mettait une catégorie vide au passif, on introduirait une différence artificielle entre le passif impersonnel (e) et le passif personnel (d).

Syntagmes exocentriques

Nous avons expliqué au chapitre 3 les raisons qui nous ont conduit à ne pas recatégoriser les noms épithètes comme adjectifs (*l'opération déménagement*, *un gâteau maison*) ou encore les adjectifs employés comme adverbes (*chanter juste*, *y croire très fort*) ; à ne pas recatégoriser arbitrairement les termes en fonction de la catégorie canonique d'une position syntaxique.

Ceci conduit à construire des syntagmes nominaux sans nom, des groupes adjectivaux sans adjectif ou des adverbiales sans adverbe.

Un nom peut avoir les propriétés d'une épithète qualificative (*le plan* ***vache folle***, *une assurance* ***dégats des eaux***) ou d'un attribut (*Jean est très* ***famille***). La fonction portée par ce nom renseigne sur le rôle syntaxique du syntagme dont il est la tête. Le principe que nous avons adopté au niveau morpho-syntaxique de ne pas recatégoriser les mots au risque de générer un lexique incohérent n'a pas lieu d'être à ce niveau d'analyse.

Nous n'analysons pas non plus dans les tours elliptiques une tête effacée qui justifierait l'étiquetage du syntagme. Les tours elliptiques des comparatives (*Luc est plus grand que toi (tu es moins grand)*), des coordonnées (*Jean mange des pommes et Luc (mange) des poires*) sont annotés comme phrases bien que leur verbe n'ait aucune réalisation de surface. C'est également le cas des phrases averbales (exclamatives sans verbes, locutions). En revanche, quand la phrase elle-même est effacée comme c'est le cas selon certaines analyses des corrélatives, (*Jean est si bête qu'il ne s'en rend pas compte (qu'il est bête)*), nous n'annotons pas l'hypothétique phrase effacée par une catégorie vide.

Ambiguïté

Nous avons fait l'expérience qu'une lecture attentive permettait toujours de lever les ambiguïtés syntaxiques dans le corpus. Les très nombreux cas d'ambiguïté qu'offre l'analyse syntaxique sont évités par les auteurs dans ce genre de texte[8]. Quand le contexte sémantique, les connaissances du monde ou la pragmatique ne permettent pas de lever l'ambiguïté, l'auteur emploie une paraphrase non ambiguë pour préciser ses dires.

Les exemples fabriqués suivants sont propres à l'intuition du linguiste et non à l'exploration de corpus telle que nous en avons l'expérience :

(a) Jean voit un homme avec un télescope.
(b) Martin peint la grille du balcon.
(c) La belle ferme le voile.

Dans l'exemple suivant, l'ambiguïté est réelle entre une apposition (b) et deux coordinations (a) seulement pour le lecteur qui ne saurait pas que Bill Clinton était président des Etats-Unis et non maire de New York.

(a) [Bill Clinton]$_{NP}$ [, [le maire de New-York]$_{NP}$]$_{COORD}$ [et [Madeleine Albright]$_{NP}$] $_{COORD}$.
(b) [Bill Clinton, [le maire de New-York]$_{NP}$]$_{NP}$ [et [Madeleine Albright]$_{NP}$] $_{COORD}$.

Dans certains cas, la structure en constituants est multiple pour exactement la même interprétation. C'est le cas par exemple des constructions à verbe support où l'élément sous-catégorisé par le nom peut être indifféremment enchâssé sous celui-ci (a) ou dépendant du verbe (b).

(a) Jean [a commis]$_{VN}$ [une agression [contre Marie]$_{PP}$]$_{NP}$.
(b) Jean [a commis]$_{VN}$ [une agression]$_{NP}$ [contre Marie]$_{PP}$.

Pour trancher, si aucun test syntaxique ne permet de préférer l'une des structures, nous optons pour celle qui est la plus *plate*, c'est-à-dire celle qui possède le moins d'enchâssements ; (b) dans notre exemple.

8. Les textes juridiques, destinés à être interprétés par les magistrats, sont connus pour être particulièrement ambigus et difficiles à traduire pour cette seule raison. Le traducteur ne devant pas interpréter les textes juridiques, ils doivent conserver les ambiguïtés, mais également les non-dits du texte original.

6.2.3 Les syntagmes retenus

Catégorie du syntagme	Description	Catégorie lexicale tête	exemples
AP	groupe adjectival	Adjectif	grand de trois mètres, français, plus grand que toi, grand et gros
AdvP	groupe Adverbial	Adverbe	très gentiment, juste après, demain ou après-demain
NP	groupe nominal	Nom, pronom, adjectif, interjection	Jean Dupont, tous les trois, une grosse vache, moi, lequel, tout, rien, l'homme qui vient, les meuilleurs, le rouge, attention
PP	groupe prépositionnel	Préposition	à Paris, juste avant le match, avec une petite cuiller, comme toi, auquel
VPinf	infinitive	Verbe à l'infinitif	à croquer, ne rien leur dire, de venir, pour leur faire plaisir, sans avoir vécu
VPpart	participiale	Verbe au participe	en arrivant, une fois arrivé, n'aimant pas les artichauts, bien noté
VN	noyau verbal	Verbe (pas au participe)	l'a bien lu, aime, on y va, le voir
Srel	relative	Verbe s'il existe	qui viendra, que tu veux voir, dont trois malades
Ssub	propositions subordonnées	Verbe s'il existe	que tu viennes, qui viendra, quand tu voudras, que toi (aussi souvent)
Sint	incises, juxtaposées	Verbe	dit-il
Coord	coordonnées	Coordonnant	et Paul, et gentil, et le mange, ou avec les doigts, voire demain

FIG. 6.8 – *Les syntagmes retenus*

La figure 6.8 dresse la liste des syntagmes retenus. Le *shallow-parser* utilise ce jeu d'étiquettes pour annoter les frontières de constituants à l'exception des incises et juxtapositions qui ne sont jamais reconnues automatiquement mais seulement ajoutées manuellement.

Le corpus de Paris 7 a été entièrement parsé grâce au *shallow-parser* et un tiers, c'est-à-dire environ 10 000 phrases, a été entièrement corrigé pour les frontières de constituants.

Les syntagmes coordonnés

Les syntagmes englobant les coordonnées ne sont pas annotés car leur catégorie est parfois discutable. D'une part on peut coordonner des syntagmes de catégories différentes (i.e. (a), (b), (c)), d'autre part le syntagme coordonné peut contenir des effacements du prédicat (i.e. (d), (e)). En revanche, on peut isoler le syntagme coordonné commençant par un coordonnant (une conjonction ou une virgule) à l'intérieur de tous les syntagmes.

(a) Paul est médecin et fier de l'être.
(b) Paul sait l'âge de Marie et qu'elle ne peut se présenter à ce concours.
(c) Une proposition intéressante et qu'on devrait considérer attentivement.
(d) Paul donne un disque à Marie et un livre à Jean.
(e) Paul va à la piscine le lundi avec Marie et au cinéma le jeudi avec Jean.

Remarquons que le premier conjoint d'une coordonnée peut ne pas avoir tous ses compléments. C'est le cas lors de la mise en facteur d'un complément (*right node raising*).

(a) Paul accepte et Jules refuse ce cadeau.

Le noyau verbal

Il comprend toute la séquence de texte entre l'auxiliaire de temps *être* et *avoir* et le verbe principal. Il contient également les clitiques du verbe (pronoms faibles et particule *ne*).

Nous n'avons pas inclus dans le noyau verbal les compléments du verbe ni les adverbes essentiels ni les circonstants. Cela aurait exigé une annotation discontinue des compléments extraits du groupe verbal traditionnel et une prise de position sur l'existence d'un syntagme verbal.

(a) les actions qu'a mises IBM sur le marché.
(b) Les actionnaires décideront certainement une augmentation de capital.

En (a), le sujet *IBM* précède un complément locatif (*sur le marché*), en (b) l'adverbe *certainement* est postverbal et précède le complément du verbe (*une augmentation de capital*).

Nous n'avons pas non plus annoté les constructions figées et les constructions à verbe support à ce niveau d'annotation.

En revanche, le verbe conjugué avec les auxiliaires de temps et les suites de clitiques ont été regroupés en un syntagme permettant de faciliter les interrogations portant sur les constructions verbales.

(a) Jean [n'en veut]$_{VN}$ plus.
(b) Elle [nous verra]$_{VN}$ bien.
(c) [regarde-moi]$_{VN}$.
(d) Jean [n'est pas encore parti]$_{VN}$.

Les incises et juxtapositions

Nous notons *Sint* (pour *internal sentence*) les phrases sans introducteur (relatif ou subordonnant) qui ne sont ni complétives ni relatives ni coordonnées. Ainsi on notera les parataxes (i.e. (b)), les coordonnées sans conjonction (i.e. (c)), les incises (i.e. (d)) et les discours rapportés (i.e. (f)).

(a) Il exagère, [vraiment il exagère]$_{Sint}$.
(b) Jean est doué, [il réussira]$_{Sint}$.
(c) Jean joue du piano, [son frère du violon]$_{Sint}$.
(d) Paul gagne, [dit Marie]$_{Sint}$, plus de 10 000 francs.
(e) Paul dit : «[je vais bien]$_{Sint}$».

6.2.4 Perspectives

Dans la suite du projet, le corpus va être annoté pour les fonctions syntaxiques. Nous associerons les fonctions aux syntagmes et non aux mots car il est souvent difficile de déterminer la tête des syntagmes qui interviennent dans un rapport de dépendance.

Il s'agira dans un premier temps d'annoter les fonctions de surface et non les fonctions profondes. Une infinitive peut réaliser son sujet hors de la phrase enchâssée comme en (a), un interrogatif peut réaliser la fonction d'objet du

verbe comme en (b), un relatif peut réaliser la fonction de complément de nom comme en (c).

(a) Jean dit à Marie$_i$ de partir(sujet=i) plus tôt.
(b) Jean demande quel film$_j$ j'ai vu(objet=j).
(c) J'ai vu un film$_k$ dont$_k$ je n'aime pas la fin(complément=k).

Pour (a), nous noterons la fonction complément (de *dire*) et non sa fonction profonde (sujet de *partir*). Pour (b), nous noterons la fonction complément (de *voir*) pour *quel film* car les deux termes font partie de la même clause. Pour (c), nous ne noterons pas la fonction de complément de nom pour *dont*.

Pour l'annotation fonctionnelle, nous utiliserons un étiqueteur qui assignera à chaque constituant une fonction syntaxique par correspondance directe ou en appliquant quelques règles contextuelles simples. Les fonctions distinguées lors de cette phase d'annotation sont *sujet*, *complément* et *modifieur*. Ces seules fonctions permettent de marquer les liens de dépendance syntaxique entre les constituants sans noter par ailleurs leurs rôles sémantiques ou casuels. Dans la plupart des cas, ces fonctions sont attribuées par simple correspondance avec les syntagmes (par exemple un groupe prépositionnel enchâssé dans un groupe nominal sera systématiquement annoté comme complément de nom), dans les autres cas, des heuristiques simples seront appliquées (voir tableau 6.9).

Pour déterminer quelques fonctions ambiguës (des groupes nominaux par exemple), le recours à l'utilisation de dictionnaires syntaxiques s'avère nécessaire. En effet, les fonctions complément et sujet sont subordonnées à la présence locale d'un prédicat. Les lexiques du projet FTAG ([Abeillé *et al.*, 1999], [Barrier, 1999]) et les tables du LADL ([Namer & Hathout, 1998] permettront de construire la sous-catégorisation maximale de chaque prédicat et de projeter cette information sur le corpus. Cependant cette condition n'est ni nécessaire ni suffisante. D'une part un verbe transitif peut ne pas sous-catégoriser un groupe nominal postposé ((a), (b))[9], d'autre part un groupe nominal local peut ne pas être le complément d'un verbe transitif (i.e. (c)).

(a) Le ministre mange le soir.
(b) Jean brise la glace.
(c) Les spécialités du boulanger que les enfants mangent.

9. Notons qu'en (b), *la glace* est un complément syntaxique sans avoir de rôle thématique, on peut donc mettre en doute qu'une fonction objet lui soit assignée.

Syntagme	**Fonction**
VN	$\emptyset$
VPinf	complément si introduit par *de* ou *à* ou sans préposition, modifieur sinon
Ssub	complément si commence par *que* ou par un pronom interrogatif, modifieur sinon
NP	sujet, complément ou modifieur ou $\emptyset$ (pour complément de préposition)
AP	complément si frère de VN, modifieur si enchâssé dans un NP
Srel	modifieur
PP	complément ou modifieur
COORD	$\emptyset$
VPpart	modifieur
Sint	modifieur

FIG. 6.9 – *Liste des fonctions syntaxiques*

Par ailleurs, les ambiguïtés de structure rendent complexe l'automatisation de l'assignation des fonctions syntaxiques aux syntagmes. Bien souvent, le recours à des connaissances encyclopédiques, à la pragmatique ou au contexte du discours est nécessaire pour trancher.

Nous pouvons néanmoins indiquer quelques pistes exploitant les seules connaissances syntaxiques pour lever un certain nombre de cas difficiles.

L'annotation en *clusters* des syntagmes nominaux permet de distinguer des adverbiales de temps et de lieu comme modifieurs.

(a) Le ministre mange [chaque dimanche]$_{modif}$.
(b) Le coursier livre [le 3, rue Censier]$_{modif}$.

D'autres types d'annotation sont envisageables : marquage des valences verbales, des liens anaphoriques, du sens, etc. Les annotations déjà disponibles pourront être utiles, mais comme pour les autres types d'annotation, une validation manuelle sera toujours nécessaire en plus d'une phase automatique, pour une garantie de qualité.

Conclusion

Dans ce mémoire, nous avons présenté un corpus annoté syntaxiquement pour le français, pleinement désambiguïsé et validé manuellement. L'annotation comprend l'assignation de traits ou catégories à des unités textuelles, les *mots* dans une première étape, et la représentation des constituances et dépendances articulant des unités syntagmatiques dans une seconde étape. Cette seconde phase de l'annotation de corpus est en cours et soulève encore un certain nombre de problèmes théoriques (comment annoter les dépendances ambiguës par exemple).

L'annotation syntaxique comprend un ensemble de procédures qui ne sont pas sans influence sur la mise en lumière de certains phénomènes linguistiques lors de l'exploitation du corpus.

En morpho-syntaxe, les unités de segmentation, les catégories, les parties du discours ont été choisies dans le but de rendre le plus réutilisable possible le corpus pour différentes études. Ce principe a été suivi également lors de l'annotation en syntagmes. Les choix d'étiquetage des syntagmes et les choix relatifs aux dépendances et constituances ont été réalisés dans le but de fournir une ressource indépendante d'un formalisme linguistique donné.

Mais bien évidemment on ne peut définir des unités de segmentation, des catégories syntaxiques et des principes d'articulation syntaxiques en s'affranchissant de toute théorie linguistique comme le souligne [Blache, 2000]. Pour que le Corpus de Paris 7 soit exploitable par la communauté, il fallait, au delà des choix que nous avons retenus, qu'ils soient expliqués et qu'ils accompagnent le Corpus.

Le corpus de Paris 7 est donc un corpus de *référence* par la taille et par la richesse de l'annotation. La taille d'un million de mots suffit à faire varier les constructions syntaxiques et à produire une liste de mots importante. La richesse de l'annotation et la documentation qui accompagne ce corpus constitue un standard utilisable pour d'autres corpus. Mais le corpus de Paris 7 n'est pas représentatif de tous les usages. Les méthodes d'échantillonnage de

corpus, la volonté de couvrir de nombreux genres et niveaux de langue ont été écartées de cette étude. L'usage que l'on peut faire du corpus de Paris 7 est donc limité à l'interrogation sur les productions syntaxiques, soit pour mettre en lumière des connaissances linguistiques, soit pour entraîner des analyseurs et étiqueteurs automatiques probabilistes.

Il faut donc prendre avec précaution les résultats obtenus à partir de ce corpus, mais ceux des interrogations qu'on a menées semblent transposables à d'autres corpus écrits. Certaines de ces investigations ont confirmé des fréquences et des préférences lexicales bien connues, d'autres ont apporté la lumière sur de nouvelles fréquences et de nouvelles préférences qui devraient être confirmées sur d'autres corpus.

Annexe A

Échantillon du corpus étiqueté – format pour annotateurs après correction (format interne)

Six d'entre eux, seulement blessés, ont pu se réfugier sur l'autre rive de la Nipoué, en Côte-d'Ivoire, avant d'être transférés à Man, dans un hôpital. La tâche des secouristes est immense, faute de moyens matériels et humains. À Danané, le chef du secteur de santé rurale travaille en étroite collaboration avec une équipe très réduite de médecins sans frontières en attendant le renfort imminent de la Croix-rouge nationale et internationale. Sur les trente-cinq mille réfugiés répartis dans la plupart des villages frontaliers, on note une majorité de femmes et d'enfants. Selon des témoignages, les hommes sont systématiquement arrêtés par les soldats libériens craignant d'avoir affaire à des maquisards.

Six	PROmp
d'	P
entre	P
eux	PROmp
,	PONCTW
seulement	ADV

blessés	VKmp
,	PONCTW
ont	VP3p
pu	VKms
se	CL3mp
réfugier	VW
sur	P
l'	Dfs
autre	Afs
rive	NCfs
de	P
la	Dfs
Nipoué	NPfs
,	PONCTW
en	P
Côte-d'Ivoire	NPfs
,	PONCTW
avant_d'	P
être	VW
transférés	VKmp
à	P
Man	NPms
,	PONCTW
dans	P
un	Dms
hôpital	NCms
.	PONCTS
La	Dfs
tâche	NCfs
de	P
les	Dmp
secouristes	NCmp
est	VP3s
immense	Afs
,	PONCTW
faute_de_moyens	ADV
matériels	Amp
et	CC
humains	Amp
.	PONCTS

À	P
Danané	NPms
,	PONCTW
le	Dms
chef	NCms
de	P
le	Dms
secteur	NCms
de	P
santé	NCfs
rurale	Afs
travaille	VP3s
en	P
étroite	Afs
collaboration	NCfs
avec	P
une	Dfs
équipe	NCfs
très	ADV
réduite	Afs
de	P
médecins_sans_frontières	NCmp
en	P
attendant	VG
le	Dms
renfort	NCms
imminent	Ams
de	P
la	Dfs
Croix-rouge	NPfs
nationale	Afs
et	CC
internationale	Afs
.	PONCTS
Sur	P
les	Dmp
trente-cinq_mille	Amp
réfugiés	NCmp
répartis	VKmp
dans	P
la_plupart_de	Dmp

les	Dmp
villages	NCmp
frontaliers	Amp
,	PONCTW
on	CL3ms
note	VP3s
une	Dfs
majorité	NCfs
de	P
femmes	NCfp
et	CC
d'	P
enfants	NCmp
.	PONCTS
Selon	P
de_les	Dmp
témoignages	NCmp
,	PONCTW
les	Dmp
hommes	NCmp
sont	VP3p
systématiquement	ADV
arrêtés	VKmp
par	P
les	Dmp
soldats	NCmp
libériens	Amp
craignant	VG
d'	P
avoir_affaire	VW
à	P
de_les	Dmp
maquisards	NCmp
.	PONCTS

Annexe B

Échantillon du corpus étiqueté – format final

⟨?xml version="1.0" encoding="ISO-8859-1"?⟩
⟨tei.2⟩
⟨teiHeader date.created="25-07-00"
creator="Anne Abeillé, Lionel Clément
and P7 team" type="corpus"⟩

⟨fileDesc⟩
⟨sourcedesc⟩
⟨/sourcedesc⟩

⟨titleStmt⟩
⟨title type="main" lang="FR"⟩Le Monde⟨/title⟩
⟨title type="gmd" lang="FR"⟩Corpus Annoté de P7⟨/title⟩
⟨respStmt⟩
⟨resp⟩sources fournies par⟨/resp⟩
⟨name⟩Linguistic Data Consortium⟨/name⟩
⟨/respStmt⟩
⟨respStmt⟩
⟨resp⟩annoté, encodé et validé par⟨/resp⟩
⟨name⟩P7⟨/name⟩
⟨/respStmt⟩
⟨/titleStmt⟩

⟨publicationStmt⟩
⟨idno type="P7"⟩LMF300⟨/idno⟩
⟨availability status="restricted"⟩

```
            ⟨p⟩accès restreint⟨/p⟩
            ⟨/availability⟩
            ⟨distributor⟩
            ⟨name⟩P7-LATTICE⟨/name⟩
            ⟨address⟩
                  ⟨addrLine⟩
                  UFRL, Université Paris 7
                  Case 7003
                  2, Place Jussieu
                  F-75251 Paris cedex 05
                  ⟨/addrLine⟩
            ⟨/address⟩
            ⟨/distributor⟩
            ⟨/publicationStmt⟩
      ⟨/fileDesc⟩

      ⟨encodingDesc⟩
      ⟨/encodingDesc⟩

      ⟨profiledesc⟩
      ⟨langusage⟩⟨language id="fr"⟩French⟨/language/⟩
      ⟨/profiledesc⟩

      ⟨revisiondesc⟩
      ⟨change⟩
            ⟨date⟩25 Juillet 2000⟨/date⟩
            ⟨respstmt⟩⟨name⟩Lionel Clément⟨/name⟩⟨resp/⟩⟨/respstmt⟩
            ⟨item⟩
                  processing of original corpus files into tei conformance.
            ⟨/item⟩
      ⟨/change⟩
      ⟨/revisiondesc⟩

⟨/teiHeader⟩
⟨text⟩
⟨s⟩
⟨w lemma="six" cat="PRO" subcat="card" mph="mp"⟩Six⟨/w⟩
⟨w lemma="de" cat="P"⟩d'⟨/w⟩
⟨w lemma="entre" cat="P"⟩entre⟨/w⟩
⟨w lemma="eux" cat="PRO" subcat="" mph="3mp"⟩eux⟨/w⟩
⟨w lemma="," cat="PONCT" subcat="W"⟩,⟨/w⟩
```

⟨w lemma="seulement" cat="ADV"⟩seulement⟨/w⟩
⟨w lemma="blesser" cat="V" subcat="" mph="Kmp"⟩blessés⟨/w⟩
⟨w lemma="," cat="PONCT" subcat="W"⟩,⟨/w⟩
⟨w lemma="avoir" cat="V" subcat="" mph="P3p"⟩ont⟨/w⟩
⟨w lemma="pouvoir" cat="V" subcat="" mph="Kms"⟩pu⟨/w⟩
⟨w lemma="se" cat="CL" subcat="refl" mph="3mp"⟩se⟨/w⟩
⟨w lemma="réfugier" cat="V" subcat="" mph="W"⟩réfugier⟨/w⟩
⟨w lemma="sur" cat="P"⟩sur⟨/w⟩
⟨w lemma="le" cat="D" subcat="def" mph="fs"⟩l'⟨/w⟩
⟨w lemma="autre" cat="A" subcat="qual" mph="fs"⟩autre⟨/w⟩
⟨w lemma="rive" cat="N" subcat="C" mph="fs"⟩rive⟨/w⟩
⟨w lemma="de" cat="P"⟩de⟨/w⟩
⟨w lemma="le" cat="D" subcat="def" mph="fs"⟩la⟨/w⟩
⟨w lemma="Nipoué" cat="N" subcat="P" mph="fs"⟩Nipoué⟨/w⟩
⟨w lemma="," cat="PONCT" subcat="W"⟩,⟨/w⟩
⟨w lemma="en" cat="P"⟩en⟨/w⟩
⟨w lemma="Côte-d'Ivoire" cat="N" subcat="P" mph="fs"⟩
    ⟨w cat=""⟩Côte⟨/w⟩
    ⟨w cat="PONCT"⟩-⟨/w⟩
    ⟨w cat=""⟩d'⟨/w⟩
    ⟨w cat=""⟩Ivoire⟨/w⟩
⟨/w⟩
⟨w lemma="," cat="PONCT" subcat="W"⟩,⟨/w⟩
⟨w lemma="avant de" cat="P"⟩
    ⟨w cat="P"⟩avant⟨/w⟩
    ⟨w cat="P"⟩d'⟨/w⟩
⟨/w⟩
⟨w lemma="être" cat="V" subcat="" mph="W"⟩être⟨/w⟩
⟨w lemma="transférer" cat="V" subcat="" mph="Kmp"⟩transférés⟨/w⟩
⟨w lemma="à" cat="P"⟩à⟨/w⟩
⟨w lemma="Man" cat="N" subcat="P" mph="ms"⟩man⟨/w⟩
⟨w lemma="," cat="PONCT" subcat="W"⟩,⟨/w⟩
⟨w lemma="dans" cat="P"⟩dans⟨/w⟩
⟨w lemma="un" cat="D" subcat="ind" mph="ms"⟩un⟨/w⟩
⟨w lemma="hôpital" cat="N" subcat="C" mph="ms"⟩hôpital⟨/w⟩
⟨w lemma="." cat="PONCT" subcat="S"⟩.⟨/w⟩
⟨/s⟩
⟨s⟩
⟨w lemma="le" cat="D" subcat="def" mph="fs"⟩La⟨/w⟩
⟨w lemma="tâche" cat="N" subcat="C" mph="fs"⟩tâche⟨/w⟩
⟨w lemma="de" cat="P"⟩des⟨/w⟩

⟨w lemma="le" cat="D" subcat="def" mph="mp"/⟩
⟨w lemma="secouriste" cat="N" subcat="C" mph="mp"⟩secouristes⟨/w⟩
⟨w lemma="être" cat="V" subcat="" mph="P3s"⟩est⟨/w⟩
⟨w lemma="immense" cat="A" subcat="qual" mph="fs"⟩immense⟨/w⟩
⟨w lemma="," cat="PONCT" subcat="W"⟩,⟨/w⟩
⟨w lemma="faute de moyens" cat="ADV"⟩
    ⟨w cat="N"⟩faute⟨/w⟩
    ⟨w cat="P"⟩de⟨/w⟩
    ⟨w cat="N"⟩moyens⟨/w⟩
⟨/w⟩
⟨w lemma="matériel" cat="A" subcat="qual" mph="mp"⟩matériels⟨/w⟩
⟨w lemma="et" cat="C" subcat="C"⟩et⟨/w⟩
⟨w lemma="humain" cat="A" subcat="qual" mph="mp"⟩humains⟨/w⟩
⟨w lemma="." cat="PONCT" subcat="S"⟩.⟨/w⟩
⟨/s⟩
⟨s⟩
⟨w lemma="à" cat="P"⟩À⟨/w⟩
⟨w lemma="Danané" cat="N" subcat="P" mph="ms"⟩Danané⟨/w⟩
⟨w lemma="," cat="PONCT" subcat="W"⟩,⟨/w⟩
⟨w lemma="le" cat="D" subcat="def" mph="ms"⟩le⟨/w⟩
⟨w lemma="chef" cat="N" subcat="C" mph="ms"⟩chef⟨/w⟩
⟨w lemma="de" cat="P"⟩du⟨/w⟩
⟨w lemma="le" cat="D" subcat="def" mph="ms"/⟩
⟨w lemma="secteur" cat="N" subcat="C" mph="ms"⟩secteur⟨/w⟩
⟨w lemma="de" cat="P"⟩de⟨/w⟩
⟨w lemma="santé" cat="N" subcat="C" mph="fs"⟩santé⟨/w⟩
⟨w lemma="rural" cat="A" subcat="qual" mph="fs"⟩rurale⟨/w⟩
⟨w lemma="travailler" cat="V" subcat="" mph="P3s"⟩travaille⟨/w⟩
⟨w lemma="en" cat="P"⟩en⟨/w⟩
⟨w lemma="étroit" cat="A" subcat="qual" mph="fs"⟩étroite⟨/w⟩
⟨w lemma="collaboration" cat="N" subcat="C" mph="fs"⟩collaboration⟨/w⟩
⟨w lemma="avec" cat="P"⟩avec⟨/w⟩
⟨w lemma="un" cat="D" subcat="ind" mph="fs"⟩une⟨/w⟩
⟨w lemma="équipe" cat="N" subcat="C" mph="fs"⟩équipe⟨/w⟩
⟨w lemma="très" cat="ADV"⟩très⟨/w⟩
⟨w lemma="réduit" cat="A" subcat="qual" mph="fs"⟩réduite⟨/w⟩
⟨w lemma="de" cat="P"⟩de⟨/w⟩
⟨w lemma="médecin sans frontière" cat="N" subcat="C" mph="mp"⟩
    ⟨w cat="N"⟩médecins⟨/w⟩
    ⟨w cat="P"⟩sans⟨/w⟩

```
        ⟨w cat="N"⟩frontières⟨/w⟩
⟨/w⟩
⟨w lemma="en" cat="P"⟩en⟨/w⟩
⟨w lemma="attendre" cat="V" subcat="" mph="G"⟩attendant⟨/w⟩
⟨w lemma="le" cat="D" subcat="def" mph="ms"⟩le⟨/w⟩
⟨w lemma="renfort" cat="N" subcat="C" mph="ms"⟩renfort⟨/w⟩
⟨w lemma="imminent" cat="A" subcat="qual" mph="ms"⟩imminent⟨/w⟩
⟨w lemma="de" cat="P"⟩de⟨/w⟩
⟨w lemma="le" cat="D" subcat="def" mph="fs"⟩la⟨/w⟩
⟨w lemma="Croix-Rouge" cat="N" subcat="P" mph="fs"⟩
        ⟨w cat="N"⟩Croix⟨/w⟩
        ⟨w cat="PONCT"⟩-⟨/w⟩
        ⟨w cat="A"⟩rouge⟨/w⟩
⟨/w⟩
⟨w lemma="national" cat="A" subcat="qual" mph="fs"⟩nationale⟨/w⟩
⟨w lemma="et" cat="C" subcat="C"⟩et⟨/w⟩
⟨w lemma="international" cat="A" subcat="qual" mph="fs"⟩internationale⟨/w⟩
⟨w lemma="." cat="PONCT" subcat="S"⟩.⟨/w⟩
⟨/s⟩
⟨s⟩
⟨w lemma="sur" cat="P"⟩Sur⟨/w⟩
⟨w lemma="le" cat="D" subcat="def" mph="mp"⟩les⟨/w⟩
⟨w lemma="trente-cinq mille" cat="A" subcat="card" mph="mp"⟩
        ⟨w cat=""⟩trente⟨/w⟩
        ⟨w cat="PONCT"⟩-⟨/w⟩
        ⟨w cat=""⟩cinq⟨/w⟩
        ⟨w cat=""⟩mille⟨/w⟩
⟨/w⟩
⟨w lemma="réfugié" cat="N" subcat="C" mph="mp"⟩réfugiés⟨/w⟩
⟨w lemma="répartir" cat="V" subcat="" mph="Kmp"⟩répartis⟨/w⟩
⟨w lemma="dans" cat="P"⟩dans⟨/w⟩
⟨w lemma="la plupart de" cat="D" subcat="" mph="mp"⟩
        ⟨w cat=""⟩la⟨/w⟩
        ⟨w cat=""⟩plupart⟨/w⟩
        ⟨w cat=""⟩des⟨/w⟩
⟨/w⟩
⟨w lemma="le" cat="D" subcat="def" mph="mp"/⟩
⟨w lemma="village" cat="N" subcat="C" mph="mp"⟩villages⟨/w⟩
⟨w lemma="frontalier" cat="A" subcat="qual" mph="mp"⟩frontaliers⟨/w⟩
⟨w lemma="," cat="PONCT" subcat="W"⟩,⟨/w⟩
⟨w lemma="il" cat="CL" subcat="suj" mph="3ms"⟩on⟨/w⟩
```

⟨w lemma="noter" cat="V" subcat="" mph="P3s"⟩note⟨/w⟩
⟨w lemma="un" cat="D" subcat="ind" mph="fs"⟩une⟨/w⟩
⟨w lemma="majorité" cat="N" subcat="C" mph="fs"⟩majorité⟨/w⟩
⟨w lemma="de" cat="P"⟩de⟨/w⟩
⟨w lemma="femme" cat="N" subcat="C" mph="fp"⟩femmes⟨/w⟩
⟨w lemma="et" cat="C" subcat="C"⟩et⟨/w⟩
⟨w lemma="de" cat="P"⟩d'⟨/w⟩
⟨w lemma="enfant" cat="N" subcat="C" mph="mp"⟩enfants⟨/w⟩
⟨w lemma="." cat="PONCT" subcat="S"⟩.⟨/w⟩
⟨/s⟩
⟨s⟩
⟨w lemma="selon" cat="P"⟩Selon⟨/w⟩
⟨w lemma="un" cat="D" subcat="ind" mph="mp"⟩
⟨w cat="D"⟩des⟨/w⟩
⟨w cat="D"/⟩
⟨/w⟩
⟨w lemma="témoignage" cat="N" subcat="C" mph="mp"⟩témoignages⟨/w⟩
⟨w lemma="," cat="PONCT" subcat="W"⟩,⟨/w⟩
⟨w lemma="le" cat="D" subcat="def" mph="mp"⟩les⟨/w⟩
⟨w lemma="homme" cat="N" subcat="C" mph="mp"⟩hommes⟨/w⟩
⟨w lemma="être" cat="V" subcat="" mph="P3p"⟩sont⟨/w⟩
⟨w lemma="systématiquement" cat="ADV"⟩systématiquement⟨/w⟩
⟨w lemma="arrêter" cat="V" subcat="" mph="Kmp"⟩arrêtés⟨/w⟩
⟨w lemma="par" cat="P"⟩par⟨/w⟩
⟨w lemma="le" cat="D" subcat="def" mph="mp"⟩les⟨/w⟩
⟨w lemma="soldat" cat="N" subcat="C" mph="mp"⟩soldats⟨/w⟩
⟨w lemma="libérien" cat="A" subcat="qual" mph="mp"⟩libériens⟨/w⟩
⟨w lemma="craindre" cat="V" subcat="" mph="G"⟩craignant⟨/w⟩
⟨w lemma="de" cat="P"⟩d'⟨/w⟩
⟨w lemma="avoir affaire" cat="V" subcat="" mph="W"⟩
⟨w cat=""⟩avoir⟨/w⟩
⟨w cat=""⟩affaire⟨/w⟩
⟨/w⟩
⟨w lemma="à" cat="P"⟩à⟨/w⟩
⟨w lemma="un" cat="D" subcat="ind" mph="mp"⟩
⟨w cat="D"⟩de⟨/w⟩
⟨w cat="D"⟩les⟨/w⟩
⟨/w⟩
⟨w lemma="maquisard" cat="N" subcat="C" mph="mp"⟩maquisards⟨/w⟩
⟨w lemma="." cat="PONCT" subcat="S"⟩.⟨/w⟩
⟨/s⟩ ⟨/text⟩ ⟨/tei.2⟩

Annexe C

Échantillon du corpus annoté en constituants (avant correction)

⟨SENT nb="1000"⟩
⟨w lemma="six" cat="PRO" subcat="card" mph="mp"⟩Six⟨/w⟩
  ⟨PP⟩
    ⟨w lemma="de" cat="P" ⟩d'⟨/w⟩
    ⟨/PP⟩
  ⟨PP⟩
    ⟨w lemma="entre" cat="P" ⟩entre⟨/w⟩
    ⟨w lemma="eux" cat="PRO" subcat="3mp" ⟩eux⟨/w⟩
    ⟨w lemma="," cat="PONCT" subcat="W" ⟩,⟨/w⟩
    ⟨w lemma="seulement" cat="ADV" ⟩seulement⟨/w⟩
    ⟨/PP⟩
  ⟨VPpart⟩
    ⟨w lemma="blesser" cat="V" subcat="Kmp" ⟩blessés⟨/w⟩
    ⟨w lemma="," cat="PONCT" subcat="W" ⟩,⟨/w⟩
    ⟨/VPpart⟩
  ⟨VN⟩
    ⟨w lemma="avoir" cat="V" subcat="P3p" ⟩ont⟨/w⟩
    ⟨w lemma="pouvoir" cat="V" subcat="Kms" ⟩pu⟨/w⟩
    ⟨w lemma="se" cat="CL" subcat="refl" mph="3mp"⟩se⟨/w⟩
    ⟨VPinf⟩
      ⟨w lemma="réfugier" cat="V" subcat="W" ⟩réfugier⟨/w⟩
      ⟨PP⟩

```
          ⟨w lemma="sur" cat="P" ⟩sur⟨/w⟩
          ⟨NP⟩
            ⟨w lemma="le" cat="D" subcat="def" mph="fs"⟩l'⟨/w⟩
            ⟨w lemma="autre" cat="A" subcat="qual" mph="fs"⟩autre⟨/w⟩
            ⟨w lemma="rive" cat="N" subcat="C" mph="fs"⟩rive⟨/w⟩
            ⟨/NP⟩
          ⟨/PP⟩
        ⟨/VPinf⟩
      ⟨PP⟩
        ⟨w lemma="de" cat="P" ⟩de⟨/w⟩
        ⟨NP⟩
          ⟨w lemma="le" cat="D" subcat="def" mph="fs"⟩la⟨/w⟩
          ⟨w lemma="Nipoué" cat="N" subcat="P" mph="fs"⟩Nipoué⟨/w⟩
          ⟨/NP⟩
        ⟨/PP⟩
      ⟨/VN⟩
        ⟨w lemma="," cat="PONCT" subcat="W" ⟩,⟨/w⟩
    ⟨PP⟩
      ⟨w lemma="en" cat="P" ⟩en⟨/w⟩
      ⟨NP⟩
        ⟨w lemma="Côte-d'Ivoire" cat="N" subcat="P" mph="fs"⟩
        ⟨w ⟩Côte⟨/w⟩
        ⟨w catint="PONCT"⟩-⟨/w⟩
        ⟨w ⟩d'⟨/w⟩
        ⟨w ⟩Ivoire⟨/w⟩
⟨/w⟩
        ⟨/NP⟩
      ⟨/PP⟩
        ⟨w lemma="," cat="PONCT" subcat="W" ⟩,⟨/w⟩
    ⟨VPinf⟩
      ⟨VN⟩
      ⟨w lemma="avant de" cat="P" ⟩avant_d'⟨/w⟩
      ⟨w lemma="être" cat="V" subcat="W" ⟩être⟨/w⟩
      ⟨w lemma="transférer" cat="V" subcat="Kmp" ⟩transférés⟨/w⟩
        ⟨/VN⟩
      ⟨PP⟩
        ⟨w lemma="à" cat="P" ⟩à⟨/w⟩
        ⟨NP⟩
          ⟨w lemma="Man" cat="N" subcat="P" mph="ms"⟩man⟨/w⟩
          ⟨/NP⟩
        ⟨/PP⟩
```

```
      ⟨/VPinf⟩
        ⟨w lemma="," cat="PONCT" subcat="W" ⟩,⟨/w⟩
    ⟨PP⟩
      ⟨w lemma="dans" cat="P" ⟩dans⟨/w⟩
      ⟨NP⟩
        ⟨w lemma="un" cat="D" subcat="ms" ⟩un⟨/w⟩
        ⟨w lemma="hôpital" cat="N" subcat="C" mph="ms"⟩hôpital⟨/w⟩
        ⟨w lemma="." cat="PONCT" subcat="S" ⟩.⟨/w⟩
        ⟨/NP⟩
      ⟨/PP⟩
    ⟨/SENT⟩
⟨SENT nb="1001"⟩
⟨NP⟩
⟨w lemma="le" cat="D" subcat="def" mph="fs"⟩La⟨/w⟩
⟨w lemma="tâche" cat="N" subcat="C" mph="fs"⟩tâche⟨/w⟩
      ⟨/NP⟩
    ⟨PP⟩
      ⟨w lemma="de" cat="P" ⟩des⟨/w⟩
      ⟨NP⟩
        ⟨w lemma="le" cat="D" subcat="def" mph="mp"⟩⟨/w⟩
        ⟨w lemma="secouriste" cat="N" subcat="C" mph="mp"⟩secouristes⟨/w⟩
        ⟨/NP⟩
      ⟨/PP⟩
    ⟨VN⟩
      ⟨w lemma="être" cat="V" subcat="P3s" ⟩est⟨/w⟩
      ⟨w lemma="immense" cat="A" subcat="qual" mph="fs"⟩immense⟨/w⟩
      ⟨w lemma="," cat="PONCT" subcat="W" ⟩,⟨/w⟩
      ⟨w lemma="faute de moyens" cat="ADV" ⟩
      ⟨w catint="N"⟩faute⟨/w⟩
      ⟨w catint="P"⟩de⟨/w⟩
      ⟨w catint="N"⟩moyens⟨/w⟩
⟨/w⟩
⟨w lemma="matériel" cat="A" subcat="qual" mph="mp"⟩matériels⟨/w⟩
      ⟨/VN⟩
    ⟨COORD⟩
      ⟨w lemma="et" cat="C" subcat="C" ⟩et⟨/w⟩
      ⟨w lemma="humain" cat="A" subcat="qual" mph="mp"⟩humains⟨/w⟩
      ⟨/COORD⟩
        ⟨w lemma="." cat="PONCT" subcat="S" ⟩.⟨/w⟩
    ⟨/SENT⟩
⟨SENT nb="1002"⟩
```

```
⟨PP⟩
⟨w lemma="à" cat="P" ⟩À⟨/w⟩
    ⟨NP⟩
      ⟨w lemma="Danané" cat="N" subcat="P" mph="ms"⟩Danané⟨/w⟩
      ⟨/NP⟩
    ⟨/PP⟩
      ⟨w lemma="," cat="PONCT" subcat="W" ⟩,⟨/w⟩
  ⟨NP⟩
    ⟨w lemma="le" cat="D" subcat="def" mph="ms"⟩le⟨/w⟩
    ⟨w lemma="chef" cat="N" subcat="C" mph="ms"⟩chef⟨/w⟩
    ⟨/NP⟩
  ⟨PP⟩
    ⟨w lemma="de" cat="P" ⟩du⟨/w⟩
    ⟨NP⟩
      ⟨w lemma="le" cat="D" subcat="def" mph="ms"⟩⟨/w⟩
      ⟨w lemma="secteur" cat="N" subcat="C" mph="ms"⟩secteur⟨/w⟩
      ⟨/NP⟩
    ⟨/PP⟩
  ⟨PP⟩
    ⟨w lemma="de" cat="P" ⟩de⟨/w⟩
    ⟨NP⟩
      ⟨w lemma="santé" cat="N" subcat="C" mph="fs"⟩santé⟨/w⟩
      ⟨w lemma="rural" cat="A" subcat="qual" mph="fs"⟩rurale⟨/w⟩
      ⟨/NP⟩
    ⟨/PP⟩
  ⟨VN⟩
    ⟨w lemma="travailler" cat="V" subcat="P3s" ⟩travaille⟨/w⟩
    ⟨/VN⟩
  ⟨PP⟩
    ⟨w lemma="en" cat="P" ⟩en⟨/w⟩
    ⟨w lemma="étroit" cat="Afs" ⟩étroite⟨/w⟩
    ⟨NP⟩
      ⟨w lemma="collaboration" cat="N" subcat="C" mph="fs"⟩
        collaboration⟨/w⟩
      ⟨/NP⟩
    ⟨/PP⟩
  ⟨PP⟩
    ⟨w lemma="avec" cat="P" ⟩avec⟨/w⟩
    ⟨NP⟩
      ⟨w lemma="un" cat="D" subcat="ind" mph="fs"⟩une⟨/w⟩
```

```
        ⟨w lemma="équipe" cat="N" subcat="C" mph="fs"⟩équipe⟨/w⟩
        ⟨AP⟩
          ⟨w lemma="très" cat="ADV" ⟩très⟨/w⟩
          ⟨w lemma="réduit" cat="A" subcat="qual" mph="fs"⟩réduite⟨/w⟩
          ⟨/AP⟩
        ⟨/NP⟩
      ⟨/PP⟩
    ⟨PP⟩
      ⟨w lemma="de" cat="P" ⟩de⟨/w⟩
      ⟨NP⟩
        ⟨w lemma="médecin sans frontière" cat="N" subcat="C" mph="mp"⟩
        ⟨w catint="N"⟩médecins⟨/w⟩
        ⟨w catint="P"⟩sans⟨/w⟩
        ⟨w catint="N"⟩frontières⟨/w⟩
⟨/w⟩
        ⟨/NP⟩
      ⟨/PP⟩
    ⟨VPinf⟩
      ⟨VN⟩
      ⟨w lemma="en" cat="P" ⟩en⟨/w⟩
      ⟨w lemma="attendre" cat="V" subcat="G" ⟩attendant⟨/w⟩
        ⟨/VN⟩
      ⟨NP⟩
        ⟨w lemma="le" cat="D" subcat="def" mph="ms"⟩le⟨/w⟩
        ⟨w lemma="renfort" cat="N" subcat="C" mph="ms"⟩renfort⟨/w⟩
        ⟨w lemma="imminent" cat="A" subcat="qual" mph="ms"⟩imminent⟨/w⟩
        ⟨/NP⟩
      ⟨/VPinf⟩
    ⟨PP⟩
      ⟨w lemma="de" cat="P" ⟩de⟨/w⟩
      ⟨NP⟩
        ⟨w lemma="le" cat="D" subcat="def" mph="fs"⟩la⟨/w⟩
        ⟨w lemma="Croix-Rouge" cat="N" subcat="P" mph="fs"⟩
        ⟨w catint="N"⟩Croix⟨/w⟩
        ⟨w catint="PONCT"⟩-⟨/w⟩
        ⟨w catint="A"⟩rouge⟨/w⟩
⟨/w⟩
⟨w lemma="national" cat="A" subcat="qual" mph="fs"⟩nationale⟨/w⟩
        ⟨/NP⟩
      ⟨/PP⟩
    ⟨COORD⟩
```

```
    ⟨w lemma="et" cat="C" subcat="C" ⟩et⟨/w⟩
    ⟨w lemma="international" cat="A" subcat="qual" mph="fs"⟩internationale⟨/w⟩
    ⟨/COORD⟩
      ⟨w lemma="." cat="PONCT" subcat="S" ⟩;⟨/w⟩
  ⟨/SENT⟩
⟨SENT nb="1003"⟩
⟨PP⟩
⟨w lemma="sur" cat="P" ⟩Sur⟨/w⟩
    ⟨NP⟩
      ⟨w lemma="le" cat="D" subcat="def" mph="mp"⟩les⟨/w⟩
      ⟨w lemma="trente-cinq mille" cat="A" subcat="card" mph="mp"⟩
      ⟨w ⟩trente⟨/w⟩
      ⟨w catint="PONCT"⟩-⟨/w⟩
      ⟨w ⟩cinq⟨/w⟩
      ⟨w ⟩mille⟨/w⟩
⟨/w⟩
⟨w lemma="réfugié" cat="N" subcat="C" mph="mp"⟩réfugiés⟨/w⟩
      ⟨/NP⟩
    ⟨/PP⟩
  ⟨VPpart⟩
    ⟨w lemma="répartir" cat="V" subcat="Kmp" ⟩répartis⟨/w⟩
    ⟨PP⟩
      ⟨w lemma="dans" cat="P" ⟩dans⟨/w⟩
      ⟨NP⟩
        ⟨w lemma="la plupart de" cat="D" subcat="mp" ⟩
        ⟨w ⟩la⟨/w⟩
        ⟨w ⟩plupart⟨/w⟩
⟨/w⟩
⟨w ⟩des⟨/w⟩
          ⟨NP⟩
            ⟨w lemma="le" cat="D" subcat="def" mph="mp"⟩⟨/w⟩
            ⟨w lemma="village" cat="N" subcat="C" mph="mp"⟩villages⟨/w⟩
            ⟨w lemma="frontalier" cat="A" subcat="qual" mph="mp"⟩
              frontaliers⟨/w⟩
            ⟨/NP⟩
          ⟨/NP⟩
      ⟨/PP⟩
    ⟨/VPpart⟩
      ⟨w lemma="," cat="PONCT" subcat="W" ⟩,⟨/w⟩
  ⟨VN⟩
    ⟨w lemma="il" cat="CL" subcat="suj" mph="3ms"⟩on⟨/w⟩
```

```
    ⟨w lemma="noter" cat="V" subcat="P3s" ⟩note⟨/w⟩
    ⟨/VN⟩
  ⟨NP⟩
    ⟨w lemma="un" cat="D" subcat="ind" mph="fs"⟩une⟨/w⟩
    ⟨w lemma="majorité" cat="N" subcat="C" mph="fs"⟩majorité⟨/w⟩
    ⟨/NP⟩
  ⟨PP⟩
    ⟨w lemma="de" cat="P" ⟩de⟨/w⟩
    ⟨NP⟩
      ⟨w lemma="femme" cat="N" subcat="C" mph="fp"⟩femmes⟨/w⟩
      ⟨/NP⟩
    ⟨/PP⟩
  ⟨COORD⟩
    ⟨w lemma="et" cat="C" subcat="C" ⟩et⟨/w⟩
    ⟨PP⟩
      ⟨w lemma="de" cat="P" ⟩d'⟨/w⟩
      ⟨NP⟩
        ⟨w lemma="enfant" cat="N" subcat="C" mph="mp"⟩enfants⟨/w⟩
        ⟨w lemma="." cat="PONCT" subcat="S" ⟩.⟨/w⟩
        ⟨/NP⟩
      ⟨/PP⟩
    ⟨/COORD⟩
  ⟨/SENT⟩
⟨SENT nb="1004"⟩
⟨PP⟩
⟨w lemma="selon" cat="P" ⟩Selon⟨/w⟩
    ⟨NP⟩
      ⟨w lemma="de les" cat="D" subcat="mp" ⟩
      ⟨w ⟩des⟨/w⟩
      ⟨w ⟩⟨/w⟩
⟨/w⟩
⟨w lemma="témoignage" cat="N" subcat="C" mph="mp"⟩témoignages⟨/w⟩
      ⟨/NP⟩
    ⟨/PP⟩
      ⟨w lemma="," cat="PONCT" subcat="W" ⟩,⟨/w⟩
  ⟨NP⟩
    ⟨w lemma="le" cat="D" subcat="def" mph="mp"⟩les⟨/w⟩
    ⟨w lemma="homme" cat="N" subcat="C" mph="mp"⟩hommes⟨/w⟩
    ⟨/NP⟩
  ⟨VN⟩
    ⟨w lemma="être" cat="V" subcat="P3p" ⟩sont⟨/w⟩
```

```
    ⟨w lemma="systématiquement" cat="ADV" ⟩systématiquement⟨/w⟩
    ⟨w lemma="arrêter" cat="V" subcat="Kmp" ⟩arrêtés⟨/w⟩
    ⟨/VN⟩
  ⟨PP⟩
    ⟨w lemma="par" cat="P" ⟩par⟨/w⟩
    ⟨NP⟩
      ⟨w lemma="le" cat="D" subcat="def" mph="mp"⟩les⟨/w⟩
      ⟨w lemma="soldat" cat="N" subcat="C" mph="mp"⟩soldats⟨/w⟩
      ⟨w lemma="libérien" cat="A" subcat="qual" mph="mp"⟩libériens⟨/w⟩
      ⟨/NP⟩
    ⟨/PP⟩
  ⟨VPpart⟩
    ⟨w lemma="craindre" cat="V" subcat="G" ⟩craignant⟨/w⟩
    ⟨VPinf⟩
      ⟨VN⟩
      ⟨w lemma="de" cat="P" ⟩d'⟨/w⟩
      ⟨w lemma="avoir affaire" cat="V" subcat="W" ⟩
      ⟨w ⟩avoir⟨/w⟩
      ⟨w ⟩affaire⟨/w⟩
⟨/w⟩
        ⟨/VN⟩
      ⟨/VPinf⟩
    ⟨/VPpart⟩
  ⟨NP⟩
    ⟨w lemma="à" cat="P" ⟩à⟨/w⟩
    ⟨w lemma="de les" cat="D" subcat="mp" ⟩
    ⟨w ⟩de⟨/w⟩
    ⟨w ⟩les⟨/w⟩
⟨/w⟩
⟨w lemma="maquisard" cat="N" subcat="C" mph="mp"⟩maquisards⟨/w⟩
⟨w lemma="." cat="PONCT" subcat="S" ⟩.⟨/w⟩
    ⟨/NP⟩
  ⟨/SENT⟩
```

Annexe D

Échantillon du corpus annoté en constituants (après correction)

⟨SENT nb="1000"⟩
⟨NP⟩
⟨w lemma="six" cat="PRO" subcat="card" mph="mp"⟩Six⟨/w⟩
⟨PP⟩
⟨w lemma="d'entre" cat="P" ⟩
⟨w lemma="de" cat="P" ⟩d'⟨/w⟩
⟨w lemma="entre" cat="P" ⟩entre⟨/w⟩
⟨/w⟩
⟨NP⟩
⟨w lemma="eux" cat="PRO" subcat="3mp" ⟩eux⟨/w⟩
⟨/NP⟩
⟨/PP⟩
⟨/NP⟩
⟨VPpart⟩
⟨w lemma="," cat="PONCT" subcat="W" ⟩,⟨/w⟩
⟨w lemma="seulement" cat="ADV" ⟩seulement⟨/w⟩
⟨w lemma="blesser" cat="V" subcat="Kmp" ⟩blessés⟨/w⟩
⟨w lemma="," cat="PONCT" subcat="W" ⟩,⟨/w⟩
⟨/VPpart⟩
⟨VN⟩
⟨w lemma="avoir" cat="V" subcat="P3p" ⟩ont⟨/w⟩
⟨w lemma="pouvoir" cat="V" subcat="Kms" ⟩pu⟨/w⟩

```
    ⟨/VN⟩
  ⟨VPinf⟩
    ⟨VN⟩
      ⟨w lemma="se" cat="CL" subcat="refl" mph="3mp"⟩se⟨/w⟩
      ⟨w lemma="réfugier" cat="V" subcat="W" ⟩réfugier⟨/w⟩
      ⟨/VN⟩
    ⟨PP⟩
      ⟨w lemma="sur" cat="P" ⟩sur⟨/w⟩
      ⟨NP⟩
        ⟨w lemma="le" cat="D" subcat="def" mph="fs"⟩l'⟨/w⟩
        ⟨w lemma="autre" cat="A" subcat="qual" mph="fs"⟩autre⟨/w⟩
        ⟨w lemma="rive" cat="N" subcat="C" mph="fs"⟩rive⟨/w⟩
        ⟨PP⟩
          ⟨w lemma="de" cat="P" ⟩de⟨/w⟩
          ⟨NP⟩
            ⟨w lemma="le" cat="D" subcat="def" mph="fs"⟩la⟨/w⟩
            ⟨w lemma="Nipoué" cat="N" subcat="P" mph="fs"⟩Nipoué⟨/w⟩
            ⟨/NP⟩
          ⟨/PP⟩
        ⟨/NP⟩
      ⟨/PP⟩
        ⟨w lemma="," cat="PONCT" subcat="W" ⟩,⟨/w⟩
    ⟨PP⟩
      ⟨w lemma="en" cat="P" ⟩en⟨/w⟩
      ⟨NP⟩
        ⟨w lemma="Côte-d'Ivoire" cat="N" subcat="P" mph="fs"⟩
          ⟨w ⟩Côte⟨/w⟩
          ⟨w cat="PONCT"⟩-⟨/w⟩
          ⟨w cat="P"⟩d'⟨/w⟩
          ⟨w cat="N"⟩Ivoire⟨/w⟩
        ⟨/w⟩
        ⟨/NP⟩
      ⟨/PP⟩
  ⟨/VPinf⟩
    ⟨w lemma="," cat="PONCT" subcat="W" ⟩,⟨/w⟩
⟨VPinf⟩
  ⟨w lemma="avant de" cat="P" ⟩avant_d'⟨/w⟩
  ⟨VN⟩
    ⟨w lemma="être" cat="V" subcat="W" ⟩être⟨/w⟩
    ⟨w lemma="transférer" cat="V" subcat="Kmp" ⟩transférés⟨/w⟩
    ⟨/VN⟩
```

```
    ⟨PP⟩
      ⟨w lemma="à" cat="P" ⟩à⟨/w⟩
      ⟨NP⟩
        ⟨w lemma="Man" cat="N" subcat="P" mph="ms"⟩man⟨/w⟩
        ⟨/NP⟩
      ⟨/PP⟩
        ⟨w lemma="," cat="PONCT" subcat="W" ⟩,⟨/w⟩
    ⟨PP⟩
      ⟨w lemma="dans" cat="P" ⟩dans⟨/w⟩
      ⟨NP⟩
        ⟨w lemma="un" cat="D" subcat="ms" ⟩un⟨/w⟩
        ⟨w lemma="hôpital" cat="N" subcat="C" mph="ms"⟩hôpital⟨/w⟩
        ⟨/NP⟩
      ⟨/PP⟩
    ⟨/VPinf⟩
      ⟨w lemma="." cat="PONCT" subcat="S" ⟩.⟨/w⟩
  ⟨/SENT⟩
⟨SENT nb="1001"⟩
  ⟨NP⟩
    ⟨w lemma="le" cat="D" subcat="def" mph="fs"⟩La⟨/w⟩
    ⟨w lemma="tâche" cat="N" subcat="C" mph="fs"⟩tâche⟨/w⟩
    ⟨PP⟩
      ⟨w lemma="de" cat="P" ⟩des⟨/w⟩
      ⟨NP⟩
        ⟨w lemma="le" cat="D" subcat="def" mph="mp"⟩⟨/w⟩
        ⟨w lemma="secouriste" cat="N" subcat="C" mph="mp"⟩secouristes⟨/w⟩
        ⟨/NP⟩
      ⟨/PP⟩
    ⟨/NP⟩
  ⟨VN⟩
    ⟨w lemma="être" cat="V" subcat="P3s" ⟩est⟨/w⟩
    ⟨/VN⟩
  ⟨AP⟩
    ⟨w lemma="immense" cat="A" subcat="qual" mph="fs"⟩immense⟨/w⟩
    ⟨/AP⟩
      ⟨w lemma="," cat="PONCT" subcat="W" ⟩,⟨/w⟩
  ⟨PP⟩
    ⟨w lemma="faute de" cat="P" ⟩
      ⟨w cat="N"⟩faute⟨/w⟩
      ⟨w cat="P"⟩de⟨/w⟩
    ⟨/w⟩
```

```
    ⟨NP⟩
      ⟨w cat="N" subcat="C" mph="fp"⟩moyens⟨/w⟩
      ⟨w lemma="matériel" cat="A" subcat="qual" mph="mp"⟩matériels⟨/w⟩
      ⟨COORD⟩
        ⟨w lemma="et" cat="C" subcat="C" ⟩et⟨/w⟩
        ⟨w lemma="humain" cat="A" subcat="qual" mph="mp"⟩humains⟨/w⟩
        ⟨/COORD⟩
      ⟨/NP⟩
    ⟨/PP⟩
      ⟨w lemma="." cat="PONCT" subcat="S" ⟩.⟨/w⟩
  ⟨/SENT⟩
⟨SENT nb="1002"⟩
  ⟨PP⟩
    ⟨w lemma="à" cat="P" ⟩À⟨/w⟩
    ⟨NP⟩
      ⟨w lemma="Danané" cat="N" subcat="P" mph="ms"⟩Danané⟨/w⟩
      ⟨/NP⟩
    ⟨/PP⟩
      ⟨w lemma="," cat="PONCT" subcat="W" ⟩,⟨/w⟩
  ⟨NP⟩
    ⟨w lemma="le" cat="D" subcat="def" mph="ms"⟩le⟨/w⟩
    ⟨w lemma="chef" cat="N" subcat="C" mph="ms"⟩chef⟨/w⟩
    ⟨PP⟩
      ⟨w lemma="de" cat="P" ⟩du⟨/w⟩
      ⟨NP⟩
        ⟨w lemma="le" cat="D" subcat="def" mph="ms"⟩⟨/w⟩
        ⟨w lemma="secteur" cat="N" subcat="C" mph="ms"⟩secteur⟨/w⟩
        ⟨PP⟩
          ⟨w lemma="de" cat="P" ⟩de⟨/w⟩
          ⟨NP⟩
            ⟨w lemma="santé" cat="N" subcat="C" mph="fs"⟩santé⟨/w⟩
            ⟨w lemma="rural" cat="A" subcat="qual" mph="fs"⟩rurale⟨/w⟩
            ⟨/NP⟩
          ⟨/PP⟩
        ⟨/NP⟩
      ⟨/PP⟩
    ⟨/NP⟩
  ⟨VN⟩
    ⟨w lemma="travailler" cat="V" subcat="P3s" ⟩travaille⟨/w⟩
    ⟨/VN⟩
```

```
⟨PP⟩
  ⟨w lemma="en" cat="P" ⟩en⟨/w⟩
  ⟨w lemma="étroit" cat="Afs" ⟩étroite⟨/w⟩
  ⟨NP⟩
    ⟨w lemma="collaboration" cat="N" subcat="C" mph="fs"⟩collaboration⟨/w⟩
    ⟨/NP⟩
  ⟨/PP⟩
⟨PP⟩
  ⟨w lemma="avec" cat="P" ⟩avec⟨/w⟩
  ⟨NP⟩
    ⟨w lemma="un" cat="D" subcat="ind" mph="fs"⟩une⟨/w⟩
    ⟨w lemma="équipe" cat="N" subcat="C" mph="fs"⟩équipe⟨/w⟩
    ⟨AP⟩
      ⟨w lemma="très" cat="ADV" ⟩très⟨/w⟩
      ⟨w lemma="réduit" cat="A" subcat="qual" mph="fs"⟩réduite⟨/w⟩
      ⟨/AP⟩
    ⟨PP⟩
      ⟨w lemma="de" cat="P" ⟩de⟨/w⟩
      ⟨NP⟩
        ⟨w lemma="médecin sans frontière" cat="N" subcat="C" mph="mp"⟩
          ⟨w cat="N"⟩médecins⟨/w⟩
          ⟨w cat="P"⟩sans⟨/w⟩
          ⟨w cat="N"⟩frontières⟨/w⟩
        ⟨/w⟩
        ⟨/NP⟩
      ⟨/PP⟩
    ⟨/NP⟩
  ⟨/PP⟩
⟨VPpart⟩
  ⟨w lemma="en" cat="P" ⟩en⟨/w⟩
  ⟨VN⟩
    ⟨w lemma="attendre" cat="V" subcat="G" ⟩attendant⟨/w⟩
    ⟨/VN⟩
  ⟨NP⟩
    ⟨w lemma="le" cat="D" subcat="def" mph="ms"⟩le⟨/w⟩
    ⟨w lemma="renfort" cat="N" subcat="C" mph="ms"⟩renfort⟨/w⟩
    ⟨AP⟩
      ⟨w lemma="imminent" cat="A" subcat="qual" mph="ms"⟩imminent⟨/w⟩
      ⟨/AP⟩
    ⟨PP⟩
      ⟨w lemma="de" cat="P" ⟩de⟨/w⟩
```

```
          ⟨NP⟩
            ⟨w lemma="le" cat="D" subcat="def" mph="fs"⟩la⟨/w⟩
            ⟨w lemma="Croix-Rouge" cat="N" subcat="P" mph="fs"⟩
              ⟨w cat="N"⟩Croix⟨/w⟩
              ⟨w cat="PONCT"⟩-⟨/w⟩
              ⟨w cat="A"⟩rouge⟨/w⟩
            ⟨/w⟩
            ⟨AP⟩
              ⟨w lemma="national" cat="A" subcat="qual" mph="fs"⟩nationale⟨/w⟩
              ⟨/AP⟩
            ⟨COORD⟩
              ⟨w lemma="et" cat="C" subcat="C" ⟩et⟨/w⟩
              ⟨AP⟩
                ⟨w lemma="international" cat="A" subcat="qual" mph="fs"⟩
                  internationale⟨/w⟩
                ⟨/AP⟩
              ⟨/COORD⟩
            ⟨/NP⟩
          ⟨/PP⟩
        ⟨/NP⟩
      ⟨/VPpart⟩
        ⟨w lemma="." cat="PONCT" subcat="S" ⟩.⟨/w⟩
  ⟨/SENT⟩
⟨SENT nb="1003"⟩
  ⟨PP⟩
    ⟨w lemma="sur" cat="P" ⟩Sur⟨/w⟩
    ⟨NP⟩
      ⟨w lemma="le" cat="D" subcat="def" mph="mp"⟩les⟨/w⟩
      ⟨w lemma="trente-cinq mille" cat="A" subcat="card" mph="mp"⟩
      ⟨w ⟩trente⟨/w⟩
      ⟨w cat="PONCT"⟩-⟨/w⟩
      ⟨w ⟩cinq⟨/w⟩
      ⟨w ⟩mille⟨/w⟩
⟨/w⟩
⟨w lemma="réfugié" cat="N" subcat="C" mph="mp"⟩réfugiés⟨/w⟩
        ⟨VPpart⟩
          ⟨w lemma="répartir" cat="V" subcat="Kmp" ⟩répartis⟨/w⟩
          ⟨PP⟩
            ⟨w lemma="dans" cat="P" ⟩dans⟨/w⟩
            ⟨NP⟩
              ⟨w lemma="la plupart de" cat="D" subcat="mp" ⟩
```

```
                ⟨w ⟩la⟨/w⟩
                ⟨w ⟩plupart⟨/w⟩
                ⟨w ⟩des⟨/w⟩
              ⟨/w⟩
              ⟨NP⟩
                ⟨w lemma="le" cat="D" subcat="def" mph="mp"⟩⟨/w⟩
                ⟨w lemma="village" cat="N" subcat="C" mph="mp"⟩villages⟨/w⟩
                ⟨w lemma="frontalier" cat="A" subcat="qual" mph="mp"⟩
                  frontaliers⟨/w⟩
                ⟨/NP⟩
              ⟨/NP⟩
            ⟨/PP⟩
          ⟨/VPpart⟩
        ⟨/NP⟩
      ⟨/PP⟩
        ⟨w lemma="," cat="PONCT" subcat="W" ⟩,⟨/w⟩
    ⟨/VN⟩
      ⟨w lemma="il" cat="CL" subcat="suj" mph="3ms"⟩on⟨/w⟩
      ⟨w lemma="noter" cat="V" subcat="P3s" ⟩note⟨/w⟩
      ⟨/VN⟩
    ⟨NP⟩
      ⟨w lemma="un" cat="D" subcat="ind" mph="fs"⟩une⟨/w⟩
      ⟨w lemma="majorité" cat="N" subcat="C" mph="fs"⟩majorité⟨/w⟩
      ⟨PP⟩
        ⟨w lemma="de" cat="P" ⟩de⟨/w⟩
        ⟨NP⟩
          ⟨w lemma="femme" cat="N" subcat="C" mph="fp"⟩femmes⟨/w⟩
          ⟨/NP⟩
        ⟨/PP⟩
      ⟨COORD⟩
        ⟨w lemma="et" cat="C" subcat="C" ⟩et⟨/w⟩
        ⟨PP⟩
          ⟨w lemma="de" cat="P" ⟩d'⟨/w⟩
          ⟨NP⟩
            ⟨w lemma="enfant" cat="N" subcat="C" mph="mp"⟩enfants⟨/w⟩
            ⟨/NP⟩
          ⟨/PP⟩
        ⟨/COORD⟩
      ⟨/NP⟩
        ⟨w lemma="." cat="PONCT" subcat="S" ⟩.⟨/w⟩
    ⟨/SENT⟩
```

```
⟨SENT nb="1004"⟩
  ⟨PP⟩
    ⟨w lemma="selon" cat="P" ⟩Selon⟨/w⟩
    ⟨NP⟩
      ⟨w lemma="de les" cat="D" subcat="mp" ⟩
        ⟨w ⟩des⟨/w⟩
        ⟨w ⟩⟨/w⟩
      ⟨/w⟩
      ⟨w lemma="témoignage" cat="N" subcat="C" mph="mp"⟩témoignages⟨/w⟩
      ⟨/NP⟩
    ⟨/PP⟩
      ⟨w lemma="," cat="PONCT" subcat="W" ⟩,⟨/w⟩
  ⟨NP⟩
    ⟨w lemma="le" cat="D" subcat="def" mph="mp"⟩les⟨/w⟩
    ⟨w lemma="homme" cat="N" subcat="C" mph="mp"⟩hommes⟨/w⟩
    ⟨/NP⟩
  ⟨VN⟩
    ⟨w lemma="être" cat="V" subcat="P3p" ⟩sont⟨/w⟩
    ⟨w lemma="systématiquement" cat="ADV" ⟩systématiquement⟨/w⟩
    ⟨w lemma="arrêter" cat="V" subcat="Kmp" ⟩arrêtés⟨/w⟩
    ⟨/VN⟩
  ⟨PP⟩
    ⟨w lemma="par" cat="P" ⟩par⟨/w⟩
    ⟨NP⟩
      ⟨w lemma="le" cat="D" subcat="def" mph="mp"⟩les⟨/w⟩
      ⟨w lemma="soldat" cat="N" subcat="C" mph="mp"⟩soldats⟨/w⟩
      ⟨AP⟩
        ⟨w lemma="libérien" cat="A" subcat="qual" mph="mp"⟩libériens⟨/w⟩
        ⟨/AP⟩
      ⟨VPpart⟩
        ⟨w lemma="craindre" cat="V" subcat="G" ⟩craignant⟨/w⟩
        ⟨VPinf⟩
          ⟨w lemma="de" cat="P" ⟩d'⟨/w⟩
          ⟨VN⟩
            ⟨w lemma="avoir affaire" cat="V" subcat="W" ⟩
              ⟨w ⟩avoir⟨/w⟩
              ⟨w ⟩affaire⟨/w⟩
            ⟨/w⟩
            ⟨/VN⟩
          ⟨PP⟩
            ⟨w lemma="à" cat="P" ⟩à⟨/w⟩
```

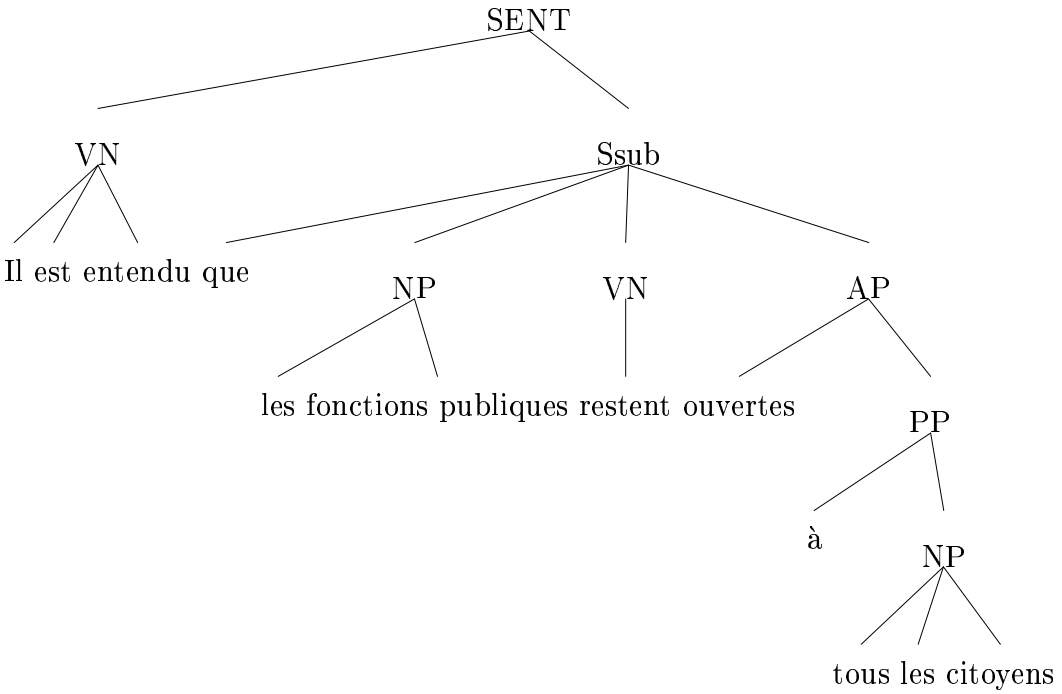
```
            ⟨NP⟩
              ⟨w lemma="de les" cat="D" subcat="mp" ⟩
                ⟨w ⟩de⟨/w⟩
                ⟨w ⟩les⟨/w⟩
              ⟨/w⟩
            ⟨w lemma="maquisard" cat="N" subcat="C" mph="mp"⟩
              maquisards⟨/w⟩
              ⟨/NP⟩
            ⟨/PP⟩
          ⟨/VPinf⟩
        ⟨/VPpart⟩
      ⟨/NP⟩
    ⟨/PP⟩
      ⟨w lemma="." cat="PONCT" subcat="S" ⟩.⟨/w⟩
⟨/SENT⟩
```

Annexe E

Échantillon du corpus annoté en constituants (correspondance avec l'arbre syntaxique)

```
<SENT>
  <VN>
    <w lemma="il" cat="CL" subcat="suj" mph="3ms">Il</w>
    <w lemma="être" cat="V" subcat="" mph="P3s">est</w>
    <w lemma="entendre" cat="V" subcat="" mph="Kms">entendu</w>
  </VN>
  <Ssub>
    <w lemma="que" cat="C" subcat="S">que</w>
    <NP>
      <w lemma="le" cat="D" subcat="def" mph="fp">les</w>
      <w lemma="fonction publique" cat="N" subcat="C" mph="fp">
        <w cat="N">fonctions</w>
        <w cat="A">publiques</w>
      </w>
    </NP>
    <VN>
      <w lemma="rester" cat="V" subcat="" mph="P3p">restent</w>
    </VN>
    <AP>
```

```
      <w lemma="ouvert" cat="A" subcat="qual" mph="fp">ouvertes</w>
      <w lemma="à" cat="P">à</w>
      <NP>
        <w lemma="tout" cat="A" subcat="ind" mph="mp">tous</w>
        <w lemma="le" cat="D" subcat="def" mph="mp">les</w>
        <w lemma="citoyen" cat="N" subcat="C" mph="mp">citoyens</w>
      </NP>
    </AP>
    <w lemma="." cat="PONCT" subcat="S">.</w>
  </Ssub>
</SENT>
```



Annexe F

Etiquettes internes

Les étiquettes internes sont codées suivant la catégorie, la sous-catégorie syntaxique et les traits morphologiques.

Format des étiquettes

Les étiquettes respectent toutes le même format :

1. La catégorie en majuscules : ADJ, ADV, N, ...
2. La sous-catégorie syntaxique en minuscules : ind, card, ...
3. Pour les verbes un codes indiquant le mode et le temps : P, S, I, ...
4. Un tiret, les personnes, genre et nombre
5. Un tiret, le genre et la personne du possédé pour les possessifs.

Note : 4 et 5 sont optionnels

Exemple :

- DETposs-2fs-p désigne un déterminant possessif deuxième personne du féminin singulier. Le nombre du possesseur est pluriel. Exemple : Votre dans : “Je vous indique votre chambre, Mademoiselle”
- VP-3s Désigne un verbe au présent de l'indicatif 3e personne du singulier

Description des traits employés

Les traits morphologiques et syntaxiques sont :

- la personne
  1, 2 ou 3
- le genre
  masculin ou féminin
- le nombre
  singulier ou pluriel
- le nombre du possesseur
  singulier ou pluriel
- le temps
  présent, passé-simple, imparfait, futur, passe
- le mode
  indicatif, conditionnel, subjonctif, imperatif, participe, infinitif

Nous notons que ces deux derniers traits sont codés grâce à une unique étiquette : P, S, I, Y, W ...

Description des catégories

Nous dressons ici la liste des catégories et sous-catégories accompagnées d'exemples.

Liste des étiquettes

122 étiquettes

ADJECTIF

Ams	il est INTELLIGENT
	le PETIT chien de la voisine
	du fait que le candidat est déclaré SEUL sur la liste
	le XIXe siècle
	Ce vélo est MIEN

	Cet individu est tout AUTRE
	Un stylo NEUF
	QUEL est ce vainqueur?
Amp	Ces enfants sont déjà GRANDS
	dans les CINQ ans à venir
	TOUS les ans
Afs	une EVIDENTE défaite
	une TELLE défaite
	C'est la prudence MÊME
Afp	trois disquettes ENDOMMAGEES
	les TROIS disquettes
	TOUTES les trois
	TELLES les trois Grâces

Cardinal (seulement pour NEUF)

ACfp	les NEUF différences
ACmp	les NEUF stylos

Indéfini

seulement pour: autre, autres, certain, certaine, différentes,
différents, divers, diverses, même, mêmes, quelconque, seul, seule, seules, seuls

AIms	Un AUTRE individu
AImp	Les MÊMES ordinateurs
AIfs	une CERTAINE angoisse
AIfp	les MÊMES filles

ADVERBE

ADV	il est PRESQUE midi
	le ministre est ENSUITE nommé
	il s'avance LENTEMENT DESSOUS
	pour ne prendre QUE des exemples
	il me demande COMBIEN ils sont
	COMBIEN sont ils?
	ce garçon-CI

BIEN des années ont passé
BEAUCOUP de ces étudiants sont diplomés
RIEN moins que trente francs
il en veut D_AVANTAGE
DEPUIS, il est parti
OÙ vas-tu?
QUAND pars-tu?
Jean ne boit PLUS

Exclamatif

ADVE
seulement pour **que**, **comme** lorsqu'il s'agit de l'adverbe exclamatif
QUE oui!
COMME elle est belle!

CONJONCTION

Conjonction de coordination

CC
La table AINSI_QUE les chaises sont en orme.
Le combustible est mélangé à l'air ET injecté dans la chambre

Conjonction de subordination

CS
QUAND Marie travaille, elle n'écoute personne
Le député n'a pas fait de déclaration LORSQU'il a été battu
Jean est gros ALORS_QUE Marie ne l'est pas.
Il est aussi petit QUE Marie.
Je pense QU elle a tort.
Je me demande SI elle m'écoute.

DÉTERMINANT

Dms	LE ministre a pris la parole
	CE vélo est-il a vendre?
	UN ordinateur est en panne
	j'ai acheté DE_LE pain
	QUEL pain veut-tu?
Dmp	TROIS éléphants sont traqués par les chasseurs
	Marie garde LES enfants
	CES ustensiles sont pratiques
	la grogne parmi CERTAINS anciens cadres
Dfs	Il y a ZÉRO femme dans ce comité
	LA linguistique, c'est passionnant
	Je regarde CETTE peinture
	QUELLE meilleure méthode que celle du puzzle?
Dfp	Jean demande QUELLES filles sont candidates
	TROIS amies sont venues
	LES ordinateurs sont allumés
	CES peintures sont laides
	Tu achèteras DE_LES prunes
	DIFFÉRENTES malversations
DPms	Je n'ai pas D' argent
	Je n'ai plus DE monnaie
DPfs	Jean boit DE_LA bière

Exclamatif

seulement pour quel, quels, quelle, quelles lorsque le déterminant est Exclamatif

DEms	QUEL homme que cet homme là !
DEmp	QUELS
DEfs	QUELLE
DEfp	QUELLES

INTERJECTION

I	HÉLAS!

NOM

Nom Commun

NCfp	Trois VACHES
	Les FRANÇAISES
	Les PME
NCfs	Une VACHE
NCmp	TAUREAUX
	MILLIONS
NCms	TAUREAU
	MILLIARD
	le FRANÇAIS

Nom propre

NPfp	les ALPES
NPfs	MARIE
	la FRANCE
	le SNCF
NPmp	les GOBELINS
	Les ÉTATS-UNIS
NPms	PAUL
	le JAPON

PRÉPOSITION

P	DEPUIS midi je viens DE sortir QUANT_À vous, vous restez elle est partie VOICI 3 ans Le sel est SUR la table Il jette un oeil AVANT_DE partir

PRONOM

Clitiques

CL1ms	je, -moi, j'
CLO1ms, CLR1ms	me, m'
CL2ms	tu
CLO2ms, CLR2ms	te, -toi, t'
CL3ms	il, on, en, y, se, s', c', ce, le, l', lui
CLS1mp, CLO1mp, CLR1mp	nous
CLS2mp, CLO2mp, CLR2mp	vous
CL3mp	ils, se, en , y, s', les, leur
CL1fs	je, me, moi, j', m'
CL2fs	tu, te, toi, t'
CL3fs	elle, se, s', la, l', lui
CLS1fp, CLO1fp, CLR1fp	nous
CLS2fp, CLO2fp, CLR2fp	vous
CL3fp	elles, se, s', les, leur

Pronoms indéfinis, cardinaux et possessifs

PROms	L'UN de_les notres, TOUT va bien, le MIEN, le TIEN, le SIEN, le NOTRE, le VOTRE, le LEUR
PROmp	les AUTRES, j'en veux DEUX, les MIENS, les TIENS, les SIENS, les NOTRES, les VOTRES, les LEURS

	BEAUCOUP de ces étudiants
	PLUSIEURS sont arrivés
PROfs	j'en veux UNE, la MIENNE, la TIENNE,
	la SIENNE, la NOTRE,
	la VOTRE, la LEUR
PROfp	DEUX sont arrivées, elles sont TOUTES là,
les MIENNEs, les TIENNES,	
	les SIENNES, les NOTRES, les VOTRES,
	les LEURS, CERTAINES sont là

Pronoms personnels forts

PRO1ms	moi
PRO2ms	toi
PRO3ms	lui, soi
PRO1mp	nous
PRO2mp	vous
PRO3mp	eux
PRO1fp	nous
PRO2fp	vous
PRO3fp	elles
PRO1fs	moi
PRO2fs	toi
PRO3fs	elle, soi

Pronoms relatifs

Seulement pour dont, lequel, qui, que, quoi lorsqu'ils sont pronoms relatifs.

PROR1ms	qui que dont
PROR2ms	qui que dont
PROR3ms	qui que quoi dont où lequel
PROR1mp	qui que dont
PROR2mp	qui que dont
PROR3mp	qui que dont où lesquels
PROR1fs	qui que dont
PROR2fs	qui que dont

PROR3fs	qui que quoi dont où laquelle
PROR1fp	qui que dont
PROR2fp	qui que dont
PROR3fp	qui que dont où lesquelles

Pronoms interrogatifs

PROIms	lequel, qui, que, quoi
PROImp	lesquels
PROIfs	laquelle
PROIfp	lesquelles

MOTS ETRANGERS

ET

VERBES

Pour les formes composées, nous notons les traits morphologiques du ou des auxiliaires puis du participe.
Exemple :
il:CL3ms a:VP3s été:VKms mangé:VKms
j:CL1ms eusse:VT1s mangé:VKms
j:CL1ms ai:VP1s eu:VKms mangé:VKms

Présent conditionnel

VC1s	MANGERAIS
VC2s	MANGERAIS
VC3s	MANGERAIT

VC1p	MANGERIONS
VC2p	MANGERIEZ
VC3p	MANGERAIENT

Futur indicatif

VF1s	MANGERAI
VF2s	MANGERAS
VF3s	MANGERA
VF1p	MANGERONS
VF2p	MANGEREZ
VF3p	MANGERONT

Participe présent

VG	MANGEANT

Imparfait indicatif

VI1s	MANGEAIS
VI2s	MANGEAIS
VI3s	MANGEAIT
VI1p	MANGIONS
VI2p	MANGIEZ
VI3p	MANGEAIENT

Passé simple indicatif

VJ1s	MANGEAI
VJ3s	MANGEA

VJ3p	MANGÈRENT

Participe passé

VKms	MANGÉ
VKmp	MANGÉS
VKfs	MANGÉE
VKfp	MANGÉES

Présent indicatif

VP1s	MANGE
VP2s	MANGES
VP3s	MANGE, VOICI
VP1p	MANGEONS
VP2p	MANGEZ
VP3p	MANGENT

Présent subjonctif

VS1s	MANGE
VS2s	MANGES
VS3s	MANGE
VS1p	MANGIONS
VS2p	MANGIEZ
VS3p	MANGENT

Imparfait subjonctif

VT1s	MANGEASSE, EUSSE

VT3s	MANGEÂT
VT1p	MANGEASSIONS, CHOISISSIONS
VT3p	MANGEASSENT

Infinitif

VW	MANGER

Impératif

VY1p	MANGEONS
VY2s	MANGE
VY2p	MANGEZ

PRÉFIXE

PREF	ANTI, ARCHI, PRÉ, POST...

Annexe G

Liste des formes fléchies ambiguës sur le lemme

choient	choir/choyer
crue	croire/croître
crues	croire/croître
crus	croire/croître
crûmes	croire/croître
crût	croire/croître
crûtes	croire/croître
dépeignaient	dépeigner/dépeindre
dépeignais	dépeigner/dépeigner
dépeignais	dépeindre/dépeindre
dépeignait	dépeigner/dépeindre
dépeignant	dépeigner/dépeindre
dépeigne	dépeigner/dépeigner
dépeigne	dépeindre/dépeindre
dépeignent	dépeigner/dépeigner
dépeignent	dépeindre/dépeindre
dépeignes	dépeigner/dépeindre
dépeignez	dépeigner/dépeigner
dépeignez	dépeindre/dépeindre
dépeigniez	dépeigner/dépeigner
dépeigniez	dépeindre/dépeindre
dépeignions	dépeigner/dépeigner
dépeignions	dépeindre/dépeindre
dépeignons	dépeigner/dépeigner
dépeignons	dépeindre/dépeindre
faille	faillir/falloir
faut	faillir/falloir
fils	fil/fils
fois	foi/fois
fondaient	fonder/fondre
fondais	fonder/fondre
fondais	fondre/fonder
fondait	fondre/fonder
fondant	fondre/fonder
fonde	fonder/fondre
fondent	fonder/fondre
fondes	fonder/fondre
fondez	fonder/fondre
fondiez	fonder/fondre
fondions	fonder/fondre
fondons	fonder/fondre
frais	frai/frais
gris-gris	gri-gri/gris-gris
gueuses	gueuse/gueux
héroïnes	héros/héroïne
loupiotes	loupiot/loupiote
maraudes	maraud/maraude
moulaient	moudre/mouler
moulais	moudre/mouler
moulait	moudre/mouler
moulant	moudre/mouler
moule	moudre/mouler
moulent	moudre/mouler

moules	moudre/mouler
moulez	moudre/mouler
mouliez	moudre/mouler
moulions	moudre/mouler
moulons	moudre/mouler
ouvraient	ouvrer/ouvrir
ouvrais	ouvrer/ouvrir
ouvrait	ouvrer/ouvrir
ouvrant	ouvrer/ouvrir
ouvre	ouvrer/ouvrir
ouvrent	ouvrer/ouvrir
ouvres	ouvrer/ouvrir
ouvrez	ouvrer/ouvrir
ouvriez	ouvrer/ouvrir
ouvrions	ouvrer/ouvrir
ouvrons	ouvrer/ouvrir
peignaient	peigner/peindre
peignais	peigner/peindre
peignait	peigner/peindre
peignant	peigner/peindre
peigne	peigner/peindre
peignent	peigner/peindre
peignes	peigner/peindre
peignez	peigner/peindre
peigniez	peigner/peindre
peignions	peigner/peindre
peignons	peigner/peindre
plu	plaire/pleuvoir
plurent	plaire/pleuvoir
plussent	plaire/pleuvoir
plut	plaire/pleuvoir
plût	plaire/pleuvoir
pu	paître/pouvoir
pâtissaient	pâtir/pâtisser
pâtissais	pâtir/pâtisser
pâtissait	pâtir/pâtisser
pâtissant	pâtir/pâtisser
pâtisse	pâtir/pâtisser
pâtissent	pâtir/pâtisser
pâtisses	pâtir/pâtisser
pâtissez	pâtir/pâtisser
pâtissiez	pâtir/pâtisser
pâtissions	pâtir/pâtisser
pâtissons	pâtir/pâtisser
recouvraient	recouvrer/recouvrir
recouvrais	recouvrer/recouvrir
recouvrait	recouvrer/recouvrir
recouvrant	recouvrer/recouvrir
recouvre	recouvrer/recouvrir
recouvrent	recouvrer/recouvrir
recouvres	recouvrer/recouvrir
recouvrez	recouvrer/recouvrir
recouvriez	recouvrer/recouvrir
recouvrions	recouvrer/recouvrir
recouvrons	recouvrer/recouvrir
refondaient	refonder/refondre
refondais	refonder/refondre
refondait	refonder/refondre
refondant	refonder/refondre
refonde	refonder/refondre
refondent	refonder/refondre
refondes	refonder/refondre
refondez	refonder/refondre
refondiez	refonder/refondre
refondions	refonder/refondre
refondons	refonder/refondre
remoulaient	remoudre/remouler
remoulais	remoudre/remouler
remoulait	remoudre/remouler
remoulant	remoudre/remouler
remoule	remoudre/remouler
remoulent	remoudre/remouler
remoules	remoudre/remouler
remoulez	remoudre/remouler
remouliez	remoudre/remouler
remoulions	remoudre/remouler
remoulons	remoudre/remouler
repeignaient	repeigner/repeindre
repeignais	repeigner/repeindre
repeignait	repeigner/repeindre
repeignant	repeigner/repeindre
repeigne	repeigner/repeindre

repeignent	repeigner/repeindre
repeignes	repeigner/repeindre
repeignez	repeigner/repeindre
repeigniez	repeigner/repeindre
repeignions	repeigner/repeindre
repeignons	repeigner/repeindre
suis	suivre/être
surfera	surfaire/surfer
surferai	surfaire/surfer
surferaient	surfaire/surfer
surferais	surfaire/surfer
surferait	surfaire/surfer
surferas	surfaire/surfer
surferez	surfaire/surfer
surferiez	surfaire/surfer
surferions	surfaire/surfer
surferons	surfaire/surfer
surferont	surfaire/surfer
tapissaient	tapir/tapisser
tapissais	tapir/tapisser
tapissait	tapir/tapisser
tapissant	tapir/tapisser
tapisse	tapir/tapisser
tapissent	tapir/tapisser
tapisses	tapir/tapisser
tapissez	tapir/tapisser
tapissiez	tapir/tapisser
tapissions	tapir/tapisser
tapissons	tapir/tapisser
vernissaient	vernir/vernisser
vernissais	vernir/vernisser
vernissait	vernir/vernisser
vernissant	vernir/vernisser
vernisse	vernir/vernisser
vernissent	vernir/vernisser
vernisses	vernir/vernisser
vernissez	vernir/vernisser
vernissiez	vernir/vernisser
vernissions	vernir/vernisser
vernissons	vernir/vernisser
vers	ver/vers

Table des figures

Index

– U –

– V –

– Z –

Références bibliographiques

Anne Abeillé (1991). *Une grammaire lexicalisée d'arbres adjoints pour le français: application à l'analyse automatique.* Thèse de doctorat, Université Paris 7.

Anne Abeillé (1993). *Les nouvelles syntaxes.* Paris : Armand Colin.

Anne Abeillé (1996). Corpus et syntaxe : l'apport de l'informatique linguistique. *Revue française de linguistique appliquée*, (pp. 7-24).

Anne Abeillé, Marie-Hélène Candito, et Alexandra Kinyon (1999). FTAG: current status and parsing scheme. In *VEXTAL'99* Venise.

Anne Abeillé et Lionel Clément (1997). *Désambiguïsation morphosyntaxique; 1 Les mots simples; 2 Les mots composés.* TALANA, Université Paris 7.

Anne Abeillé et Lionel Clément (1999). A tagged reference corpus for French. In *Proceedings LINC-EACL'99* Bergen.

Anne Abeillé, Lionel Clément, et Alexandra Kinyon (2000). *Building and using a syntactically annoted corpora.* Dordrech : Kluwer Academic Publishers.

Anne Abeillé, Lionel Clément, et Alexandra Kinyon (2001). Building a Treebank for French. *Treebanks.* Kluwer Academic Publishers.

Anne Abeillé, Lionel Clément, et Rodrigo Reyes (1998). TALaNa Annotated Corpus: the first results. In *Proceedings First Conference on Linguistic Resources* (pp. 992-999). Grenade.

Anne Abeillé, Lionel Clément, et François Toussenel (2000). *Corpus Le Monde - Annotations en constituants - Guide pour les correcteurs.* Université Paris 7.

Steven Abney (1990). *Principle-based Parsing*, Parsing by Chunks. Kluwer Academic Publishers, berwick edition.

Steven Abney (1996). Chunk Stylebook. http://whorf.sfs.nphil.uni-tuebingen.de/abney/Papers.html.

Gilles Adda, Joseph Mariani, Patrick Paroubek, Martin Rajman, et Josette

Lecomte (1999). L'action GRACE d'évaluation de l'assignation de parties du discours pour le français. *Langues*, 2-2, 119-130.

Antoine Arnauld et Claude Lancelot (1676). *Grammaire générale et raisonnée*. Paris : Allia, 1997 edition.

Michel Arrivé, Françoise Gadet, et Michel Galmiche (1986). *La grammaire d'aujourd'hui*. Paris : Flammarion.

Sébastien Barrier (1999). Repérage et classification de valences verbales, expériences avec FTAG. Mémoire de DEA, Université Paris 7.

Douglas Biber (1991). *Variation across speech and writing*. Cambridge : Cambridge University Press.

Mireille Bilger (2000). *Corpus - Méthodologie et applications linguistiques*. Paris : Honoré Champion.

Mireille Bilger (2000). *Linguistique sur corpus*. Presses Universitaires de Perpignan.

Philippe Blache (2000). *A quoi sert l'annotation syntaxique de corpus ?*, in Corpus, méthodologie et applications linguistiques (éd. Mireille Bilger). Honoré Champion : Paris.

Claire Blanche-Benveniste (1996). De l'utilité du corpus linguistique. *Revue Française de linguistique appliquée*, 12, 25-42.

Claire Blanche-Benveniste (1997). *Approches de la langue parlés en français*. Paris : Ophrys.

Claire Blanche-Benveniste (1999). Constitution et exploitation d'un grand corpus. *Revue Française de linguistique appliquée*, 12, 25-42.

Claire Blanche-Benveniste, Christine Rouget, et Karel van den Eynde (1991). *Le Français Parlé. Etudes Grammaticales*. Paris : Editions du CNRS.

Didier Bourigault (1992). Surface Grammatical analysis for the extraction of terminological noun phrases. In *Proceedings COLING'92* (pp. 977-981).

Thorsten Brants, Wojciech Skut, et Hans Uszkorcit (2001). Syntactic Annotation of a German Newpaper Corpus. In *Treebanks*. Kluwer Academic Publishers.

Éric Brill (1992). A simple rule-based part of speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing* : ACL.

Éric Brill (1993). *A Corpus-Based Approach to Language Learning*. Thèse de PhD, Université de Pennsylvanie, Computer and Information Science.

Ted Briscoe, John Carroll, et Antonio Sanfilippo (1998). Parser evaluation : A survey and a new proposal. In *Proceedings First Conference on Linguistic Resources* (pp. 447-455). Granada.

Lou Burnard et Michael Sperberg-McQueen (1996). *La TEI simplifiée : une*

introduction au codage des textes électroniques en vue de leur échange, cahiers gutemberg edition. n. 24.

Marie-Hélène Candito (1996). A principle-based hierarchical representation of LTAGs. In *Proceedings 19th COLING* Kopenhagen.

Marie-Hélène Candito (1999). *Représentation hiérarchique de grammaires lexicalisées: application au français et à l'italien.* Thèse de doctorat, Université Paris 7.

Marie-Hélène Candito et Sylvain Kahane (1998). Can the TAG derivation tree represent a semantic graph?
An answer in the light of Meaning-Text Theory. In *TAG+4*.

Nina Catach (1984). *Les listes orthographiques de base.* Collection Recherche. Paris : Nathan.

Noam Chomsky (1957). *Syntactic structures.* Janua linguarum. The Hague : Mouton.

Noam Chomsky et George A. Miller (1968). *L'analyse formelle des langues naturelles.* Paris : Mouton Gauthier-Villars.

Oliver Christ (1994). A Modular and Flexible Architecture for an Integred Corpus Query System. In *COMPLEX'94* (pp. 23-32). Budapest.

Kenneth Church (1988). A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. In *2nd ANLP Conference* (pp. 136-143). Austin.

Lionel Clément (2000). XLFG - Une plate-forme de développement de Grammaires Lexicales Fonctionnelles. In *TALN 2000* (pp. 405-408). Lausanne.

Lionel Clément et Alexandra Kinyon (2000). Chunking, marking and searching a morpho-syntactically annotated corpus for French. In *ACIDCA* Monastir, Tunisia.

Lionel Clément et Alexandra Kinyon (2001). XLFG-an LFG parsing scheme for French. In *LFG 2001* Hong Kong.

Anne Condamines, Cécile Fabre, et Marie-Paule Péry-Woodley (éds.) (1999). *Corpus et TAL : Pour une réflexion méthodologique*, Cargèse, Corse.

Dan Cristea, Nancy Ide, et Laurent Romary (1998). Marking up multiple views of a text : discourse and reference. In *Proceedings First Conference on Linguistic Resources* (pp. 483-488). Granada.

Anne Daladier (1999). Auxiliation des noms d'action. *Langages*, 135, 87-10.

Jacques Damourette et Édouard Pichon (19111927). *Des mots à la pensée.* Éditions d'Artrey.

Laurence Danlos (éd.) (1988). *Les locutions figées*, in 90. Paris : Larousse.

Laurence Danlos (1998). G-TAG : un formalisme lexicalisé pour la génération

de textes inspiré de TAG. *Traitement Automatique des Langues (TAL)*, 39.

Ferdinand de Saussure (1912). *Cours de linguistique générale.* Paris : Payot, 1997 edition.

Alin Deutsh, Mary Fernandez, Daniela Florescu, et Alon Levy ans Dan Suciu (1999). A query language for XML. In *Proceedings of the International Wold Wide Web Conference*, in 31 (pp. 1155-1169).

Bonnie Jean Dorr (1991). Principle-Based Parsing for Machine Translation. In R. C. Berwick, S. P. Abney, et C. Tenny (éds.), *Principle-based Parsing: Computation and Psycholinguistics.* Norwell, MA : Kluwer Academic Publishers.

Jean Dubois (1994). *Dictionnaire de linguistique et des sciences du langage.* Larousse.

Nelson Francis et Henry Kučera (Revised 1989). *Manual of Information to accompany a Standard Corpus of Present-day Edited American English, for use with Digital Computers.* Providence, Rhode Island : Brown University.

Eric Gaussier, Gregory Grefenstette, David Hull, et Claude Roux (2000). Recherche d'information en français et Traitement Automatique des Langues. *TAL*, 41(2).

Genelex 93 (1993). *Projet Eureka Genelex - Rapport sur la couche Syntaxique - Rapport sur la couche morphologique.* Consortium Genelex.

R. Gibs (1985). On the Process of Understanding Idioms. *Journal of Psycholingistic Research*, 14.

Edward Gibson et Carson Schütze (1999). Disambiguation Preferences in Noun Phrase Conjunction Do Not Mirror Corpus Frequency. *Journal of Memory and Language*, 40, 263-279.

Emmanuel Giguet (1998). *Méthodes pour l'analyse automatique de structures formelles sur documents multilingues.* Thèse de doctorat, Université de Caen.

Maurice Grévisse (1964). *Le bon usage.* Paris : Duculot. 1939 8e édition.

Maurice Grévisse (1993). *Le bon usage.* Duculot.

Gaston Gross (1996). *Les expressions figées en français.* Paris : Ophrys.

Maurice Gross (1968). *Grammaire transformationnelle du français. Syntaxe du verbe.* Paris : Larousse.

Maurice Gross (1975). *Méthodes en syntaxe. Régime des constructions complétives.* Paris : Hermann.

Maurice Gross (1995). Une grammaire locale pour l'expression des sentiments. *Langue Française, Larousse, Paris*, 105.

Benoît Habert, Cécile Fabre, et Fabrice Issac (1998). *De l'écrit au numérique : constituer, normaliser, exploiter les corpus électroniques.* Paris : InterÉditions/Masson.

Benoît Habert, Adeline Nazarenko, et André Salem (1997). *Les linguistiques de corpus.* Paris : Armand Colin.

Eva Hajicova, Jarmila Panevova, et Petr Sgall (1998). Language ressources need annotations to make them reusable: the Prague dependency Treebank. In *Proceedings First Conference on Linguistic Resources* (pp. 713-718). Granada.

Zellig S. Harris (1968). *Mathematical Structures of Language.* New York : Wiley. trad. fr. Structures mathématiques du langage, Paris, Dunod, 1971.

Virgina Holmes et John Kevin O'Regan (1981). Eye Fixation Patterns During the Reading of Relative Clause Sentences. *Journal of Verbal Learning and Verbal Behaviour*, 20, 417-430.

Paul J. Hopper et Elizabeth Traugott (1993). *Grammaticalisation.* Cambridge : Cambridge University Press.

Hélène Huot (1981). *Constructions infinitives du français. Le subordonnant «de».* Genève : Droz.

Nancy Ide, Jean Véronis, et Greg Priest-Dorman (1996). *Corpus Encoding Standard.* Rapport technique, EAGLES/MULTEX.

Timo Järvinen (1994). Annotating 200 Millions words: the Bank of English project. In *Proceedings 15th COLING* (pp. 565-568). Kyoto.

Timo Järvinen (2000). *Bank of English and beyond*, Treebanks (éd. Anne Abeillé). Kluwer Academic Publishers.

Laura Kallmeyer (2000). A query tool for syntactically annoted corpora. In *ACL 2000* (pp. 190-198). Hong Kong.

Richard S. Kayne (1975). *French syntax: the transformational cycle.* Cambridge, MA : MIT Press.

Edward L. Keenan et Sarah Hawkins (1987). The psychological validity of the accessibility hierarchy. In Keenan (éd.), *Universal Grammar* (pp. 60-85). Routledge London.

Graeme Kennedy (1998). *An Introduction to Corpus Linguistics.* London : Addison Wesley Longman.

Alexandra Kinyon (2001). A Language-Independent Shallow-Parser Compiler. In *ACL'01* Toulouse.

Anthony S. Kroch et Aravind K. Joshi (1985). *The linguistic relevance of tree adjoining grammars.* Technical report MS-CIS-85-16, Department of Computer and Information Science, Université de Pennsylvanie.

Henry Kučera et Nelson Francis (1967). *Computational analysis of present-day American English.* Providence, RI : Brown University Press.

Éric Laporte (2000). *Mots et niveau lexical*, in Ingénierie des langues (éd J. M. Pierrel), (pp. 25-49). Hermès.

Josette Lecomte (1997). *Codage Multext pour Grace/Multitag - Critère d'assignation des étiquettes morpho-syntaxiques.* Rapport technique, INaLF.

Josette Lecomte et Patrick Paroubek (1996). *Le catégoriseur d'E Brill : mise en oeuvre d'une version entraînée pour le français.* Rapport technique, INaLF, Nancy.

Geoffrey Leech (1991). *The state of the art in corpus linguistics*, in English Corpus Linguistics (éd. K. Aijmer, B. Altenberg), (pp. 8-29). Longman : Londres.

Danielle Leeman (1999). La préposition: un "auxiliaire" du nom? *Langages*, 135, 75-86.

Christiane Marchello-Nizia (1999). *Le français en diachronie - douze siècles d'évolution.* Paris : Ophrys.

Mitchell P. Marcus, Mary-Ann Marcinkiewicz, et Beatrice Santorini (1993). Building a large annotated corpus of English : the Penn Treebank. *Computational Linguistics*, 19(2), 313-330.

André Martinet (1980). *Éléments de linguistique générale.* Paris : Armand Colin.

Jacques Le Maître, Elisabeth Murisasco, et Monique Rolbert (1998). *From annotated Corpora to Databases: the SgmlQL Language*, in Linguistic databases (éd. John Nerbonne). CSLI.

Denis Maurel (1989). *Reconnaissance de séquences de mots par automate.* Thèse de doctorat, Université Paris 7, Paris.

Denis Maurel (1991). Préanalyse des adverbes de date du français. *TA Informations*, (pp. 5-17).

Tony McEnery et Andrew Wilson (1996). *Corpus Linguistics.* Edinburgh : Edinburgh University Press.

Philippe Miller (1991). *Clitics and Constituents in phrase structure grammar.* Thèse de PhD, UTR.

Jean-Claude Milner (1978). *De la syntaxe à l'interprétation. Quantités, insultes, exclamations.* Paris : Seuil.

Jean-Claude Milner (1982). *Ordre et raisons de langue.* Paris : Seuil.

Elisabeth Murisasco (1996). Manipulation de documents SGML : le langage SgmlQL. GECT, Toulon.

Fiametta Namer et Nabil Hathout (1998). Automatic construction and vali-

dation of French large lexical resources: reuse of verb theoretical descriptions. In *Proceedings First Conference on Linguistic Resources* Granada.

Michèle Noailly (1990). *Le substantif épithète.* Presses Universitaires de France.

Patrick Paroubek et Martin Rajman (2000). *Etiquetage morpho-syntaxique*, in Ingénierie des langues, in Ingénierie des Langues (éd Jean-Marie Pierrel). HERMES-Science, Paris.

Marie-Paule Péry-Woodley (1995). Quels corpus pour quels traitements automatiques? *TAL*, 36, 213-232.

Mireille Piot (1993). Les connecteurs du français. *Linguistic*æ*Investigationes*, 17, 141-160.

Joël Pynte (1998). The time-course of attachment decisions: Evidence from French. *Syntax and Semantics*, 31, 227-245.

Owen Rambow et Aravind Joshi (1992). A Formal Look at Dependency Grammars and Phrase-Structure Grammars, with Special Consideration of Word-Order Phenomena. In *International Workshop on The Meaning-Text Theory* Darmstadt. Arbeitspapiere der GMD 671. To appear in *Current Issues in Meaning-Text Theory*, Leo Wanner, editor.

Rodrigo Reyes (1997). Un Etiqueteur du français inspiré du taggeur de Brill. Rapport de stage - TALaNa, Paris 7.

Martin Riegel, Jean-Christophe Pellat, et René Rioul (1994). *Grammaire méthodique du français.* Presses Universitaires de France.

Emmanuel Roche et Yves Schabes (1995). Deterministic Part-of-Speech Tagging with Finite-State Transducers. In *Association for Computational Linguistics.*

Laurent Romary (2000). *Outils d'accès à des ressources linguistiques*, in Ingénierie des Langues (éd Jean-Marie Pierrel). HERMES-Science, Paris.

Geoffrey Sampson (1994). Suzanne, a domesday book of English Grammar. In P. d. H. N. Oostdijk (éd.), *Corpus-based research into language* (pp. 169–187). Rodopi.

Antonio Sanfilippo (1996). *EAGLES Subcategorization Standards.* http://www.icl.pi.cnr.it/EAGLES96/syntax/syntax.html.

Yves Schabes et Stuart M. Shieber (1992). *An Alternative Conception of Tree-Adjoining Derivation.* Rapport technique TR-08-92, Harvard Univ. Center for Research in Computing Technology.

Charlotte Schapira (1999). *Les stéréotypes en français.* Paris : Ophrys.

J. Senellart (1999). *Localisation d'expressions linguistiques complexes dans de gros corpus.* Thèse de doctorat, Université Paris 7.

Vijay K. Shanker (1987). *A study of Tree Adjoining Grammars.* Thèse de PhD, Department of Computer and Information Science, Université de Pennsylvanie, Philadelphia, PA.

Vijay K. Shanker, David Weir, et Owen Rambow (1995). Parsing D-Tree Grammars. In *International Workshop on Parsing Technologies.*

Hava Bat-Zeev Shyldkrot (1999). Les auxiliaires : délimitation grammaticalisation et analyse. *Langages*, 135, 3-7.

Max Silberztein (1993). *Dictionnaires électroniques et analyse automatique de textes: le système INTEX.* Paris : Masson.

Max Silberztein (1996). *INTEX 3.3 reference manual.* LADL, Paris.

Kevin Sinclair (1996). *Preliminary recommendations on Corpus Typology.* Rapport technique, EAGLES.

Wojciech Skut, Thorsten Brants, Brigitte Krenn, et Hans Uszkoreit (1998). A linguistically interpreted corpus of German newspaper texts. In *Proceedings First Conference on Linguistic Resources* (pp. 705-712). Granada.

Bangalore Srinivas (1997). *Complexity of lexical descriptions and its relevance for partial parsing.* Thèse de PhD, Université de Pennsylvanie, Philadelphia.

Christian Touratier (1996). *Le système verbal français.* Paris : Armand Colin.

Evelyne Tzoukermann, Dragomir R. Radev, et William Gale (1995). Tagging French without lexical probabilities - combining linguistic knowledge and statistical learning. In *Proceedings EACL SIGDAT Workshop* Dublin.

Hans van Halteren (2000). *Syntactic wordclass tagging.* Kluwer Academic Publishers.

Jean Véronis (1998). Annotation automatique de corpus: état de l'art. In *Questions de méthode dans la linguistique de corpus* Perpignan.

Jean Véronis (1999). *Guide d'étiquetage Multitag.*

Jean Véronis (2000). *Annotation automatique de corpus : panorama et état de la technique*, in Ingénierie des langues, 4. HERMES-Science, Paris.

Jean Véronis et Liliane Khouri (1995). Étiquetage grammatical multilingue : le projet MULTEX. *TAL*, 36.

Jean Véronis et Philippe Langlais (2000). *Evaluation of parallel text alignment systems: The ARCADE project*, in Parallel Text Processing, Kluwer Academic Publishers, Text, Speech and Language Technology Series, (éd. Jean Véronis), (pp. 369-388). Aupelf-Uref.

Ursula von Rekowski (1996). ELM-FR : Specifications for French morpho-syntax, lexicon specification and classification guidelines. EAGLES document.

Sean Wallis (2000). *Completing parsed corpora : from correcton to evolution*, in Treebanks (éd Anne Abeillé). Kluwer Academic Publishers.

Éric Wehrli (1997). *L'analyse syntaxique des langues naturelles.* Paris : Masson.

Marc Wilmet (1981). La place de l'épithète qualificative en français contemporain - étude grammaticale et stylistique. *Revue de linguistique romane*, 45, 17-73.

Marc Wilmet (1997). *Grammaire critique du français.* Paris : Hachette - Duculot.

Daniel Zagar, Joël Pynte, et Sylvie Rativeau (1997). Evidence for Early-closure Attachment on First-pass Reading Times in French. *The Quarterly Journal of Experimental Psychology*, 50, 421-438.

Arnold Zwicky (1977). *On clitics.* Bloomington : IULC.

Arnold Zwicky (1985). Clitics and particles. *Language*, 61, 283-305.

Arnold M. Zwicky et Geoffrey K. Pullum (1983). Cliticisation vs. Inflection. *Language*, 59, 502-513.