

DE LA TRANSCRIPTION À L'ÉTIQUETAGE MORPHOSYNTAXIQUE
Le cas de la banque de données textuelles orales VALIBEL

Thèse présentée par Anne Dister en
vue de l'obtention du titre de Docteur
en Philosophie et Lettres (langues et
littératures romanes).

Jury : Pr. Cédric Fairon (co-promoteur)
Pr. Michel Francard (co-promoteur)
Pr. Benoît Habert
Pr. Eric Laporte
Pr. Jean-Claude Polet (président)
Pr. Anne Catherine Simon



LOUVAIN-LA-NEUVE
ANNÉE ACADÉMIQUE 2006-2007

Cette thèse, le lecteur s'en apercevra rapidement, s'inscrit dans le prolongement du travail de plusieurs équipes de recherche. Durant toute son élaboration, j'ai ainsi pu compter sur le soutien et l'aide concrète de nombreuses personnes que je voudrais remercier ici avec chaleur : l'équipe du CENTAL (UCL) dirigée par Cédrick Fairon, celle du Centre de recherche VALIBEL (UCL) dirigée par Michel Francard et les membres de l'Institut Gaspard-Monge (Université de Marne-la-Vallée) dirigé par Éric Laporte.

J'éprouve également beaucoup de gratitude à l'égard de Gérard Purnelle (ULg), qui a guidé mes premiers pas dans la recherche et n'a cessé depuis de me soutenir. L'aide et la disponibilité d'Olivier Blanc (IGM) m'ont été également très précieuses. Laurent Kevers (UCL) m'a sauvé la vie dans la dernière ligne droite.

Enfin, outre mes co-promoteurs, je tiens à remercier les membres de mon jury, qui ont accepté de lire cette thèse : Anne Catherine Simon, Benoît Habert et Jean-Claude Polet.

Et Benoît, qui sait pourquoi.

– Vous me paraissez fort adroite, déclara Poirot.

La petite sourit d'un air penaud.

– Miss Bird dit pourtant que je serai jamais reçue à l'examen de grammaire.

– La grammaire ne mène pas à tout.

(A. Christie, *Poirot joue le jeu*, Paris, Librairie des Champs-Élysées, 1957, p. 222)

INTRODUCTION

Depuis plus de quinze ans, le Centre de recherche VALIBEL constitue et exploite des corpus oraux. Les enregistrements des productions orales sont transcrits, et ces transcriptions peuvent alors servir de base à des recherches, indépendamment de leur source sonore. C'est dire l'importance que prennent ces données secondaires et le soin qui est apporté à leur « fabrication ».

Au fil des ans et des recherches, cette base de données s'est considérablement agrandie, au point d'être actuellement, avec ses quatre millions de mots, la plus importante ressource de données textuelles orales pour le français.

Grâce au développement d'outils informatiques performants, il est aujourd'hui facile de stocker ces données ; néanmoins, leur taille les rend de moins en moins maniables pour le chercheur, qui y trouve parfois difficilement l'information désirée. Ainsi en est-il des attestations de formes ou de constructions syntaxiques pour le linguiste. La nécessité se fait donc de plus en plus pressante d'ajouter à ces textes une série d'annotations qui permettront de retrouver facilement l'information pertinente. C'est dans ce cadre que s'inscrit notre thèse. En effet, notre objectif est d'annoter les données en assignant à chaque mot ou groupe de mots du texte transcrit un lemme (c'est-à-dire une forme canonique) et une information morphosyntaxique (c'est-à-dire quant à la catégorie grammaticale et à la flexion).

Cette annotation constituera à n'en pas douter une plus-value pour ces corpus.

En effet, alors que depuis plusieurs années sont apparus de nombreux systèmes qui donnent des résultats corrects sur des textes du français standard écrit (Véronis 2000, Clément 2001), très rares sont les tentatives d'étiquetage de corpus oraux retranscrits (voir cependant Mertens 2002, Valli et Véronis 1999).

Les logiciels faisant actuellement le type d'étiquetage auquel nous souhaitons aboutir s'avèrent inadéquats pour notre corpus, et ce pour plusieurs raisons :

- les conventions de transcription notent des éléments qui posent des problèmes aux logiciels habitués à analyser des textes de français écrit « standard » : amorces de mots, répétitions, chevauchements de parole, etc., entravent en effet l'analyse linéaire du texte ;
- pour des raisons théoriques (Blanche-Benveniste et Jeanjean 1987), les transcriptions ne sont pas ponctuées ; or la plupart des logiciels se basent sur la ponctuation lors de l'étiquetage ;
- notre corpus contient des particularités lexicales qui sont soit propres à l'oral, soit propres aux variétés du français en Belgique.

Ainsi, si l'annotation de corpus oraux ne nous semble pas devoir être vue comme un problème spécifique (Benzitoun *et al.* 2004, Véronis 2000), dans la mesure où il n'y a pas de grammaire de l'oral par opposition à une grammaire de l'écrit (Blanche-Benveniste *et al.* 1990), les problèmes posés ci-dessus se doivent néanmoins d'être résolus pour rendre le système d'annotation performant.

Indépendamment de la caractéristique « données textuelles orales » de notre corpus, le travail de catégorisation et de lemmatisation n'est pas un travail aisé, qu'il soit fait par un humain ou par une machine. Le titre d'un article d'Étienne Brunet (2000) est à cet égard évocateur : « Qui lemmatise, dilemme attise ». En effet, se pose non seulement le problème des catégories grammaticales à prendre en considération – ainsi que des sous-catégories et des traits spécifiques associés à ces catégories et sous-catégories –, mais aussi celui de leur extension. Et cela d'autant plus si l'on ne se limite pas au mot graphique mais que l'on prend en compte les mots composés.

Nous venons de dire que le travail de catégorisation était complexe en raison des choix linguistiques à poser. S'ajoute encore à cette difficulté de choix celle rencontrée lorsque l'annotation est faite automatiquement par ordinateur. En effet, l'ordinateur est confronté à une masse d'ambiguïtés que ne perçoit même pas l'humain.

La grande majorité des systèmes d'étiquetage morphosyntaxique actuels sont basés sur des méthodes statistiques, avec peu ou pas de données linguistiques (Brill 1993, Schmid 1994). Le but de l'annotation est de fournir une et une seule étiquette au final.

Notre approche se veut radicalement différente : non seulement elle met en avant les données linguistiques (des dictionnaires et des grammaires de désambiguïsation construits par des humains), mais elle ne vise pas à obtenir un texte totalement désambiguïsé. Nous préférons conserver plusieurs étiquettes qu'obtenir des étiquettes erronées.

Le système de levée d'ambiguïtés que nous développons est le système ELAG (Laporte et Monceaux 1999), distribué avec le logiciel de traitement de corpus UNITEX (Paumier 2006). Ce système offre l'avantage de donner prise sur des décisions importantes à plusieurs niveaux : le choix des étiquettes, l'extension de celles-ci, la possibilité de corriger le processus d'étiquetage quand celui-ci est erroné (ce que ne permettent pas les systèmes actuels qui fonctionnent comme des boîtes noires) ou encore le fait de fournir un résultat partiel qui pourra être complété par d'autres chercheurs.

On a donc la pleine maîtrise d'un processus qui part du texte transcrit pour aboutir à un texte annoté morphosyntaxiquement. C'est ce processus que ce travail vise à décrire, dans six chapitres distincts, dont les interactions les uns avec les autres sont étroites.

En effet, la réflexion a procédé en des allers-retours constants entre les données textuelles de base, l'objectif final, les enjeux théoriques et les contraintes de faisabilité. De plus, ce qui est, dans le cadre de ce travail, l'objectif final d'annotation constitue en fait les prémices de recherches ultérieures qui devront pouvoir être faites par d'autres chercheurs. C'est en tenant compte de tous ces paramètres que nous avons élaboré cette recherche.

Le premier chapitre s'intéresse au problème de la transcription des données orales. Envisagée par certains comme une simple activité de copiste, on a montré depuis Ochs (1979) que la transcription engage une théorie.

Dans le deuxième chapitre, nous caractérisons le sous-corpus que nous avons constitué à partir de la banque de données VALIBEL. Nous avons sélectionné 60 textes, sur la base de la qualité sonore des enregistrements ainsi qu'à partir du profil sociologique des locuteurs. Ce sous-corpus comprend près de 450 000 mots.

Dans le troisième chapitre, nous étudions de manière approfondie certaines marques typiques des énoncés en cours d'élaboration, marques appelées fréquemment *disfluences* : les répétitions, les autocorrections, les amorces de mots et les pauses dites pleines (*eah*). Ces marques constituent un piétinement sur l'axe syntagmatique de l'énoncé et elles nécessitent d'être prises en compte par le système d'étiquetage. Nous analysons chaque disfluence individuellement, ainsi que les interactions entre elles. Nous tentons également de caractériser les locuteurs en fonction de ces disfluences.

Le quatrième chapitre détaille le prétraitement auquel nous soumettons le corpus. En effet, pour être utilisé par UNITEX, le texte doit subir un certain nombre de modifications. Celles-ci concernent les disfluences étudiées au chapitre 3, mais également la référencement des locuteurs ou encore le découpage du texte en unités de traitement.

Dans le cinquième chapitre, nous décrivons les ressources lexicales que nous utilisons. Celles-ci sont constituées des dictionnaires électroniques développés au LADL (Courtois et Silberztein 1990), auxquels nous avons ajouté le DELFRABEL, dictionnaire électronique du français en Belgique que nous avons construit. Nous détaillons nos choix quant aux catégories grammaticales que nous utilisons, ainsi qu'à leur extension, pour ce qui concerne les mots grammaticaux.

Dans le dernier chapitre, nous présentons le système de levée d'ambiguïtés ELAG ainsi que certaines grammaires que nous avons construites. Nous montrons l'efficacité d'un tel système, basé sur des règles contextuelles écrites par des linguistes, sur un corpus tel que le nôtre. L'intérêt d'ELAG est de travailler sur l'automate du texte et non sur un texte linéaire, et de permettre des résultats partiels de désambiguïsation.

Ce texte fait usage des rectifications orthographiques proposées par le Conseil supérieur de la Langue française (1990) et approuvées par toutes les instances francophones compétentes, dont l'Académie française.

CHAPITRE 1

LA TRANSCRIPTION DE DONNÉES ORALES

Un, deux, trois, essai, essai... ce dictaphone a l'air de fonctionner, un Olympus Pearlcordeur, j'avais acheté ça au duty-free de Heathrow avant de m'envoler vers... Où déjà ? Me rappelle plus, peu importe... Le but de l'exercice consistant à enregistrer le plus fidèlement possible les pensées qui me traversent l'esprit en ce moment précis, à savoir, voyons voir – dix heures treize, le dimanche 23 fév... San Diego, voilà ! Je l'avais acheté en partant pour ce colloque... (...) Le but de l'exercice consistant à tenter de décrire la structure de... ou plutôt à fournir un spécimen, c'est-à-dire des données brutes sur la base desquelles on pourrait commencer à tenter de décrire la structure... ou à partir desquelles on pourrait induire la structure de la pensée. Ce courant de conscience, comme l'a appelée William James, qui l'a aussi comparée, assez joliment, à un oiseau fendant les airs, se perchait un instant et prenant à nouveau son vol, un vol ponctué par des instants de... À propos de ponctuation, comment va se débrouiller la personne chargée de taper ceci ? Il faudra que je donne des instructions : mettre des points de suspension pour une brève interruption, un point pour une pause plus marquée, changer de paragraphe si la pause se prolonge... Ce dictaphone est à déclenchement vocal automatique, il s'arrête si on se tait plus de trois secondes, mais il y aura des pauses d'une durée inférieure perceptibles dans le flux de la parole... Malin, ce petit gadget...

(David Lodge, *Mémoires intimes*, Paris, Éditions Payot et Rivages, 2002, pp. 11-12)

1 INTRODUCTION

Dans ce premier chapitre, nous voudrions nous pencher sur la pratique de consignation des données orales, consignation qui passe par la mise en écrit du son. Souvent considérée comme une étape nécessaire mais peu valorisée, la transcription est au cœur de la recherche sur l'oral car dans bien des études, comme c'est le cas dans celle-ci, le texte transcrit constitue les seules données que manipulera le chercheur. Le transcripteur a donc de ce fait une position centrale, et ses choix de transcription auront des répercussions inévitables sur l'analyse.

2 DE LA NÉCESSITÉ D'ENREGISTRER¹

Afin d'étudier des productions langagières orales, il est nécessaire d'en fixer le flux, par essence évanescent².

Pour le français, ce travail de fixation, dans un but de pérennisation du matériau langagier, a été entrepris dès 1911 par Ferdinand Brunot. Celui-ci procède en effet à des enregistrements sonores sur phonogramme, à Paris et dans toute la France. C'est également à cette date que sont créées, sous l'impulsion du linguiste et de l'industriel Émile Pathé, « Les Archives de la Parole », qui seront remplacées en 1927 par le « Musée de la Parole et du Geste ».

En ce qui concerne la consignation de données sonores, l'invention du magnétophone en Allemagne dans les années 1930 peut être considérée comme une véritable révolution. Dès lors, enregistrer devient une tâche relativement aisée. Au fil des années et des progrès technologiques, le matériel se fait plus léger, plus solide, plus maniable (invention du magnétophone portable dans les années 1950) et son prix est de moins en moins élevé. L'objet fugace qu'est l'oral peut ainsi être pérennisé sur des bandes. Le développement de l'informatique et les possibilités offertes par la numérisation des données constituent quant à elles une autre révolution : le document original peut être reproduit à l'identique, à l'infini, de manière

¹ Nous n'évoquons pas ici les problèmes, eux aussi cruciaux, de recueil des données. Sur ce sujet, voir notamment Gadet (2000a et 2000b), dont les titres des articles sont à cet égard évocateurs : « On n'en a pas fini avec les problèmes de recueil de corpus » et « Derrière les problèmes méthodologiques du recueil des données ». Pour les questions éthiques et juridiques qui lient le chercheur et les locuteurs, nous renvoyons au *Guide des bonnes pratiques* (2005) ; les aspects légaux ont été à la base des réflexions du groupe de travail qui a rédigé ce guide.

² Notons que, pour des études ponctuelles, cette étape première de fixation du son par enregistrement peut parfois être omise. C'est le cas du corpus de lapsus recueillis « à la volée » par Berthille Pallaud (1999). Mais ces cas sont extrêmement rares.

automatisée et sans perte de qualité. Le concept même d'original perd de son sens.

3 DE LA NÉCESSITÉ DE TRANSCRIRE

Tout travail sérieux sur les productions orales est impossible à partir de la seule source sonore. En effet, le chercheur aura beau écouter et réécouter les enregistrements, il ne peut les appréhender uniquement par le biais du son. L'oral ne pourra devenir objet d'étude à part entière qu'à partir de sa mise en/par écrit. La parole reste fluide, essentiellement fugace, même après avoir été captée sur bande sonore :

On ne peut pas étudier l'oral par l'oral, en se fiant à la mémoire qu'on en garde. On ne peut pas, sans le secours de la représentation visuelle, parcourir l'oral en tous sens et en comparer les morceaux. (Blanche-Benveniste 2000c : 24)

On est donc là devant un paradoxe, mis en évidence par de nombreux chercheurs (Blanche-Benveniste et Jeanjean 1987, Blanche-Benveniste 2000c, Gadet 2003, Raingeard et Lorscheider 1977) : pour approcher l'oral, on doit « en passer » par l'écrit.

3.1 Le mythe du copiste

La transcription des données apparaît donc comme une étape indispensable dans l'étude de l'oral. Or, on constate bien souvent que cette démarche cruciale est généralement sous-estimée : en effet, régulièrement, cette tâche est confiée à des personnes qui ne seront pas véritablement engagées dans les processus ultérieurs de la recherche³. Dans une telle perspective, transcrire est considéré comme une tâche subalterne, qui ne relève pas véritablement de la recherche, même si elle en est un passage obligé. Cette dévalorisation est liée à la croyance, assez répandue parmi les personnes ne s'étant jamais attelées à la tâche, que pour bien transcrire, il n'y aurait qu'à bien écouter (Blanche-Benveniste et Jeanjean 1987, Blanche-Benveniste 2002c). La mise en mots de productions orales est vue comme une activité mécanique, dans laquelle le transcripateur serait un simple exécutant totalement indifférent à la tâche à laquelle il se livre. Le but de ce chapitre est de mettre en évidence qu'il n'en est rien : le transcripateur

³. Le corpus étudié par Diane Vincent (1993) est transcrit par des secrétaires ; les transcripateurs du *COBUILD spoken corpus* « have secretarial experience » (Payne 1995 : 203).

participe à la création des données, car son travail est à la fois analytique et interprétatif.

3.2 Le statut du texte transcrit

Les termes qui dénomment le passage de l'oral vers un support écrit varient d'un auteur à l'autre : *transcription* est le plus couramment utilisé, mais on trouve aussi *retranscription*, *fixation*, *notation* (Houdebine 1979), *matérialisation*, *transformation*, *transcodage* (Leroy 1985 : 7), *traduction* (*translation*, Cheepen 1995 : 135), *représentation heuristique* (Mondada 2000 : 131) ou encore *déformation* (Bally 1935, Giovannoni et Savelli 1990, citant Jespersen).

Certains de ces termes, non neutres, marquent les modifications qui s'opèrent lors de la mise en écrit. Le statut du texte transcrit est ainsi interrogé :

En migrant du terrain à la bande puis de la bande à la page, l'oral se transforme donc radicalement. (Mondada 2000 : 132)

Raingard et Lorscheider (1977) mettent en avant le décalage qui existe entre le texte transcrit et sa source sonore. Ce décalage s'effectue en deux temps : d'abord lors de l'enregistrement, qui procède à une première sélection dans les données (choix du moment enregistré, positionnement du ou des micro(s), de la caméra, etc.), et ensuite dans le filtrage qui s'opère dans les choix de transcription adoptés (les phénomènes concernés par ces choix)⁴.

Leroy (1985) se pose la question de la correspondance entre les deux documents, la source sonore, donnée primaire, et sa transcription, donnée secondaire. Pour elle, le texte transcrit « n'est rien en lui-même »⁵. De nombreux auteurs (notamment Chafe 1995, Mondada 2000) insistent sur le statut secondaire que revêtent les transcriptions et sur la nécessité de recourir aux données sonores pour effectuer l'analyse, dont la transcription ne peut être plus qu'un support⁶.

Pourtant, on constate que, bien souvent, la transcription devient la donnée même (Ochs faisait déjà ce constat en 1979) :

⁴ Les auteures opposent ce filtrage, inévitable, au « toilettage » qui consisterait notamment à ne pas transcrire les scories de l'oral et « défigurerait » ainsi les données.

⁵ L'auteure parle même de textes « morts ».

⁶ Le paradoxe de la mise en écrit des données orales est encore plus prégnant dans le cadre de notre travail, qui se base exclusivement sur les données transcrites, que nous considérons comme des données à part entière.

In fact, I suspect that all too many spoken language corpus users are quite content to ignore the sound of language completely, doing whatever they do on the basis of transcription alone, as if they were dealing with written language. I find that too bad (...) (Chafe 1995 : 54).

Dans notre recherche, les transcriptions constituent les seules données que nous exploitons. De ce fait, la nécessité d'en théoriser les enjeux devient encore plus fondamentale.

Le texte transcrit est interrogé (ou revendiqué par les auteurs de corpus) en termes de **qualité**, de **fidélité** (Cosnier et Kerbrat 1987 : 371, Roubeau 2004a : 163, sur le fait de noter les amorces), d'**authenticité**⁷. Mais qu'entend-on par là exactement ? La plupart des auteurs ne s'en expliquent pas toujours. Pour Roubeau (2004b : 249), il s'agit de « restitue[r] le texte écouté avec un maximum de fidélité, sans aucun toilettage ». Souvent, on constate aussi que *authenticité* implique *spontanéité* dans l'interaction.

Or, on s'en rend vite compte en transcrivant, « le texte authentique fuit, du moins par certains bouts » (Blanche-Benveniste et Jeanjean 1987 : 112), « le texte authentique est un mythe rassurant, il n'est pas vérifiable » (idem : 115). Les auteures comparent cette recherche de l'authenticité dans la transcription à l'édition de textes médiévaux par les philologues, pour lesquels le texte authentique « se dérobe » également. Nous pourrions également comparer ce travail à l'édition de textes plus récents, et des choix que pose l'éditeur pour l'établissement du texte dans la collection de la Pléiade par exemple.

De plus, dans sa pratique, le transcripateur engage ses savoirs et ses représentations : sur le monde, sur la langue (écrite, orale), sur les locuteurs (cf. § 5.4.3.).

Le transcripateur n'est donc pas qu'un copiste, même un copiste de très grande qualité :

Pour le linguiste praticien averti de la transcription, le rêve du copiste qui saisirait la matière vivante telle qu'elle est, objectivement, et la restituerait ipso facto, est illusoire. (...) Le réel transcrit n'est jamais le réel émis mais un réel perçu et traduit. Au cours de ce processus, on assiste à une série de filtrages, d'appauvrissements, d'interprétations. (Giovannoni et Savelli 1990)

⁷ Authenticité ici par rapport à la transcription. Quand on parle de données authentiques, en ce qui concerne les données primaires (l'enregistrement), il est fait référence à des données non sollicitées, non provoquées par le chercheur (on parle aussi de données « écologiques »).

4 DIFFICULTÉS DE TRANSCRIRE : LES DIFFICULTÉS D'ÉCOUTE LIÉES À DES ASPECTS TECHNIQUES⁸

Nous abordons tout d'abord les problèmes de transcription étroitement liés à la perception et pour lesquels les aspects techniques entrent en jeu.

Si l'on excepte les enregistrements réalisés en laboratoire (« speech corpora »), dans une sorte de vase clos imperméable aux bruits de l'extérieur, la plupart des enregistrements sont faits en milieu naturel (« corpora of spoken discourse » ou « spoken language corpora »⁹). C'est évidemment le cas lorsqu'il s'agit de recueillir de l'oral spontané (repas en famille, réunion d'amis, etc.), mais également lors de situations de communication plus contraintes (prises de paroles publiques tels que discours politiques ou conférences, débats télévisés, entretiens d'embauche, etc.). De nombreux paramètres entrent en jeu, qui parasitent d'une certaine manière l'enregistrement et diminuent sa qualité sonore :

- Il peut s'agir de bruits de fond qui passaient inaperçus lors de l'enregistrement : bruits de couverts lors des repas, peu perceptibles sur le moment mais qui résonnent etaturent l'enregistrement ; photocopieuse qui semble très silencieuse, mais dont le bruit sur la bande masque les paroles des locuteurs ; sonnerie de téléphone ; locuteur qui tousse, bébé qui pleure, chien qui aboie, etc. Si l'on peut réduire ou éliminer certaines de ces perturbations (veiller à éteindre la radio qui fonctionne en sourdine dans la pièce, à fermer les fenêtres, etc.), on ne peut les prévoir et les maîtriser toutes (cloches qui sonnent dans l'école, par exemple). Il arrive aussi que prime la situation de communication dans le choix de l'enregistrement et que l'on soit conscient des bruits générés par le cadre choisi : qu'on pense à l'enquête de Labov dans les grands magasins.
- À moins de travailler avec un petit nombre de locuteurs qui portent chacun un micro-cravate, on ne peut garantir que les locuteurs ne s'éloigneront pas du (ou des) micro(s), rendant leurs paroles parfois peu audibles¹⁰.

⁸ Sur ces questions, voir l'incontournable livre de Blanche-Benveniste et Jeanjean, *Le français parlé. Transcription et édition* (1987).

⁹ Sur cette différence, voir Leech *et al.* (1995 : 5).

¹⁰ Dans le cas où le but est de recueillir de l'oral spontané, et que l'on veut faire oublier aux locuteurs l'existence du micro, leur demander de ne pas s'en éloigner risque de rappeler constamment la présence de celui-ci. Dans le cadre d'un séminaire sur le recueil et la transcription de données orales, dont le but était de recueillir de l'oral spontané lors de conversations familiales, les étudiants se sont trouvés confrontés à ce problème. Ils ont d'ailleurs fait remarquer, comparant leur travail aux enregistrements de l'émission de

- Pour des enregistrements relativement longs, le changement de cassette cause inévitablement une coupure lors de l'enregistrement.
- Les enregistrements sont souvent monophoniques, c'est-à-dire réalisés avec un seul micro. De ce fait, « ce ne sont pas les mêmes données [que dans l'écoute directe] qui arrivent à nos oreilles. (...) Donc toute une hiérarchisation des sons se perd, ce qui rend le travail du transcripateur très ardu » (Blanche-Benveniste et Jeanjean 1987 : 94).
- Des enregistrements faits avec du matériel de mauvaise qualité seront évidemment plus difficiles à écouter et à transcrire.

Des difficultés d'écoute peuvent également être liées au nombre de personnes prenant part à la conversation¹¹. En effet, un nombre élevé de participants constitue un écueil majeur dans le travail du transcripateur : plus il y a de participants, plus la transcription risque d'être malaisée. En effet, plus nombreux seront les chevauchements de parole et plus difficile sera l'identification des différents locuteurs.

5 « TRANSCRIPTION AS THEORY »

Nous l'avons dit, la transcription n'est pas une activité neutre (Bilger 2000 : 77) et transparente faite par un quelconque copiste. Au contraire, elle est une activité sélective et interprétative :

Far from being an objective and exhaustive mirroring of events of an interaction, a transcript is fundamentally selective and interpretative. (Edwards 1995 : 19)

Depuis l'article fondateur de Ochs (1979), on s'accorde pour dire qu'elle engage, ou révèle, une théorie. La transcription, ainsi que l'établissement par le chercheur de conventions qui la sous-tendent, font l'objet d'une véritable réflexion dans laquelle apparaissent des enjeux théoriques¹² :

(...) transcription is a selective process reflecting theoretical goals and definitions. (Ochs 1979 : 44)

télévision *Strip-Tease*, dans laquelle les protagonistes sont filmés dans des situations quotidiennes et oublient la présence de la caméra, qu'il est plus facile pour une équipe de télévision de se déplacer avec caméras et micros que pour eux, seuls, avec un seul micro.

¹¹ Le terme de conversation est ici employé dans son sens général d'échange langagier entre au moins deux locuteurs.

¹² Outre à l'article de Ochs (1979), on se reportera, entre autres, à Blanche-Benveniste et Jeanjean (1987), Edwards (1993 et 1995), Gumperz et Berenz (1993), Mondada (2000).

Il n'y a pas, aussi bien dans le monde anglo-saxon que dans la francophonie, de modèle unique de transcription qui se soit imposé, qui fasse autorité, et cela malgré les tentatives d'uniformisation et les recommandations qui émanent d'organismes de normalisation¹³.

C'est en effet qu'il n'y a pas qu'une seule manière de matérialiser le son dans de l'écrit, qu'une seule façon de représenter le flux oral dans la matière graphique.

A first question might be whether any transcription system can be ideal. Probably the answer is no. How we transcribe depends on our purposes, which are inevitably going to differ. (Chafe 1995: 54-55)

Choisir un système de transcription plutôt qu'un autre n'est donc pas innocent. Comme le dit Mondada (2000 : 132), ces choix

impliquent – de façon souvent implicite – des positionnements spécifiques, à rapporter aux fins pratiques et théoriques poursuivies par l'analyste qui les adopte.

Selon Chafe (1995 : 55),

any transcription system is a theory of what is significant about language.

Les phénomènes visés par les conventions peuvent (doivent même) varier selon les objectifs de recherche, selon les modèles analytiques :

(...) les choix présidant à l'édition d'un corpus doivent être explicités, maîtrisés et intégrés dans les objectifs de la recherche. (Raingeard et Lorscheider 1977 : 14)

Dans le cadre de transcriptions faites pour des analyses linguistiques, cette correspondance entre les choix de transcription d'une part et les présupposés théoriques et les objectifs de la recherche d'autre part posent parfois un problème de circularité mis en évidence par Mondada (2000 : 133) :

De nombreuses interprétations de phénomènes se trouvent en effet incorporées dans la transcription a priori, alors qu'elles sont le but déclaré de l'analyse censée s'exercer a posteriori sur cette transcription.

¹³ La TEI (Text encoding Initiative, voir TEI (2002)), NERC (Network of European Reference Corpora, voir French (1992)) ou encore EAGLES (Expert Advisory Group on Language Engineering Standards, voir Llisterri (1996)).

Il faut être attentif au fait que la mise à plat opérée par la transcription tend à donner à celle-ci une apparente objectivité sur laquelle se base ensuite l'analyse. À trop intégrer les objectifs de la recherche dans la transcription, on risque d'obtenir les résultats de l'analyse avant de l'avoir faite. C'est la raison pour laquelle il est nécessaire d'explicitier ses conventions de transcription.

Pour faire de la transcription une démarche scientifique, le chercheur doit se poser un certain nombre de questions : que transcrire ? dans quel but de recherche ? qui travaillera sur les données transcrites ? qui sera susceptible de lire ces transcriptions, dans l'exposé des résultats de la recherche par exemple ? En résumé, la question devrait être : « Quelle transcription, par qui, pour quel usage ? » Il faut trouver un juste équilibre, en sachant qu'il s'agit effectivement d'un équilibre, qui résulte de choix. Ces choix sont posés en tenant compte des hypothèses de recherche ainsi que des exploitations futures du texte transcrit. Les différentes applications faites à partir des corpus permettent d'ailleurs bien souvent d'éclairer les aspects théoriques pris en considération dans les choix de transcription (Leech *et al.* 1995).

Ainsi, la transcription d'un même évènement conversationnel ne se fera pas de façon identique par un journaliste, un linguiste, un historien ou un sociologue. Les buts visés ne sont pas les mêmes, les moyens d'y parvenir seront donc différents.

On le voit dans l'exemple suivant emprunté à Blanche-Benveniste (2003a : 340). Il s'agit de trois transcriptions faites d'un même évènement langagier, à savoir une intervention télévisée de François Mitterrand en 1988. La première version est celle d'un transcripateur non professionnel qui s'attache à « tout transcrire » ; la seconde est celle du journal *Libération* et la troisième celle du *Monde*¹⁴.

Version 1 : transcripateur non professionnel

Le congrès d'Valence s'est tenu... pendant que j'étais...
au Mexique... pour la conférence de Cancun... et quand
je suis revenu... accueilli par les journalistes... très ému...
euuh dès ma *descent'* d'avion... j'ai dit : *ben c'est pas ça*
qu'i faut faire... c'est pas ça qu'i faut faire... i faut pas...
euuh... essayer d'diviser les Français... et... certaines
expressions excessives de mes amis *eh ben... iz ont eu*
tort.

¹⁴ Les italiques sont de Blanche-Benveniste.

Version 2 : *Libération*

Le congrès de Valence s'est tenu pendant que j'étais au Mexique, pour la conférence de Cancun et, quand je suis revenu, accueilli par les journalistes, très ému, dès ma descente d'avion, j'ai dit : « *Eh bien, ce n'est pas ça qu'il faut faire, il ne faut pas essayer de diviser les Français, et certaines expressions excessives de mes amis, eh bien, ils ont eu tort.* »

Version 3 : *Le Monde*

(...) j'ai dit : « *Ce n'est pas cela qu'il faut faire, il ne faut pas essayer de diviser les Français, et certaines expressions excessives de mes amis, eh bien, ils ont eu tort.* »

Dans la version 1, le transcripteur note les pauses (points de suspension, *eah*), les élisions au moyen de l'apostrophe (*descent', d'diviser*) (cf. § 6.2.1.), les répétitions (deux fois *c'est pas ça qu'i faut faire*) ou la prononciation des pronoms (*i, iz*).

Les versions des journalistes tendent quant à elles à supprimer ces particularités de l'oral : l'orthographe standard est rétablie, les pauses pleines et les répétitions sont supprimées, la ponctuation est introduite. L'écrit procède à une standardisation des paroles : le *ne* de négation est réintroduit, et *Le Monde* transforme même le *ça* en *cela*.

On a donc ici trois représentations relativement différentes d'un même flux sonore. Choisir l'une ou l'autre produit inévitablement des effets différents sur le lecteur.

De la même manière, une transcription faite par un linguiste peut prendre des formes différentes selon qu'elle sera le matériau même de la recherche ou qu'elle sera destinée à être diffusée auprès d'un public plus large de non-spécialistes. Ainsi par exemple, la ponctuation est absente des transcriptions du Groupe Aixois de Recherches en Syntaxe (GARS)¹⁵, mais réintégrée dans des publications destinées à des lecteurs non spécialistes de ce type de texte (Blanche-Benveniste 2003a : 34).

¹⁵ Le GARS, pionnier dans les études de syntaxe de l'oral et la constitution de corpus oraux en français, s'est constitué en 1977 autour de Claire-Blanche Benveniste. Il est devenu aujourd'hui DELIC (DEscription Linguistique Informatisée sur Corpus). Par commodité, dans la suite de travail, nous continuerons à parler du GARS.

5.1 Segmenter et catégoriser

The process of discourse transcription is never mechanical, but crucially relies on interpretation within a theoretical frame of reference to arrive at functionally significant categories, rather than acoustic facts. (Du Bois 1991 : 72)

Dans un premier temps, le transcripteur découpe, parmi le continuum de la chaîne sonore, des unités qui correspondent aux unités du code écrit. Il tente de représenter dans le moule graphique de l'écrit¹⁶ la suite, uniquement interrompue par des pauses, du flux oral.

On sait que ce travail de mise en mots nécessite un apprentissage et qu'il est lié à notre culture (Goody 1979).

Discrétiser la chaîne sonore nécessite de segmenter, d'identifier et de catégoriser : segmenter un continuum, identifier des unités fonctionnelles, catégoriser ces unités. Ce travail de catégorisation se révèle notamment dans les choix que pose le transcripteur devant des variantes : ces variantes seront-elles représentées comme des allomorphes d'un même morphème, ou des allophones d'un même phonème (auxquels elles ne seront pas reliées explicitement) ou transcrites par une graphie qui les standardisera ? (cf. § 6) La catégorisation se fait encore dans l'attribution des paroles à un locuteur.

5.2 Que transcrire ?

La transcription oblige à opérer un choix parmi les divers éléments qui participent à une conversation :

A more useful transcript is a more selective one. Selectivity, then, is to be encouraged. But selectivity should not be random and implicit. Rather, the transcriber should be conscious of the filtering process. The basis for the transcription should be clear (Ochs 1979 : 44).

Les éléments susceptibles d'être transcrits, qui entrent en jeu dans une conversation ordinaire, peuvent être classés de la manière suivante¹⁷ :

¹⁶ Lorsque c'est la transcription orthographique qui est choisie (cf. § 6.2)

¹⁷ Nous reprenons ces divers éléments à Atkinson et Heritage 1999, Bilger 2000, Blanche-Benveniste et al. 1991, Crowdy 1995, Cook 1995, Du Bois 1991, Du Bois *et al.* 1993a et 1993b, Edwards 1993 et 1995, Ehlich 1993, Gumperz et Berenz 1993, Llisterri 1996, Ochs 1979, Payne 1995, Simon 2004, TEI 2002.

le vocal

- le verbal, c'est-à-dire les paroles (éléments lexicaux) ;
- le verbal, semi-lexical : principalement les pauses dites remplies (*euh*), les onomatopées (*boum*, *bang*, etc.) ;
- le vocal mais non-verbal : cette catégorie reprend des phénomènes tels que le rire, les pleurs, la toux, les éternuements, etc. ;
- la pause, vue comme une interruption dans le flux verbal ;
- vocal mais non segmental : les faits prosodiques ;

le non-vocal lié aux locuteurs

- les informations sur les locuteurs (âge, sexe, lien de parenté, degré d'intimité entre les locuteurs, etc.)¹⁸ ;
- le kinésique : les gestes, les regards, les déplacements, etc. ;

le non-vocal lié au contexte de la conversation

- les événements qui peuvent se produire lors de la conversation mais ne sont pas le fait des locuteurs : sonnerie de téléphone, bruits extérieurs, etc. ;
- conditions générales de la conversation (lieu, temps, etc.).

L'objectif de la TEI (Text Encoding Initiative¹⁹) est de définir un format de création des données qui permette leur échange et leur stockage. Le but de la standardisation proposée par la TEI est de fournir des versions « computer readable » des textes.

Les directives recensent, pour chaque type de texte, les traits à encoder ainsi qu'une normalisation des balises pour chacun de ces traits. En ce qui concerne les textes transcrits de l'oral, les recommandations distinguent sept éléments caractéristiques de ces textes, et balisés comme tels dans les transcriptions :

- <u>²⁰ : les paroles (utterance) ;
- <pause> : les pauses ;
- <vocal> : phénomènes vocalisés mais pas nécessairement lexicaux (ex : pauses remplies) ;

¹⁸ Pour autant que ces informations soient disponibles.

¹⁹ Voir <http://www.tei-c.org/>.

²⁰ À l'intérieur des balises <u> peuvent apparaître les six autres phénomènes répertoriés.

- <kinesic> : phénomènes communicatifs mais non vocalisés (ex: gestes) ;
- <event> : phénomènes non nécessairement vocalisés ou communicatifs (ex. : bruits divers) ;
- <writing> : passage de texte écrit oralisé par l'un des participants (texte lu) ;
- <shift> : balise qui marque des variations de prosodie dans les paroles d'un locuteur ;

La limite entre ces catégories peut n'être pas tout à fait étanche : en effet, le bruit de la fermeture d'une porte peut être classé comme un <event>, mais dans certains cas, il peut avoir une valeur communicative et figurer alors parmi les phénomènes <kinesic>. Ici, le transcripteur est déjà dans l'interprétation. Classer l'un ou l'autre événement dans <event> ou dans <kinesic> aura inévitablement des répercussions sur l'analyse ultérieure des données.

5.3 Finesse de la transcription

Chacun de ces sept éléments peut être accompagné d'attributs, définis préalablement.

Ainsi, l'élément <shift> peut être accompagné d'un certain nombre d'attributs, qui concernent le rythme, le tempo, la tension, la qualité de la voix, etc. Ces attributs ont chacun plusieurs valeurs prédéfinies. Par exemple, pour la qualité de la voix²¹, on a : *en sanglotant* (*trem*, pour *tremulous*), *avec une voix enrouée* (*husk*, pour *husky*), *en chuchotant* (*whisp*, pour *whisper*), *avec une voix de fausset* (*fals*, pour *falsetto*), etc.

Le principe même des recommandations de la TEI est de laisser le choix au transcripteur : parmi ces attributs et ces valeurs, il sélectionne ceux qui correspondent aux phénomènes qu'il juge pertinent de transcrire pour sa recherche. C'est le concept de la « Chicago Pizza » : une pizza, quels que soient les éléments qu'on y ajoute, reste une pizza. L'ingrédient de base, la pâte à pizza dans les conventions de la TEI, ce sont les « mots »²².

²¹ Pour la transcription de la qualité de la voix, voir aussi Du Bois *et al.* (1993b : 68-71).

²² Ce que critique Cook, dont l'article porte sur « the untranscribable » (1995: 41-42) : « (...) the assumption that the language of oral interaction can be analysed as words abstracted and isolated from other features of the original event may fundamentally distort the object of analysis, imposing upon it a feature of written communication which is alien to the nature of oral interaction. The pizza analogy and the concept of plain text are misleading, for in speech the non-verbal aspects of interaction are not mere toppings, but essential, even constitutive of what is happening. »

Un même fait langagier ou paralangagier peut donc être transcrit/décrit plus ou moins finement, voire être omis. Ainsi, le hochement de tête, au centre de la gestuelle d'un personnage dans l'extrait de roman suivant :

Probst montra son doigt et, pour la vingtième fois, raconta l'histoire du stade. Les dix dernières fois, son récit avait été parfaitement identique, au mot près.

À chaque phrase, Thompson acquiesçait d'un hochement de tête. D'amples hochements de haut en bas, d'autres plus petits et horizontaux, d'autres rotatifs qui exprimaient une compréhension, une approbation sans réserve, d'autres plus insolites avec des angles variés, et aussi quelques hochements oscillants semblables au mouvement d'un essuie-glace. (Jonathan Franzen, *La vingt-septième ville* Paris, L'Olivier-Le Seuil, 2004, p. 259)

Là où certaines transcriptions pourraient omettre totalement l'information – le hochement de tête n'est accompagné d'aucune vocalisation et ne peut être perçu à la seule écoute de la bande sonore –, ou se contenter de l'indication générale « hochement de tête », une analyse fine de la gestuelle des locuteurs nécessitera de catégoriser ces différents hochements de tête dont la différence physique et fonctionnelle apparaît clairement à la lecture du texte de Franzen.

Un chercheur peut établir des niveaux de transcription différents, selon que certains faits sont ou non notés. Ochs (1979) parle de « basic transcript », par opposition à une transcription plus détaillée²³.

5.4 La recherche du sens

Dans son activité cognitive de segmentation et de catégorisation, le transcripteur cherche à faire sens :

Phonéticiens et psychologues s'accordent à dire que la perception est un processus actif qui s'appuie sur la compréhension des énoncés (Blanche-Benveniste et Jeanjean 1987 : 103).

L'homme a un rôle actif dans la perception des énoncés oraux. Il cherche à construire, voire à reconstruire du sens. Le transcripteur, dans ce travail de recherche de sens, part en général du principe que le locuteur tient des propos sensés, qu'il est cohérent, qu'il cherche lui-même à faire sens, etc. Ce

²³ Voir aussi Du Bois *et al.* (1993b : 46) pour cette même opposition et une description des éléments qui figurent dans chacun de ces « niveaux ».

crédit accordé au locuteur²⁴ permet bien souvent de choisir entre des interprétations concurrentes, et ainsi, de ne pas multiplier à l'envi les transcriptions en parallèle, appelées aussi *multitranscriptions*.

Ainsi, devant une suite sonore homophone, c'est après une analyse, qui explore notamment le contexte et garantit une cohérence textuelle, que le transcripateur sélectionnera une solution parmi les multiples possibilités qui théoriquement s'offraient à lui.

5.4.1 Les homophonies

Un certain nombre de difficultés de perception sont le fait de formes auditivement très proches, sur lesquelles achoppent bien souvent les transcripateurs. Il s'agit de pièges récurrents, souvent non perçus comme tels d'ailleurs, et au sujet desquels il est utile d'avertir le transcripateur débutant.

Certaines divergences reviennent très régulièrement et nous mettent en alerte ; il s'agit de « points fragiles » sur lesquels il convient d'être particulièrement vigilants dans la transcription. Les éléments clitiques (comme articles, pronoms, prépositions), les conjonctions, les auxiliaires, les finales verbales, sont fréquemment interprétés avec des divergences. C'est d'autant plus important que nous touchons là à des secteurs grammaticalisés de la langue. (Blanche-Benveniste et Jeanjean 1987 : 107)

De nombreux articles consacrés à l'oral et/ou à sa transcription donnent des exemples de ces pièges classiques de la perception, pièges dus pour la plupart à des ambiguïtés phonétiques²⁵.

Pallaud (2002a : 268) met en évidence les conséquences de ces difficultés de transcription²⁶ :

En fait, la transcription recèle parfois, en ces points contestés, l'indice d'une deuxième voix qui s'introduit dans l'énoncé du locuteur ; celle du transcripateur. C'est

²⁴ Pallaud (2002b : 80-81) pointe le fait que les hypothèses de cohérence et de congruence ont pour effet de sous-estimer les phénomènes de lapsus.

²⁵ Pour le français, voir notamment Blanche-Benveniste et Jeanjean 1987 : 105-110, Blanche-Benveniste 2002c, Cappeau 1997, Fuchs 1996, Gedeo 1997, Giovannoni 1990, Giovannoni et Savelli 1990, Pallaud 2002. Remarquons que, dans ces typologies des ambiguïtés, certaines ne valent pas pour l'ensemble de la communauté francophone. Voir à ce sujet Raingeard et Lorscheider (1977).

²⁶ Cook (1995 : 38) pense quant à lui que la différence entre des transcriptions faites par plusieurs transcripateurs est souvent exagérée.

ce que révèle le travail de vérification des documents linguistiques obtenus même après une transcription soigneuse. Le texte obtenu contient en quelques endroits des erreurs d'écoute qui, n'étant pas la représentation écrite de ce que le locuteur a dit, sont la transcription de ce que le transcripteur a cru percevoir. En ces endroits, le texte n'appartient donc pas au locuteur mais bien au transcripteur. En ce sens, on peut dire que le discours du locuteur se trouve "ajouré" par un autre discours, celui du transcripteur.

Illustration : les amorces de morphème

Pour illustrer ce problème d'homophonie, nous voudrions nous pencher plus particulièrement ici sur la transcription des amorces de morphèmes (Pallaud 1999 et 2002b), afin de mettre en évidence qu'un phénomène, simple en apparence, pose des questions au transcripteur²⁷ et l'oblige à des choix qui risquent de se répercuter dans l'analyse.

Les amorces de morphèmes se rencontrent très fréquemment dans le parler des locuteurs. Elles sont la marque d'un achoppement dans l'énonciation, et des recherches qui veulent rendre compte de l'élaboration des énoncés et des modes de productions propres à l'oral doivent les intégrer dans leurs transcriptions. Les conventions de transcription de VALIBEL notent ce phénomène par une barre oblique collée directement au mot amorcé ; le mot amorcé conserve, jusqu'au point de troncation, son orthographe traditionnelle. Cette convention est non ambiguë, et les amorces sont donc aisément identifiables, par le lecteur, mais aussi de manière totalement automatisée, dans le texte transcrit. Voici quelques exemples d'amorces²⁸ :

(1) **norKJ1** ah bè ils l'auront le euh il est joué de toute manière par / ils rassemblent des gens // euh / ils créent

²⁷ Pour les erreurs commises par les transcripteurs dans la représentation des amorces, voir Pallaud (2002b).

²⁸ Sauf indication contraire, tous nos exemples sont attestés et issus de la banque de données VALIBEL. Nous les présentons dans un tour de parole complet (cf. § 7), sauf lorsque celui-ci est trop long et que sa citation complète n'apporte rien au propos. Le tour de parole commence ou s'achève alors par les points de suspension entre parenthèses (...), notation qui ne figure pas dans le texte original.

Le code de 5 lettres et 1 chiffre qui précède l'exemple est le code du locuteur (3 lettres minuscules pour le corpus, 2 pour les initiales du locuteur) ; le code entre parenthèse à la fin est celui de l'enregistrement. Les conventions de transcription adoptées par le centre de recherche VALIBEL sont détaillées à l'adresse suivante : <http://valibel.fltr.ucl.ac.be/>.

Les principales conventions utilisées dans les transcriptions reprises dans ce travail sont décrites dans la section 9 de ce chapitre.

un corps // et dès le moment où on crée un corps on on
s'ada/ on s'adapte (...) [nokJ1r]

(2) **ilrMS1** en Autriche / un mois / enfin en Hongrie un
 mois et en en **Autri/** en Autriche une semaine / c'est
 amusant [ilrMS1r]

Dans ces deux exemples, la transcription de l'amorce semble s'imposer d'elle-même dans la mesure où celle-ci est immédiatement reprise et complétée dans la suite de la séquence. Peu de travail d'interprétation est demandé au transcripateur, aussi bien dans l'interprétation du fait langagier comme une amorce de morphème que dans la manière de l'orthographier. Par contre, dans l'exemple 3, c'est en interprétant le sens global de l'énoncé que le transcripateur décode le son [ko] comme étant l'amorce de *quotidien*, et qu'il l'orthographie alors *quo/*, et non *ko/* ou *co/*, ce qui aurait été théoriquement possible pour un mot du français :

(3) **ilePA2** (...) le le le le contraire est vrai aussi bon il
 est certain qu'un qu'un journaliste qui est habitué à lire
 un jou/ un grand journal populaire à lire des grands
 titres et cetera bè si on lui dit voilà à partir de
 maintenant vous allez lire La-Libre-Belgique ben ça ça va
 l'ennuyer c'est un langage qu'il ne va pas aimer / qu'il ne
 va comprendre qui va être trop hermétique pour lui /
 donc indépendamment de la communication il y a tout
 de même une certaine forme forme de langage qui
 correspond plus ou moins à ce qu'on en attend / donc il
 y a une question à la fois de un peu didactique / mais
 peu parce que ce n'est pas parce que un quelqu'un qui
 est un qui préfère un **quo/** va lire La-Libre tout le temps
 qui pour ça va devenir plus intellectuel ça je ne crois pas
 que ça aura (rire) une influence jusque-là ni et vice-
 versa (...) [ilePA2r]

Dans les exemples qui suivent, c'est le contexte, mais également les phénomènes suprasegmentaux qui ont guidé le transcripateur vers une interprétation de [la] en amorce de *l'habitude* (exemple 4) et le *il n'a* (exemple 5) comme un lexème complet dans un énoncé inachevé et non comme une amorce de morphème.

(4) **iljDV1** eh ben on a trop trop **l'ha/** l'habitude je
 pense d'utiliser des expressions consacrées [iljDV1r]

(5) **marRI1** Francis **il n'a** il ne sait pas un mot de
 flamand hein et il / hein c'est un bruxellois / pur / et il
 parle pas un mot / tous les élever en français [marRI1r]

En général, ce travail d'interprétation se fait aisément, et l'orthographe standard peut être conservée.

Des cas difficilement interprétables sont ceux pour lesquels seule la première consonne est amorcée. Ainsi l'exemple suivant, dans lequel nous notons l'amorce par une transcription phonétique :

(6) **iljDV1** négativement // par rapport à un employeur sûrement (silence) quand l'accent est trop est trop important hein (xx) un petit accent mais ça je crois que c'est c'est normal / mais mais vraiment quand l'accent est / quand quand finalement on ne remarque que l'accent euh euh local euh je crois que **c'est [s] ça** peut influencer négativement euh l'avis des gens [iljDV1r]

Dans ce cas, l'amorce peut recevoir trois interprétations différentes :

1. la forme pleine est *c'est*, dont seul le début est répété : *c'est c' / ça* ;
2. la forme pleine est *ça*, qui est amorcé avant d'être complété : *c'est ç / ça* ;
3. l'amorce est inachevée : le son [s] est l'amorce d'un morphème qui ne sera ni corrigée ni complétée dans la suite de l'énoncé ; ce morphème commence par le son [s], qui dans la graphie du français est représenté par les lettres *s*, *c*, *sc* ou *ç*. Dans ce cas, pour respecter la règle de l'orthographe standard des transcriptions, on devrait donc proposer les quatre réalisations graphiques possibles.

La solution préconisée par le GARS lorsque plusieurs interprétations s'offrent au transcripteur est de les conserver en parallèle, dans des multitranscriptions (Raingeard et Lorscheider 1977).

Mais on le voit, ici, les possibilités d'interprétations sont nombreuses. Choisir l'une ou l'autre des deux premières solutions risque de guider la lecture vers une interprétation de l'amorce relevant d'un phénomène de répétition, soit de ce qui précède, soit de ce qui suit. Noter toutes les possibilités en parallèle risque de surcharger considérablement la transcription, demandant au lecteur un effort de décodage considérable. La solution préconisée par les conventions VALIBEL, dans ce cas, est de noter les cas indécidables de transcription d'amorces en [s] par la lettre *s*/. Évidemment, ce mode de transcription lui-même risque de guider la lecture vers cette seule interprétation. Mais des conventions détaillées qui explicitent ces choix permettent bien souvent d'éviter cet écueil²⁹.

Des amorces constituées par une seule consonne posent le cas particulier de la liaison que nous allons maintenant aborder.

²⁹ De la même manière, une amorce en [k] dont le transcripteur ne peut interpréter la forme pleine sera notée *q*/.

(7) **accCV1** on le reconnaît comme ça mais on n/ / on ne va jamais je crois en profondeur pour savoir d'où // d'où ça viendrait hein [accCV1r]

(8) **chaBR1** (...) et et et quand ma grand-mère est décédée à quatre-vingt-quatre ans il y a eu mille personnes à son enterrement / un enterrement comme on n/ on ne verra jamais plus à Herstal ou nulle part faudrait un roi pour voir ça [chaBR1r]

(9) **ileBH1** euh ça c'est sûr / mais on n/ non je s/ non euh là vraiment on ne sait pas / notre lecteur on sait qui il est euh c'est c'est vraiment le Belge moyen dans toute l'exception du terme ce n'est pas péjoratif / euh il eut un temps euh il y a une vingtaine d'années d'ici qu'on que je renvoyais encore bien un papier en disant ça mon vieux le chef porion du Borinage n'y comprendra rien [ileBH1]

Dans les exemples qui précèdent, le son [n] perçu a été transcrit par le chercheur comme une amorce de morphème, qui peut être interprétée comme la particule de négation *ne*, négation présente dans la suite de l'énoncé, ce qui induit cette interprétation de l'amorce : la forme complète est soit immédiatement reprise (dans les exemples 7 et 8), soit plus tardivement (dans l'exemple 9). Dans ce dernier cas, l'amorce peut aussi être interprétée comme l'amorce de *non*, complétée immédiatement après. Dans tous les cas, l'interprétation de [n] comme amorce ne semble pas poser de problème, et celle-ci, dans tous les cas, est complétée.

Dans les deux exemples qui suivent, le pronom *on* est également suivi du son [n], et le transcripteur a choisi également de l'indiquer dans le texte comme une amorce de morphème :

(10) **ilrDA1** mais un petit peu quand même je crois / au sein par exemple de l'école quoi oui je crois / quand on doit s'exprimer oralement ou euh devant les élèves ou quoi je sais pas on n/ on aurait tendance à à se moquer euh de la personne qui a un accent je crois oui [ilrDA1r]

(11) **norHJ1** hein en France donc je crois qu'il y a il y a évidemment une certaine souplesse mais qu'on manque d'indications enfin / je crois que si on n/ si on arrive déjà à dire // telle prononciation ou / cela // marquant les deux extrêmes enfin n'est-ce pas c'est c'est bien (xxx) quand vous aurez toujours des nuances selon les individus aussi n'est-ce pas et [norHJ1r]

Or, dans ces deux cas, on pourrait tout à fait interpréter la consonne perçue par le transcripteur non pas une amorce de morphème en tant que

telle, mais comme la marque de la liaison obligatoire du pronom *on* avec la voyelle du mot qui suit.

Dans une transcription qui ne note pas la liaison, ni la phonétique de manière générale, les notations proposées ci-dessus invitent le lecteur à interpréter la séquence comme une amorce corrigée : le locuteur amorce une forme commençant par le son [n], s'interrompt, reprend au pronom *on* (ou à la forme qui le précède, à la conjonction *si* dans l'exemple 11) et corrige en énonçant un autre morphème que celui qu'il avait précédemment amorcé. Dans ce cas-ci, une autre transcription serait possible, qui ne noterait tout simplement pas la liaison. La transcription proposerait ainsi au lecteur une alternative, qui ne forcerait pas la lecture dans le sens d'une amorce de morphème corrigée. La multitranscription, dans ce cas, se présente de la manière suivante, où la forme (ici l'amorce) en concurrence avec rien (puisque le phénomène phonétique de liaison n'est pas noté) apparaît entre accolades :

(10bis) **ilrDA1** mais un petit peu quand même je crois /
au sein par exemple de l'école quoi oui je crois / quand
on doit s'exprimer oralement ou euh devant les élèves
ou quoi je sais pas on {n/} on aurait tendance à à se
moquer euh de la personne qui a un accent je crois oui
oui [ilrDA1r]

(11bis) **norHJ1** hein en France donc je crois qu'il y a il y
a évidemment une certaine souplesse mais qu'on
manque d'indications enfin / je crois que si on {n/} si
on arrive déjà à dire // telle prononciation ou / cela //
marquant les deux extrêmes enfin n'est-ce pas c'est
c'est bien (xxx) quand vous aurez toujours des nuances
selon les individus aussi n'est-ce pas et [norHJ1r]

Une situation parallèle se présente dans l'exemple 12, où le transcripateur a noté une amorce de morphème, là où l'interprétation pourrait aussi bien être celle de l'amorce de la liaison obligatoire après le déterminant :

(12) **ilePA2** (...) j'essaye de me souvenir mais je n'arrive
pas à donner un n/ un exemple (x) (...) [ilePA2r]

La transcription suivante a le mérite de ne pas guider le lecteur de la transcription vers une seule interprétation :

(12bis) **ilePA2** (...) j'essaye de me souvenir mais je
n'arrive pas à donner un {n/} un exemple (x) (...)
[ilePA2r]

Néanmoins, si le transcripateur décide de classer cette liaison comme un phénomène purement phonétique et de ne pas le transcrire, cette marque de

l'élaboration de l'énonciation sera perdue, notée seulement, dans l'exemple ci-dessus, par la reprise du déterminant³⁰.

Les deux exemples suivants mettent également en jeu le phénomène de la liaison :

(13) **ileGG1** et -| je crois au niveau des z/ des des collègues aussi hein (silence) non je crois qu'il faut pas tellement compter sur la presse pour défendre un dialecte [ileGG1r]

(14) **ilcJG2** on maîtrise en moyenne un certain nombre de mots / euh et je ne crois que l/ la nationalité a une influence quelconque // peut-être que nous avons nous connaissons un peu plus de mots / de belgicisms de de mots de franglais que les z/ Français [ilcJG2r]

Le transcripteur a choisi, comme dans les exemples ci-dessus pour le son [n], de marquer le son [z]³¹ comme une amorce de morphème, qui dans les deux cas est aussi interprétée comme une amorce corrigée. Or, dans ces exemples, à la différence des cas 7 à 12, l'amorce est effectivement corrigée. Si supprimer la notation z/ dans l'exemple 13 n'enlève pas totalement la trace de l'élaboration de l'énonciation (la répétition reste présente dans la transcription), en supprimer la notation dans l'exemple 14 efface la trace du lapsus à peine ébauché.

Les pronoms personnels sujets de la troisième personne du masculin, singulier et pluriel, posent également des problèmes d'interprétation et de transcription. Bien souvent, en effet, les locuteurs prononcent *il* et *ils* [i], et [iz] dans les cas de liaison de *ils* avec le mot à initiale vocalique qui le suit. Dans la majorité des énoncés, ces formes sont identifiables et interprétables : elles ne peuvent en effet figurer qu'en position sujet, et la flexion du verbe, quand le contexte ne permettait pas de le faire, lève généralement l'ambiguïté sur le nombre du pronom.

Dans les trois exemples qui suivent, l'interprétation du [i] et sa transcription se fait sans trop de difficulté : *il* dans l'exemple 15, *ils* dans l'exemple 14. Par contre, dans l'exemple 17, si l'on semble bien avoir affaire au pronom sujet, il y a indécision quant à savoir si c'est le singulier ou le pluriel. Dans ce dernier cas, on proposera une multitranscription qui conserve les deux interprétations en parallèle, entre accolades {il , ils}.

³⁰ Ces traces de liaison dans l'énonciation sont en effet la marque de la planification faite par le locuteur. Voir Jefferson (1974) pour une interprétation de la prononciation de l'article anglais *the* dans les achoppements.

³¹ Tout comme le n/ ne notait pas nécessairement un morphème commençant par ce son, le mot amorcé en z/ a une forte probabilité de ne pas comporter cette lettre dans sa graphie.

(15) **marDL1** mais eux [i] n/ ils ne comprennent rien
(en criant aigu) [marDL1r]

(16) **marRI1** Francis il n'a il ne sait pas un mot de
flamand hein et [i] / hein c'est un Bruxellois / pur / et il
parle pas un mot / tous les élever en français [marRI1r]

(17) **ilrHP2** en général c'est beaucoup de ceux qui sont
fort cultivés / dans un peu surtout dans écoutez bon //
de ceux qui ont fait des études littéraires / comme des
des professeurs de français / c'est un peu leur métier les
inspecteurs / il me semble que [i] me ils ils manient
bien le français / (xx) [ilrHP2r]

Mais il est des cas, notamment dans les énoncés inachevés³², où l'interprétation du [i] en pronom sujet ne s'impose pas d'emblée, et où l'ambiguïté soit avec le pronom adverbial *y*, soit avec une amorce de mot est tout à fait possible :

(18) **ilrDA2** on ne s/ on ne s'en rend pas compte nous-
mêmes mais / si on va à l'étranger ou quoi [i] / ou ou
même non même pas à l'étranger aussi bien en Belgique
mais un peu plus loin / on verra bien qu'on a un accent
quoi mais on ne s'en rend pas compte / enfin moi je
crois que tout le monde a un accent [ilrDA2r]

(19) **ilrDC5** non je pense pas / je trouve que les
personnes allez dans chaque dans chaque ville ou quoi il
y a toujours moyen / en somme il y a plusieurs
personnes ou quand on compare des personnes comme
ça [i] je crois qu'elles se valent à peu près toutes à peu
près toutes pardon (rire) et je trouve que il y a pas
beaucoup de différence moi je crois que tout le monde a
un accent [ilrDC5r]

Dans ces deux exemples dont l'interprétation est indécidable, la multitranscription devrait donc mettre quatre formes en parallèle : *il*, *ils*, *y* et une amorce de morphème en *i/*.

La complexité d'une telle transcription rend la lecture du texte peu aisée ; le choix de l'une ou l'autre solution oriente inévitablement l'interprétation du lecteur. Peut-être que le mieux serait alors de laisser, comme nous l'avons fait dans les exemples ci-dessus (alors que ce n'était pas le cas dans les transcriptions originales), une indication de phonétique dans le texte en orthographe standard.

³² Pallaud (2002b : 89) note que les cas d'amorces les plus difficiles à interpréter sont ceux pour lesquels l'amorce est inachevée.

5.4.2 Les formes « sans signifié »

Il arrive que n'ayant pas de problème de perception en tant que tel, le transcripteur soit pourtant incapable de noter orthographiquement ce qu'il entend : c'est le cas pour les noms propres (anthroponymes, toponymes) mais également pour les titres (de livres ou de journaux) ou encore les termes techniques inconnus du transcripteur, pour lesquels celui-ci est incapable de reconstruire du sens.

Dans des transcriptions en orthographe standard, des variations considérables peuvent donc apparaître entre différents transcripteurs, pour un même passage.

5.4.3 Les croyances du transcripteur

La pratique de transcription confirme ce que prétend l'héroïne de Molière, Done Elvire³³ : « Souvent on entend mal ce qu'on croit bien entendre ». C'est ce qu'avancent également Blanche-Benveniste et Jeanjean (1987 : 102) : « l'oreille est un traître ; on écoute ce qu'on s'attend à écouter ». Les auteures (p. 6) citent Stubbs qui parle d'« hallucinations auditives » ; Cappeau (1997 : 120) utilise quant à lui les termes de « mirage auditif ».

Ainsi, bon nombre d'erreurs dans les transcriptions trouvent leur source dans l'image que le transcripteur a de la langue (sa représentation du code écrit, ses préjugés liés à l'oral, etc.) et du profil sociologique des locuteurs (Cappeau 1997, Gadet 2003). Par exemple, on constate que des étudiants qui transcrivent une conversation entre des membres de leur famille ont tendance à rétablir les *ne* de négation, là où aucun doute n'est possible quant à leur non-articulation. Ainsi, dans une analyse des erreurs que commettent les débutants, Cappeau (1997 : 124) constate que « les formes valorisées sont plus fréquemment attribuées aux locuteurs qui ont l'air d'avoir du prestige ».

6 TRANSCRIPTION PHONÉTIQUE VS TRANSCRIPTION ORTHOGRAPHIQUE

6.1 Transcription phonétique

Dans un article de 1979 publié dans *Le Français dans le monde*, Anne-Marie Houdebine préconise l'usage exclusif de la notation phonétique pour aborder l'enseignement de l'oral en classe et permettre une bonne connaissance linguistique de ses spécificités. On le sait, l'orthographe, et

³³ Dans *Dom Garcie de Navarre*, scène 3.

tout particulièrement l'orthographe française, est parfois très éloignée de la forme phonique des « mots ». Ainsi, selon Houdebine (1979 : 28-29),

loin d'être un avantage, la connaissance antérieure qu'on a de ce code [l'orthographe] est en fait une difficulté majeure : la graphie traditionnelle masque en fait plus que d'autres écritures la spécificité de l'oralité ; elle est si profondément inscrite en nous qu'elle détermine des erreurs de perception (...). En tant que linguiste, préoccupée de développer l'approche et la connaissance de la langue orale, nous ne pouvons donc que déconseiller aux enseignants de recourir à cette graphie s'ils veulent écouter, entendre et découvrir l'oralité.

Un argument avancé par Leroy (1985 : 10) est que le recours à la notation phonétique marque d'emblée la distance qu'il y a entre le texte transcrit de l'oral et un texte écrit standard : les deux ne peuvent être confondus, le décodage nécessaire de l'alphabet phonétique utilisé marque inévitablement la différence. Ainsi, oral et écrit ne peuvent être indument comparés sur la base de la transcription graphique. On peut dès lors

regarder l'écrit transcodé de l'oral autrement que comme de l'écrit, s'abstraire de la routine de l'œil façonnée par l'écrit (Gadet 2003 : 31).

C'est ainsi que, bien que privilégiant les transcriptions orthographiques, Raingeard et Lorscheider (1977 : 23) mettent en garde l'effet produit par le texte transcrit de l'oral (appelé ici corpus) :

Un corpus est nécessairement un texte bâtard dans la mesure où il restitue sous une forme écrite une organisation discursive et syntaxique propre à l'oral. Sa nature de texte écrit provoque chez le lecteur un réflexe culturel de référence plus ou moins consciente aux caractéristiques de l'écrit : il trouvera les « textes » du corpus peu économiques, mal organisés, etc. alors que ces jugements perdent toute pertinence lorsqu'il s'agit d'aborder un « texte » oral. Celui-ci a ses lois organisationnelles propres qui ne peuvent être reconnues qu'en dehors de toute référence à l'écrit.

Leroy avance également un autre argument en faveur de la transcription phonétique, argument qui concerne les homophonies. Le recours à une transcription phonétique possède l'avantage de ne pas orienter le sens et de masquer à tout jamais une interprétation qui n'aurait pas été perçue par le transcripteur :

(...) il est fort peu probable qu'un lecteur-utilisateur de la transcription pense à reconstituer l'oralité

correspondante jusqu'à retrouver l'ambiguïté recouverte
par la graphie (Leroy 1985 : 12).

Notons néanmoins que cet argument peut être utilisé dans l'autre sens. En effet, Leroy cite le cas de la séquence

[jav_:yyng_ãdf_t]

qui peut être interprétée soit comme « y avait eu une grande fête », soit comme « y avait u-une grande fête »³⁴, avec amorce du déterminant. Il n'est pas certain qu'à la lecture de la transcription phonétique, cette seconde interprétation serait émise par le lecteur de la transcription.

De plus, de nombreux cas problématiques dans les transcriptions ne sont pas liés à une homophonie, mais à une difficulté de perception qui empêche de trancher entre des solutions concurrentes et qui semblent pourtant parfois bien éloignées. Leroy cite [d'bons/d'beaux], [ben/mais], [alors/oh] et [main(te)nant/alors] en concluant : « il ne reste plus donc qu'à noter l'ambiguïté, et la seule graphie traditionnelle permet de le faire sans ridicule »³⁵.

S'il est vrai que nos habitudes graphiques orientent l'image que nous avons de la langue³⁶, et que cela est sans doute plus patent encore dans le cadre scolaire, où c'est sous sa forme quasi exclusivement écrite que la langue est apprise (combien d'enfants³⁷ savent que le son [o~], orthographié *on*, est en fait une voyelle ?), il semble que l'oral peut également être appréhendé à travers des transcriptions en orthographe standard.

Actuellement, aucun centre de recherche n'a retenu la transcription phonétique pour un enregistrement de plus de quelques minutes. En effet, celle-ci est :

- particulièrement couteuse en temps. Transcrire phonétiquement nécessite un apprentissage relativement long, et le travail de codage en lui-même prend un temps considérable ;

³⁴ Nous reprenons ici la transcription orthographique de Leroy.

³⁵ On ne voit pas en quoi ce serait plus ou moins ridicule en phonétique ou dans la graphie traditionnelle. Selon Leroy, la double notation de deux formes si éloignées que *maintenant* et *alors* donnerait une transcription qui « ne ressemblerait plus guère à une transcription phonétique crédible » (p. 11). Il s'agit ici d'une méconnaissance des phénomènes de perception.

³⁶ Pour les effets de l'écriture sur nos modes de pensée, voir Goody (1977) et Goody et Watt (1963).

³⁷ Pas seulement les enfants d'ailleurs. Carla Bruni, dans sa chanson *Raphaël*, confond lettres et sons : « Quatre consonnes et trois voyelles c'est le prénom de Raphaël, / Je le murmure à mon oreille et chaque lettre m'émerveille. »

-
- peu lisible, et totalement illisible pour des personnes non initiées : la lecture de tels textes ne repose pas sur nos habitudes de « native reader » (Sinclair 1995 : 102) ;
 - difficilement compatible avec des applications informatiques. C'était le cas de l'alphabet phonétique international (API). Ce n'est plus le cas avec l'alphabet Sampa (Speech Assessment Methods Phonetic Alphabet)³⁸, conçu spécialement pour permettre l'échange des données sur support informatique. Néanmoins, même en choisissant l'alphabet Sampa, les possibilités d'exploitation du texte s'en trouvent réduites (notamment avec des outils qui permettent des recherches automatisées dans les textes) ;
 - peu utilisable pour des recherches indépendantes de la phonétique/phonologie. La réutilisation des données nécessite toujours d'en repasser par une transcription orthographique.

Le recours à la phonétique ne peut être totalement abandonné (cf. le cas évoqué des amorces ; voir aussi § 6.2.5.) et il s'avère même à certains moments nécessaire pour éclairer la transcription graphique.

6.2 Transcription orthographique

Actuellement, pour les recherches linguistiques autres que celles concernant directement la forme phonique (phonétique, phonologie, prosodie), on privilégie donc l'orthographe traditionnelle.

6.2.1 Les trucages orthographiques

Dans certaines transcriptions, l'orthographe traditionnelle est adaptée, dans la volonté de marquer des particularités de prononciation. Ces particularités apparaissent à travers la graphie et non dans une notation phonétique. Ainsi, le *Guide de présentation de mémoires et de thèses en linguistique et conventions pour la transcription de conversations* (Losier et al. 2002) détaille les conventions que doivent adopter les étudiants pour la transcription de corpus acadiens. Voici un extrait de ces conventions :

- **chus** pour *je suis*.

Il importe ici de garder le *s*, puisqu'il s'agit du verbe à la première personne. Il serait possible de transcrire *j'suis*, mais ce ne serait pas exact à cause de l'élision du *i*.

- **j'sais, j'savais, j'chante**, etc.

³⁸ <http://www.phon.ucl.ac.uk/home/sampa/french.htm>.

Il y a élision du *je* devant une consonne.

- **j'sais pas** pour *je ne sais pas*.

Même si le *sais* est prononcé [e], il est préférable d'adopter cette transcription plutôt que *che pas*.

- **i** pour *il* et *ils*, **a** ou **alle** pour *elle*.

- **celle-cite** pour *celle-ci*

- **à c't'heure** pour *maintenant*.

- **qu** et **ti** prononcés [t_] sont écrits *tch*, comme dans **tchequ'un** (quelqu'un), **tchèque** part (quelque part), **moitché** (moitié), **pitché** (pitié), **tcheu** fou (quel fou).

On le voit, ces déformations graphiques affectent aussi bien des pronoms, des suites pronom+verbe que des lexèmes. Le but est clairement de calquer, dans la graphie, les variantes de prononciations répandues dans le français acadien.

Ces modifications graphiques participent de ce que Blanche-Benveniste et Jeanjean (1987) appellent des « trucages orthographiques »³⁹. « Trucages », parce que ces formes ne correspondent plus à des graphies connues d'un scripteur/lecteur de la langue, telles que répertoriées dans les ouvrages de référence par exemple, tout en y renvoyant néanmoins. Dans de telles transcriptions, on est entre les deux : ni tout à fait de l'oral, ni tout à fait de l'écrit.

Ces adaptations, qui peuvent sembler bénignes, ont en fait des conséquences majeures en termes d'effets produits sur le lecteur de la transcription⁴⁰ :

Ces transformations de l'orthographe, somme toute assez mineures, font toujours un effet péjoratif : elles signalent à l'attention un texte « populaire » et relâché. Il semble qu'elles n'aient jamais été suffisamment banalisées en français pour devenir un procédé de notation non marqué (Blanche-Benveniste et Jeanjean 1987 : 130-131).

³⁹ Giovannoni et Savelli (1990) parlent quant à elles de « bricolage orthographique », Gadet (2003) d'« aménagement graphique », Raingeard et Lorscheider (1977) de « bâtards phonético-orthographiques ». Dans le corpus édité par ces dernières, on est donc étonnée de trouver *main'ant* pour *maintenant*, *pa'ce* que pour *parce que*, *c'qui faut* pour *ce qu'il faut*, et *'pis* pour *et puis*, *i' veulent* pour *ils veulent*, *i' faut* pour *il faut*, *c't-à dire* pour *c'est-à-dire*, etc.

⁴⁰ Cette orthographe modifiée (« modified orthography ») relève de l'« eye dialect » (Edwards 1993 : 20, Gumperz et Berenz 1993 : 96).

On ne peut nier l'effet de dévalorisation qu'engendre une telle transcription sur les propos de la personne, et par là, sur la personne elle-même (Gumperz et Berenz 1993 : 97).

Ces altérations graphiques (notamment l'élision de voyelles) sont d'ailleurs fréquemment utilisées par les écrivains, et ce depuis le XIX^e siècle, pour stigmatiser le parlé de personnes issues de classes populaires. Ainsi, cet extrait d'un roman de Timothy Findley (traduction française de 1995), dans lequel s'expriment deux domestiques, Iris et Alberta :

« Vois-tu, dit Alberta, Mr. Harris s'est imaginé que Teddy d'venait vieux et qu'il allait être malade là-bas, ou s' perdre – ou s' faire mordre par un serpent, ou s' faire attaquer par un lapin et enfermer dans un terrier... C'est pourquoi ici – ici, en plein ville ! – s'étale la vieille Alberta Perkins, dans les vapeurs tropicales de l'été ! Pendant ce temps, où crois-tu, bonté divine, où crois-tu que ce teigneux de chien s' trouve ? Hein ? Peux-tu me dire où il est ? Devine. Vas-y. Je t' le donne en mille ! Hooker ?

- J' sais pas, dit Iris.

- Je ne sais pas », dit Hooker.

On voit le foisonnement d'élisions censées calquer la prononciation des bonnes. Par contraste, Hooker, le fils de la maison, s'exprime selon la norme. La différence est d'autant plus marquée qu'il prononce la même phrase qu'Iris, sans élision et en rétablissant la particule négative *ne*. Notons que, dans le roman, tous les dialogues entre des personnages de la bourgeoisie ne se distinguent d'aucune manière de la langue de la narration. L'effet produit, et voulu par l'auteur, est ici très clair : les procédés graphiques utilisés, et qui abondent à chaque prise de parole des bonnes (*et t'nez, j' descends, j' dis, me v'là, p'tit, à demain qu'y m' répond*), jouent sur les stéréotypes du parlé attribué à des classes sociales défavorisées.

Le trucage orthographique, qu'on n'utilise que pour le 'populaire', peut donc difficilement passer pour un innocent procédé de transcription (Blanche-Benveniste 2000c : 26).

Dans les publications scientifiques, l'effet est le même que dans les œuvres romanesques. Regardons le texte suivant, transcription d'une conversation entre élèves de CE2.⁴¹

⁴¹ Cette transcription est extraite de Verstraete (2000 : 154-156). Nous avons respecté strictement la typographie et la mise en page. Le gras note que l'élève lit ; les majuscules marquent l'accentuation ; les doubles barres l'hésitation entre deux interprétations /« beuh », peur/ ; E indique une indétermination du locuteur. Ponctuation et didascalies ont été ajoutées par l'auteure en vue de cette publication.

Olivia : Oh ! le hamster ?

Sélalé : Hein!! Eh elle sait même pas c' que c'est qu'un hamster !!!

Lévy : C'est une bête hé !

Sélalé : C'est un p' tit truc : ça ressemble un peu à une souris !!

Olivia : (*très sure d'elle*) Mais NON ! un HAMSTER ! Pas un animal !

Lévy : Un hamster c'est un animal.

Sélalé : Ben oui c'est un p'tit animal qui ressemble un peu à une souris...

Olivia : (*rires moqueurs*) C'est un animal ? un hamster ! ...Eh r'

garde (*parle au magnétophone, en articulant bien*) Un hamster est un

animal ! (*rires moqueurs*) ... (*puis se rangeant brutalement à l'avis*

des autres) Ouais, j'en ai un chez moi.

Kévin : (*plus vite*)...**que faisaient les Gaulois pour effrayer les**

esprits ? (*lit la question 3*)

Tarik : arrête Vincent !

Vincent : Y f'saient /« Beuh », « peur »/ !

EEE : *Rires*

Vincent : Ou bien y cassaient leurs couilles

Rire de Karen

Maîtresse : chut

Vincent : y faisaient /la mort/

Tarik : Attends fais voir...

Kévin : Ils tuaient les gens...

Karen : Non !

Vincent : Ben ouais p'têt ça...

Karen : Non quand même pas ! ...Non les Gaulois y faisaient pas

ça ! ... Je sais qu'ils ont tué... mais...

On voit les trucages utilisés ici : élision marquée au moyen de l'apostrophe (*p'tit, c'que, p'têt, f'saient, r'garde*), ce qui produit un effet d'ailleurs assez étrange pour *r'garde*, dont la mise en page laisse le *r'* seul en

fin de ligne ; utilisation de la graphie *y* pour noter la prononciation du pronom *ils*⁴².

Par ailleurs, informant le lecteur du cadre sociologique dans lequel s'est effectué son travail, l'auteure dit (pp. 153-154) que la classe « est située dans une zone urbaine sensible », que six élèves (sur vingt) ont déjà doublé une fois, que « la plupart des élèves ont un rapport au savoir qui n'est pas toujours direct », que certains ont des difficultés psychologiques et sont suivis par un psychologue. Elle précise également que les paroles des enfants n'ont pas été censurées : « elles sont retranscrites telles quelles dans la crudité du langage des élèves...⁴³ ».

Inévitablement, les choix de transcription et le profil des élèves présentés ne font que cumuler leurs effets négatifs. La stigmatisation déjà possible dans la description du profil des élèves se voit renforcée par les choix de transcription.

En fait, comme le dit Gadet (2003 : 30) :

seule une graphie sans aménagement ni réécriture limite
le risque de stigmatiser un énoncé.

6.2.2 Les inconvénients des trucages

Outre cet effet de stigmatisation, nous voyons quatre grands inconvénients dans ces trucages.

1) Selon les phénomènes concernés, ces aménagements graphiques sont difficiles à systématiser. Ce manque de systématisation est parfois involontaire, parfois volontaire. À l'intérieur d'une même transcription, ils ne concernent que certaines formes et pas d'autres pour lesquelles le même trucage aurait pu être choisi. Le trucage attire donc l'attention sur les formes marquées, pouvant ainsi laisser croire que les autres ne seraient pas touchées par des variations de prononciation.

Concernant le trucage qui consiste à marquer l'élision du *e* muet, Leroy (1985 : 14) se demande dans quelle mesure on peut noter dans tous les cas. Doit-on aussi écrire *ami'* pour *amie*, *presq'* pour *presque*, *sort'* pour *sortent*, etc. ?

Exemple de non-systématisation, le corpus qui sert de base à l'ouvrage de Cosnier et Kerbrat-Orechionni (1987 : 374) : les auteurs avouent pratiquer l'élision dans leurs transcriptions, mais de façon ponctuelle : « Les voyelles

⁴² Risquant par là de rendre la transcription ambiguë, *y* étant dans nos habitudes de lecture le pronom ou la lettre de l'alphabet.

⁴³ Les points de suspension sont le fait de Verstraete.

caduques et les syllabes élidées sont indiquées, *occasionnellement, mais pas systématiquement*, au moyen d'une apostrophe, par extension de l'usage conventionnel de l'écrit. »⁴⁴

De même Luzzati (1985 : 39), alors qu'il ne transcrit pas les « amuïssements courants (/ot/, /dsy/, ...) », « respecte les altérations morphologiques majeures (/ja/, /pask/, ...) »⁴⁵, altérations pourtant tout aussi courantes que les amuïssements qu'il ne transcrit pas. De plus, cette transcription des altérations morphologiques ne concerne que les « majeures ».

2) Il est parfois difficile de distinguer des variantes de prononciation à la seule écoute du signal sonore. Ainsi, la prononciation de *il y a* présente quatre réalisations différentes ([ilija], [ilja], [ija] et [ja]) qu'il est parfois malaisé d'identifier avec certitude.

3) Ces déformations perturbent nos habitudes de lecture en compliquant le système graphique, et obligent le lecteur à oraliser pour décoder ;

4) Ce mode de transcription complique la tâche du transcripteur, surchargeant son travail par l'attention qu'il doit porter à ces phénomènes, dont la fréquence est souvent très élevée.

Il faut par ailleurs noter que ces altérations graphiques servent à noter des faits de prononciation non seulement très fréquents, mais aussi attestés depuis longtemps (Gadet 1992 et 1997). Néanmoins, fréquence et ancienneté des phénomènes concernés n'en font pas moins que ces choix de transcription sont stéréotypants, car éloignés des habitudes graphiques du français écrit normé.

6.2.3 Les recommandations EAGLES et NERC

Les organismes qui se sont préoccupés de normalisation des ressources, dans le but notamment de favoriser leur échange, se sont bien évidemment penchés, dans le cadre spécifique de la constitution de corpus oraux, sur cette question de la mise par écrit du signal sonore. Pour ce qui concerne des recherches en linguistique de corpus, par opposition aux études qui s'intéressent spécifiquement à la matière sonore (« speech research », voir Llisterri (1996)), les transcriptions orthographiques sont clairement recommandées.

Aussi bien NERC (Network of European Reference Corpora, voir French (1992)) que EAGLES (Expert Advisory Group on Language Engineering

⁴⁴ C'est nous qui soulignons.

⁴⁵ *parce que* est donc noté *pasque*.

Standards, voir Llisterri (1996)) préconisent l'orthographe standard, qui se base sur un ouvrage de référence clairement identifié comme tel par les transcrip-teurs. NERC propose pour l'anglais le *Oxford English Dictionary* ; le *Robert* ou le *Trésor de la langue française* sont traditionnellement identifiés comme ouvrages de référence pour le français, et cela d'autant plus que le dernier est maintenant consultable gratuitement via internet.

En ce qui concerne les réductions graphiques, ces organismes recommandent de les éviter et de ne pas faire apparaître les altérations phonétiques visuellement dans la forme des mots : pas de trucage orthographique donc.

Les variantes non standard, qui ne sont pas recensées dans les ouvrages de référence mais qui seraient néanmoins utilisées par les transcrip-teurs, doivent être listées et documentées. Cela permet de garantir que le corpus sera transcrit de manière uniforme.

6.2.4 Cas particuliers : l'apocope et l'aphérèse

Certaines réductions phonétiques affectent les mots, tronquant ceux-ci d'une ou de plusieurs syllabes à l'initiale (aphérèse) ou en finale (apocope). Ce phénomène, s'il peut sembler comparable aux altérations phonétiques mentionnées ci-dessus, s'avère pourtant très différent : il affecte au moins une syllabe et constitue une réduction volontaire de la part du locuteur. Certaines de ces formes sont attestées dans les dictionnaires soit parce qu'elles sont devenues la norme (*mé-tro*, *au-to*, *ki-lo*, etc.), soit parce qu'elles figurent dans l'article comme variantes morphologiques (*prof*, *math*, *per-pète*, *stat*, *ré-pète*, etc.). D'autres, par contre, sont absentes des ouvrages de référence : *unif*⁴⁶, *mayo*, *bon ap*, etc.

Plusieurs solutions s'offrent pour noter ces troncations : marquer la coupure au moyen d'une apostrophe, par exemple, aurait pour conséquence de stigmatiser la forme, alors que le procédé de création est comparable aux autres apocopés attestés ; noter la forme comme une amorce de mots rend cette catégorie non cohérente, englobant dans la même classe des phénomènes différents, ne relevant pas de conditions de productions semblables ; transcrire la forme tronquée sans autre indication a pour conséquence de ne pas faire de lien dans la transcription (et par conséquent dans des recherches automatisées) entre celles-ci et les formes pleines dont elles dérivent.

On le voit, la solution adoptée pour transcrire un phénomène particulier a des implications sur les analyses ou les possibilités de recherches ultérieures.

⁴⁶ Le *Petit Robert* donne *uni*, forme attestée en France mais pas en Belgique.

6.2.5 Les notations phonétiques qui accompagnent les transcriptions en orthographe standard

Même lorsque l'orthographe traditionnelle est utilisée, le besoin de recourir à des indications phonétiques peut se faire sentir. Nous voudrions illustrer cela dans les deux exemples suivants qui mettent en jeu des phénomènes de reprises présents dans le texte. Le recours à la phonétique permet d'éclairer la reprise.

Si certains bafouillages ou certains lapsus n'entraînent pas de modification sur l'axe linéaire de l'énonciation (le locuteur poursuit son propos « comme si de rien n'était »), il arrive que le locuteur se corrige, revenant en arrière et répétant correctement la forme sur laquelle il a achoppé. Les deux exemples suivants illustrent le phénomène de lapsus et la correction par la reprise de la bonne forme :

(1) **pluGM1** enfin je comprends tout [t_u] / je comprends tout mais je le parle pas [pluGM1r]

(2) **accCP1** (...) un un Borain va tout de suite voir celui-là c'est un Liégeois / et chacun [Saka~] / chacun prêchant un petit peu pour sa chapelle / bon les Carolos trouvent que les Carolos sont les meilleurs les Liégeois trouvent que les Liégeois sont les meilleurs (...) [accCP1r]

Ces deux exemples se distinguent par le fait que dans le premier, le locuteur répète l'énoncé depuis le sujet *je* alors que dans le second, seule la forme posant un problème est reprise. En fait, la différence fondamentale pour ce qui nous intéresse ici est que dans le deuxième exemple, le lapsus donne lieu à une forme non répertoriée parmi les lexèmes de la langue, qu'on pourrait orthographier *chacan*, ce qui n'est pas le cas dans le premier exemple où la graphie calquant la prononciation, *trou*, donnerait bien une forme du français. Cela implique que dans le premier cas, la transcription en orthographe engendre une forme existante et qui sera prise en compte comme telle (dans les travaux de statistiques, dans les logiciels, etc.), même si ce n'est pas ce lexème-là qui devrait être concerné, mais celui à l'origine du bafouillage (en l'occurrence, *tout*) ; dans le cas de *chacan*, la forme sera considérée comme inconnue et ne recevra pas d'analyse avec ces mêmes outils.

La situation, en termes d'analyse, est comparable pour toutes les « erreurs » de prononciation, qu'elles soient le fait d'un cafouillage ponctuel non corrigé ou d'une habitude du locuteur (mais comment savoir si l'on est devant l'un ou l'autre ?). Dans tous ces cas, transcrire les « mots

bafouillés »⁴⁷ selon les normes graphiques de la forme standard permet de les répertorier comme des morphèmes de la langue ; les faire suivre de la prononciation en cas d'auto-correction éclaire le phénomène en jeu.

Mais on utilisera ces annotations phonétiques avec prudence. Elles ne doivent pas surcharger le texte, et ne pas faire apparaître comme remarquables des phénomènes de prononciation tout à fait ordinaires.

6.2.6 Ponctuation

Les corpus de langue parlée transcrits à des fins de recherche linguistique ne sont en général pas ponctués. En fait, à la différence de ce que croit le héros de David Lodge cité en exergue, il n'y a pas de correspondance entre pause orale et ponctuation écrite. À une petite pause ne correspond pas nécessairement une virgule à l'écrit, à une pause plus grande une ponctuation forte⁴⁸. Il n'y a pas de relation bi-univoque entre les deux.

De plus si certains énoncés se laissent enfermer relativement facilement dans le moule de la ponctuation graphique, comment s'en sortir avec des productions dans lesquelles foisonnent répétitions de mots, inachèvements et autres « scories » propres à l'oral ?

Blanche-Benveniste et Jeanjean (1987 : 139) plaident pour des transcriptions non ponctuées :

La ponctuation, si on la met trop tôt, préjuge de l'analyse syntaxique et impose un découpage sur lequel il est difficile de revenir.

Selon elles, en ponctuant, le transcripteur « suggèr[e] une analyse avant de l'avoir faite » (id. : 142).

En fait, l'absence de ponctuation va de pair avec la remise en cause de la notion de phrase à l'oral. Ainsi que le dit Gadet (1992 : 69) :

Comme pour toutes les études de phénomènes oraux, la séquence fondamentale ne correspond généralement pas à ce que l'on entend par « phrase » à l'écrit. Il faut donc se passer de cette catégorie.

⁴⁷ Lorsque cela est possible évidemment. Nous parlons des cas où la langue a « un peu fourché », et où l'analyse est toujours possible. Par ailleurs, notre propos ne concerne pas la transcription de paroles pathologiques, par exemple.

⁴⁸ Nous parlons ici d'oral non planifié et non d'oral obtenu à partir d'un texte lu.

C'est ainsi que la notion de phrase, dans les études de productions orales, a été abandonnée et que s'est posée la question de l'unité minimale pertinente à prendre en considération⁴⁹.

6.2.7 L'unité de transcription

À l'écrit, la division d'un texte en ses différents éléments est relativement claire, instituée, codifiée, et nos habitudes de lecteurs natifs (« native reader », Sinclair 1995 : 102⁵⁰) nous permettent de les identifier aisément. La plupart du temps, un texte possède un titre et est accompagné du nom de son auteur ; il est constitué de mots séparés par des blancs, qui se regroupent eux-mêmes en phrases, puis en paragraphes et encore en chapitres :

Anyway the standardization of the representation of the written language was perhaps the most important development of that age, one that makes it so easy for us to handle the written language in machines. (Sinclair 1995 : 100)

Dans les transcriptions, cette disposition spatiale ne va pas de soi. Elle relève de choix qui reflètent des partis-pris théoriques concernant l'unité d'analyse. En effet, selon Mondada (2000), les auteurs intègrent leur position théorique dans la présentation des données ; la segmentation du flux oral se fait alors en tours de paroles, en unités intonatives, etc.

7 LA REPRÉSENTATION SPATIO-VISUELLE DES DONNÉES. LE TOUR DE PAROLE

La question de l'unité pertinente à l'oral pose un problème (cf. chap. 4, § 3), et une solution afin de ne pas faire intervenir cet aspect théorique dès la transcription est de considérer l'unité, qui serait alors envisagée comme relativement neutre, du tour de parole. On donne alors au tour de parole un statut non pas théorique mais purement graphique, qui devient une unité (visuelle) par défaut, unité pratique dotée de caractéristiques « objectives » qui ne seraient pas définies par des partis-pris théoriques⁵¹.

⁴⁹ Voir l'ouvrage collectif dirigé par Berthoud et Mondada (2000) pour la position des différents auteurs sur le sujet ; voir aussi Béguelin (2000 et 2002).

⁵⁰ Voir aussi le concept de « literate native (or similarly expert) users » chez Cook (1995 : 38).

⁵¹ Marty Laforest (1992 : 31) note à propos du tour de parole : « Cette notion est toujours donnée, jamais construite ».

Définir le tour de parole de façon « neutre » revient à l'envisager de la sorte : un tour de parole correspond à l'occupation matérielle du canal de parole par un locuteur (Simon 2004 : 97) ; le tour de parole s'achève lorsqu'un nouveau locuteur prend la parole à son tour.

Dans les transcriptions, la succession des tours de parole peut se présenter de manière horizontale (version A) : les paroles des locuteurs se succèdent, de haut en bas sur la « page » (ou à l'écran) et chaque paragraphe⁵² représente l'intervention d'un locuteur.

Version A

ilrMF0 et quand tu es sur scène euh ce qui est le plus important pour toi / c'est quoi

ilrLP1 jouer

ilrMF0 oui mais c'est-à-dire

ilrLP1 mais c'est-à-dire être à fond dans mon personnage et dans l'histoire / à la limite j'ai j'ai peur avant mais une fois que je suis sur scène ça va la peur passe et et j'oublie un peu que le public est là

ilrMF0 m parce que justement tu parlais du contact avec le public

ilrLP1 mm

ilrMF0 qui était important

ilrLP1 oui oui non je sais je s/ il y a de toute façon un contact qui doit passer et / oublier totalement c'est pas possible parce qu'on joue pour un public mais / on joue pour soi aussi (raclements de gorge)

Cette disposition, que Edwards (1995) appelle le format vertical, est cohérente avec nos habitudes de lecture : on commence par lire en haut de la page, et ce qu'on lit d'abord se déroule dans le temps avant ce qu'on lit ensuite, plus bas sur la page. Les textes des pièces de théâtre adoptent depuis longtemps ce format, qui fait partie de notre culture pour représenter la succession des paroles des protagonistes de la conversation. Edwards oppose cette disposition au format en colonnes (« columnar » format)⁵³, qui donne à la transcription ci-dessus la matérialisation suivante :

⁵² On parle ici de paragraphe au niveau graphique, car au niveau informatique il ne s'agit que d'une seule ligne.

⁵³ Notons que la disposition verticale n'est pas absente de ce format.

Version B

ilrMFO	ilrLP1
et quand tu es sur scène euh ce qui est le pluS important pour toi / c'est quoi	
	jouer
oui mais c'est-à-dire	
	mais c'est-à-dire être à fond dans mon personnage et dans l'histoire / à la limite j'ai j'ai peur avant mais une fois que je suis sur scène ça va la peur passe et et j'oublie un peu que le public est là
m parce que justement tu parlais du contact avec le public	
	mm
qui était important	
	oui oui non je sais je s/ il y a de toute façon un contact qui doit passer et / oublier totalement c'est pas possible parce qu'on joue pour un public mais / on joue pour soi aussi (raclements de gorge)

Dans les deux cas, le tour de parole constitue l'unité de transcription ; les mêmes informations sont fournies à la lecture, seule diffère la présentation spatiale des données.

L'hypothèse d'Edwards, déjà défendue par Ochs (1979), est que le format des données peut avoir une répercussion sur l'analyse, parce qu'elle induit inévitablement des effets sur le lecteur (le chercheur) : la version A impliquerait selon Edwards un statut égal des locuteurs, et une influence équivalente de l'un sur l'autre. Par contre, la disposition spatio-visuelle de la version B donnerait quant à elle une importance plus grande au locuteur le plus à gauche, en en faisant le locuteur dominant. Dans ce cas, ce sont aussi nos habitudes de lecture qui sous-tendent cette interprétation, puisqu'on lit en commençant par la gauche et en allant vers droite. Ochs, dont l'article porte plus spécifiquement sur les implications théoriques de choix de transcription dans le cadre particulier de dialogues adulte-enfant, faisait également ce constat : le locuteur auquel on attribue la colonne de gauche est celui qui semble contrôler la conversation, et auquel le transcripteur réserve la position dominante de cette colonne. Ochs constate que dans la littérature, la tendance des chercheurs est de placer l'adulte à gauche et l'enfant à droite⁵⁴ (elle note d'ailleurs que les interactions sont la plupart du temps décrites en termes de dialogue adulte-enfant, plutôt que enfant-adulte). Selon elle, si cette position de contrôle de la colonne de gauche peut refléter l'idée que l'adulte mène la conversation, par un effet de circularité, elle la renforce indubitablement. En effet, parcourant la transcription de gauche à droite, le lecteur est orienté dans l'idée que le locuteur qui initie l'échange est celui de gauche, l'adulte : c'est lui qui mène la conversation et la contrôle, même lorsque ce n'est pas le cas dans la réalité de l'interaction. Le point le plus à gauche a tendance à devenir le point de départ de l'échange dans tous les cas. Ochs parle de candidat « naturel » comme point de départ interactionnel :

Starting points become constraining rather than
constrained variables in an assessment of social control.
(Ochs 1979 : 51)

Elle propose alors, pour contrecarrer cette tendance, de mettre les paroles de l'enfant dans la colonne de gauche et celui de l'adulte à droite⁵⁵.

⁵⁴ On peut imaginer que dans le cadre d'entrevues, la tendance naturelle du transcripteur serait de mettre les paroles de l'interviewer dans la colonne de gauche, avec l'idée que c'est lui qui mène la conversation, et celle de l'interviewé à droite (c'est ce que nous avons d'ailleurs fait ci-dessus).

⁵⁵ Selon Cook (1993 : 51), "Yet it would be mistaken to suppose that such biases *can* ever be corrected; they are rather an inevitable consequence of transferring speech into writing".

7.1 Les chevauchements

La définition donnée ci-dessus du tour de parole ne prend pas en compte les cas très fréquents dans lesquels les paroles des locuteurs se chevauchent. La définition « neutre » du tour de parole tient compte de l'occupation du canal par le locuteur ; un chevauchement de parole se produit donc lorsque au moins deux locuteurs occupent le canal simultanément. Deux cas de figure se présentent : 1) un locuteur L2 prend la parole alors qu'un locuteur L1 la possède déjà ; le locuteur L2 poursuit tandis que le locuteur L1 arrête de parler ; 2) un locuteur L2 prend la parole alors qu'un locuteur L1 la possède déjà ; les paroles du locuteur L2 chevauchent les paroles de L1 qui, malgré cette intervention, poursuit et, donc, « ne perd pas la parole ».

Dans le format de transcription en colonnes, une manière de représenter ce phénomène est de noter les paroles chevauchées sur une même ligne, en vis-à-vis. Ainsi, si les paroles de ilrMF0 (*oui mais c'est-à-dire*) et de ilrLP1 (*mais c'est-à-dire être à fond (...) j'oublie un peu que le public est là*) étaient prononcées en même temps, la Version B prendrait la forme suivante :

Version C

ilrMF0	ilrLP1
et quand tu es sur scène euh ce qui est le pluS important pour toi / c'est quoi	
	jouer
oui mais c'est-à-dire	mais c'est-à-dire être à fond dans mon personnage et dans l'histoire / à la limite j'ai j'ai peur avant mais une fois que je suis sur scène ça va la peur passe et et j'oublie un peu que le public est là
m parce que justement tu parlais du contact avec le public	
	mm
qui était important	
	oui oui non je sais je s/ il y a de toute façon un contact qui doit passer et /

	oublier totalement c'est pas possible parce qu'on joue pour un public mais / on joue pour soi aussi (raclements de gorge)
--	---

Mais là encore, nos habitudes culturelles nous font débiter la lecture à gauche, donnant ainsi l'impression que c'est le locuteur de cette première colonne qui est le premier à parler et que les paroles du second locuteur chevauchent les siennes, et non l'inverse (Ochs 1979). Ainsi, la spatialisation visuelle a des répercussions sur l'interprétation temporelle du déroulement de la parole des locuteurs.

7.2 Le chevauchement : deux autres représentations

Le GARS utilise le soulignement pour noter les chevauchements de parole, comme ci-dessous :

transcription du GARS⁵⁶

L2 on le tirait par la queue

L1 il y avait il y a avait on le tirait par la queue voilà eh c'est

Le même échange, selon les conventions VALIBEL, prend la forme suivante, où la barre et le tiret délimitent la portion de texte chevauchée (tiret à droite de la barre pour le début du chevauchement, tiret à gauche de la barre pour la fin du chevauchement).

transcription de VALIBEL

L2 on le tirait |- par la queue

L1 il y avait il y a avait -| on le tirait par la queue voilà eh c'est

Dans ce cas, où, à la fin du chevauchement, L1 qui avait la parole cesse de parler et L2 poursuit, aucune différence n'apparaît à la lecture des deux formats de transcription.

Par contre, les conventions des deux centres impliquent une disposition spatiale différente lorsque, après le chevauchement de parole, le locuteur qui avait la parole la conserve :

transcription du GARS⁵⁷

⁵⁶ Cette transcription est extraite de Blanche-Benveniste *et al.* (2002 : 47).

L1 ces choses-là j'en

L2 plus ou moins

L1 parle – je ne raconte pas une je vais pas raconter
tout ça comme ça tu vois

transcription de VALIBEL

L1 ces |- choses-là j'en <L2> plus ou moins -| parle / je
ne raconte pas une je vais pas raconter tout ça comme
ça tu vois

Dans les transcriptions du GARS, le chevauchement de L2 apparaît visuellement dans un tour de parole séparé, alors que dans les conventions VALIBEL il est inséré dans la ligne d'un autre locuteur : il est de ce fait moins visible et risque de ne pas être interprété comme un tour de parole à part entière⁵⁷. Mais la différence la plus flagrante est sans doute le sort réservé au tour de parole de L1 : le découpage opéré par le GARS en fait deux tours de parole, là où VALIBEL en conserve un seul.

Ainsi, la représentation visuelle d'un même phénomène a des implications sur les unités de segmentation du discours, alors même que celles-ci avaient reçu une définition ne semblant pas devoir prendre en compte ce type de variation. Cette disposition différente a des répercussions directes sur le décompte des tours de parole : pour cette courte séquence, trois au GARS, contre un seul chez VALIBEL.

7.3 Les continueurs

Selon la définition donnée ci-dessus du tour de parole et du chevauchement, on peut conclure à la lecture de la transcription proposée ci-dessus, aussi bien la version A que la version B, que le *mm* de ilrLP1 se produit pendant une pause de ilrMF0 ; celui-ci poursuit ensuite avec une relative, qui peut être rattachée directement à son tour de parole précédent. Ce *mm*, appelé couramment *continueur*, a pour fonction d'inciter le locuteur à poursuivre son propos.

Dans certaines transcriptions, alors que le tour de parole n'a pas un statut théorique revendiqué comme tel mais semble être une unité spatio-temporelle, ces continueurs ont tendance à être systématiquement notés à l'intérieur du tour de parole du locuteur, comme s'ils étaient toujours

⁵⁷ La pause brève est symbolisée par un tiret au GARS, par une barre oblique entourée de blanc chez VALIBEL. Cet extrait est issu de Blanche-Benveniste *et al.* (2002 : 152).

⁵⁸ La définition de ce chevauchement par VALIBEL va de pair avec sa disposition spatio-visuelle : « chevauchement interne au tour de parole du locuteur ».

ilrMF0 le public qui était important
 ilrLP1 mm oui oui non je sais je s/ il y a de toute façon un

ilrMF0
 ilrLP1 contact qui doit passer et / oublier totalement c'est pas possible parce qu'on joue pour

ilrMF0
 ilrLP1 un public mais / on joue pour soi aussi (raclements de gorge)

Ici, à chaque locuteur est attribuée une ligne ; les retours en début de ligne sont matériellement générés par le manque de place à la ligne précédente : si la largeur de la page pouvait être extensible à souhait, on n'aurait qu'une unique ligne par locuteur. Selon Edwards (1995 : 11-12), l'avantage de cette présentation est qu'elle est particulièrement efficace pour représenter le déroulement temporel des paroles (sur l'axe horizontal) et leur simultanéité lors des chevauchements (sur l'axe vertical) – comme dans le format en colonne – ; de plus, elle garantit un statut égal aux participants – comme dans le format vertical. On peut douter de cette dernière affirmation et émettre l'hypothèse que, tout comme la colonne de gauche donne une prédominance au locuteur qui l'occupe, le locuteur de la ligne du haut, conformément à nos habitudes de lecture, sera considéré comme celui menant d'une certaine manière la conversation⁶¹.

Il est des cas où cette disposition ne suffit pas à noter de manière efficace le chevauchement : en effet, selon qu'un locuteur parle vite ou lentement, la longueur des paroles à transcrire peut varier considérablement pour une même durée ; à cela s'ajoute le facteur de la taille d'un mot graphique, qui ne correspond pas nécessairement à sa durée à l'oral. Ainsi, si l'alignement vertical note facilement le début du chevauchement, le système d'Ehlich insère une marque pour noter la synchronisation de la fin des paroles des deux locuteurs⁶². Nous reprenons son exemple (Ehlich 1995 : 131).

In : What happened to Bill after his disappearance ?|__
 Su : Bill had a pretty hard time.|

⁶¹ Ainsi, dans les transcriptions faites dans des classes, on verra attribuée préférentiellement la ligne du dessus au professeur, et à l'interviewer dans les entrevues.

⁶² Jefferson propose aussi des crochets pour noter le début et la fin des chevauchements (voir Atkinson et Heritage 1999 : 158-159) ; pour le français, ce système est également utilisé dans l'ouvrage dirigé par Cosnier et Kerbrat-Orecchioni (1987).

Un inconvénient majeur de ce type de transcription est qu'il relève d'une certaine manière d'un bricolage manuel pour aligner les paroles des locuteurs. Ainsi, si une nouvelle écoute du son nécessite l'introduction ou le retrait d'un passage, même très court, tout le travail d'alignement est à refaire, non seulement pour la portion concernée mais pour l'ensemble de la transcription. De la même manière, si la police de la transcription change. Voici l'échange ci-dessus non plus en Tahoma corps 9 mais en Arial corps 10.

In :	What happened to Bill after his disappearance ? __
Su :	Bill had a pretty hard time.

À une époque où la diffusion des données doit guider l'établissement de conventions non ambiguës et aisément échangeables, ce format semble peu adéquat (cf. § 8).

7.5 Les informations associées aux tours de parole

Le tour de parole caractérise l'activité langagière d'un locuteur particulier. Dès lors, l'insertion d'une indication dans un tour est interprétée comme étant le fait de ce locuteur. Or, ce n'est pas toujours le cas : de nombreuses indications soit paralinguistiques, soit contextuelles sont associées aux tours de parole, qui ne sont pas à rapporter au locuteur dont c'est le tour (interruptions par des rires, des bruits divers, etc.) (Mondada 2000).

L'indication dans les transcriptions de certains faits, attribués préférentiellement à un locuteur plutôt qu'à un autre (puisque l'indication figure dans son tour de parole), peut se répercuter dans l'analyse. Ainsi en est-il des pauses inter-tour.

La notation des pauses entre les tours de parole se fait ou non selon les conventions adoptées. Ainsi, les conventions de Jefferson pour l'analyse conversationnelle marquent explicitement l'enchaînement (ou non) des tours de parole (Atkinson et Heritage 1999), allant jusqu'à les quantifier ; les transcriptions du GARS ou de VALIBEL omettent cette information.

En ce qui concerne les pauses prolongées, appelées aussi silences, certaines conventions les notent lorsqu'elle se produisent entre deux tours de parole soit systématiquement en fin de tour (comme c'est le cas dans les conventions VALIBEL, version 1) soit, au contraire, au début du tour qui suit ce silence (Version 2).

Version 1

famHC1 ben oui euh oui au |- repas oui <famED1> le
soir -| (silence)
famES1 tiens et ta maman elle travaille toujours à la /
boucherie là |- en bas ?

Version 2

famHC1 ben oui euh oui au |- repas oui <famED1> le
soir -|
famES1 (silence) tiens et ta maman elle travaille
toujours à la / boucherie là |- en bas ?

Dans la version 1, le silence sera préférentiellement imputé aux locuteurs famHC1 et famED1 plutôt qu'à famES1. À l'inverse, la version 2 aura tendance à en attribuer la responsabilité à famES1.

Une manière d'éviter ce biais est de noter le silence inter-tour dans une ligne spécifique, en faisant par là une responsabilité partagée par les locuteurs (Tannen 1981, Edwards 1993) :

Version 3

famHC1 ben oui euh oui au |- repas oui <famED1> le
soir -|
(silence)
famES1 tiens et ta maman elle travaille toujours à la /
boucherie là |- en bas ?

Dans la version 3, on a donc aussi des paragraphes dans lesquels les informations ne sont plus associées aux paroles d'un locuteur déterminé (voir aussi Gumperz et Berenz 1993).

Dans ce cas-ci également, le changement de disposition visuelle de l'information induit des interprétations différentes chez le lecteur de la transcription.

8 LE CORPUS COMME DONNÉES PARTAGÉES

On l'a dit, la transcription est une étape obligée pour étudier les productions langagières. Mais c'est une étape couteuse en temps (« time-consuming » : Chafe 1995 : 61, Leech *et al.* 1995 : 5)⁶³. En effet, il est

⁶³ Couteuse en temps, et donc, évidemment couteuse au sens propre (time is money). Les responsables du *Corpus gesproken Nederlands* estiment le cout à 0,50 euro par mot transcrit (communication de Dominique Willems).

nécessaire d'écouter plusieurs fois la bande sonore. Fillol et Mouchon (1977, cités par Blanche-Benveniste et Jeanjean 1987 : 101) estiment que cette écoute doit se faire à cinq ou six reprises. Au centre de recherche VALIBEL, où toutes les transcriptions sont revues par un expert autre que le transcripateur, on estime qu'à une minute de signal correspond une heure de travail de transcription.

Cette relecture, qui s'accompagne d'une réécoute du signal sonore, apparaît comme une étape indispensable dans le processus. Elle permet une contre-vérification des données. Ce second transcripateur a une importance capitale, puisque c'est lui qui en définitive établit le texte, au sens philologique. Il corrige les erreurs du premier transcripateur ; il propose, le cas échéant, des multitranscriptions. Le résultat de son écoute prend alors place dans la transcription à côté de la solution proposée par le premier transcripateur.

La transcription est donc une étape qui demande de l'énergie, de l'attention, du temps. Il est de ce fait souhaitable qu'une fois le corpus constitué, il puisse être réexploité pour d'autres recherches. Les données deviennent alors une ressource commune.

C'est dans cette optique qu'il est essentiel d'obtenir le consentement éclairé⁶⁴ des locuteurs, afin d'exploiter dans d'autres recherches des corpus constitués pour une analyse particulière, à l'occasion de laquelle les enregistrements ont été faits.

C'est également en vue d'une réutilisation des données que le corpus doit être documenté. Il s'accompagne de méta-données qui concernent les locuteurs (données de type sociologique : âge, études, profession, lieu de naissance, etc.), la situation d'enregistrement (quand, où), la situation de parole, la transcription (qui, quand), les liens entre l'enquêteur, le transcripateur et les locuteurs, etc. Ces méta-données sont stockées dans un document séparé, ou figurent dans une section qui précède la transcription (Johansson 1995).

Quant à la transcription elle-même, de nombreux chercheurs (Edwards 1993 et 1995, Du Bois 1991, Du Bois et al. 1993b, Ehlich 1993) mettent en avant des points essentiels qui en permettent la diffusion. Ils insistent sur l'ergonomie du système (sa lisibilité notamment), le choix des catégories, la documentation détaillée des conventions adoptées, etc. Ainsi, les cinq maximes qu'énonce Du Bois dans son article « Transcription design principles for spoken discourse research » (1991), à suivre dans l'établissement des conventions de transcription :

⁶⁴ Sur ce concept de consentement éclairé, voir le *Guide des bonnes pratiques* (2005).

-
- Category definition : Define good categories.
 - Accessibility : Make the system accessible.
 - Robustness : Make representations robust.
 - Economy : Make representations economical.
 - Adaptability : Make the system adaptable.

Nous voudrions insister sur le fait qu'actuellement, les conventions de transcription adoptées se doivent d'être compatibles avec une utilisation informatique des données, notamment avec des logiciels d'interrogation de corpus tels que les concordanciers.

On évitera donc, par exemple, les soulignements, l'italique, l'utilisation de polices particulières, la présence d'informations en notes de bas de page, etc.

Ainsi notamment, il est très courant d'utiliser le soulignement pour noter les chevauchements de paroles (cf. conventions du GARS en § 7.2). Si cette notation apparaît comme très lisible, l'information est perdue dès que le texte est sauvegardé dans un autre format que le format classique « .doc ». Il en va de même lorsque les chevauchements sont notés par un alignement des paroles chevauchées l'une sous l'autre.

9 LES CONVENTIONS DE TRANSCRIPTION DU CENTRE DE RECHERCHE VALIBEL

Depuis sa création en 1988, le centre de recherche VALIBEL a réfléchi à la pratique de transcription, se dotant de conventions explicites (voir Francard et Péronnet (1989), pour une première formulation). Celles-ci ont suivi quatre grands principes : adoption de l'orthographe standard, non-recours à la ponctuation de l'écrit, valorisation de l'oralité des corpus et compatibilité avec un traitement informatisé. Nous entendons par « valorisation de l'oralité » le fait que le texte transcrit ne gomme pas le travail d'élaboration du discours non planifié. Ces marques du travail de formulation (Morel et Danon-Boileau 1998), appelées également *disfluences* (cf. chap. 3), sont scrupuleusement notées, notamment :

- les pauses pleines et les ponctuations (*euh, ben, bon, etc.*) ;
- les répétitions de mots ou de suites de mots ;
- les amorces de morphèmes ;
- les autocorrections.

Ces quatre grands principes ont guidé les pratiques de transcription du Centre VALIBEL même si, au cours des années, ces pratiques ont évolué parallèlement au développement de nouveaux outils (Dister et Simon 2007)

ou en fonction des caractéristiques des données recueillies (conversations bilingues ou multi-locuteurs, etc.). Ces conventions de transcription sont appliquées à l'ensemble des textes qui entrent dans la banque de données⁶⁵. Elles sont explicitées dans un manuel à destination des transpositeurs (Dister *et al.*, 2006). Nous les résumons brièvement dans le tableau ci-après :

⁶⁵ Nos conventions s'appliquent au format des données telles que nous les stockons. En ce qui concerne la publication scientifique par exemple, on peut modifier certaines conventions ou la mise en forme du texte pour des questions de meilleure lisibilité ou de mise en évidence du phénomène étudié. Cette remarque se fonde sur le principe que le format de stockage et d'exploitation (format textgrid ou texte) ne coïncide pas nécessairement avec le format de publication (dont on peut obtenir automatiquement divers types, à l'aide par exemple du logiciel Transformer, cf. Ehmer 2006).

marque	signifié
/	pause brève ⁶⁶
//	pause longue
(silence)	silence
(toux), (rire), (chuchotement)	commentaires paraverbaux de type : toux, rire, chuchotement, etc.
(x)	passage incompréhensible d'une syllabe
(xx)	passage incompréhensible de plusieurs syllabes
(xxx)	passage incompréhensible relativement long
cou/	amorce de morphème
cou/ -pure	amorce achevée sans reprise antérieure
?	question de forme déclarative à contour intonatif montant
-	début de chevauchement
-	fin de chevauchement
--	début de chevauchement impliquant plus de 2 locuteurs
--	fin de chevauchement impliquant plus de 2 locuteurs
§	début de conversations en parallèle (séquences simultanées)
§	fin de conversations en parallèle (séquences simultanées)
§	sépare 2 conversations en parallèle
{choix1 , choix2}	multitranscriptions
{incertain}	transcription incertaine
[sampa]	indications de phonétique en Sampa

Tableau 1. Conventions de transcription VALIBEL

⁶⁶ Les 3 degrés de pauses sont jugés intuitivement par le transcripateur, c'est-à-dire non corrélés à des seuils de mesures acoustiques.

10 CONCLUSIONS

Beaucoup d'auteurs, que nous avons cités, ont mis l'accent sur ce fait, reprenant le titre de l'article pionnier de Ochs (1979) : « transcription as theory ». Parce qu'elle engage le chercheur, révèle ses conceptions de la langue en général et de l'oral en particulier, parce qu'elle a des répercussions sur l'analyse même qui sera faite des données, parce qu'elle est en elle-même déjà une démarche analytique et interprétative, la transcription est une étape cruciale de la recherche sur le parlé.

Dans un travail comme le nôtre, qui se base exclusivement sur les transcriptions, sans recourir au signal sonore, en théoriser les enjeux nous semblait donc capital. Les choix de conventions de transcription, qui peuvent paraître relativement neutres, ont nécessairement des effets sur le lecteur et sur l'interprétation qu'il fera des données. Ainsi, la représentation spatio-visuelle donnée aux productions orales peut avoir des incidences sur l'analyse, même quand le choix du découpage en tours de parole semble être objectivé auditivement.

Le choix de l'orthographe est aussi fondamental : de nombreuses transcriptions, qui adoptent pourtant explicitement l'orthographe standard, ont néanmoins recours, pour certains faits phonétiques, à une orthographe aménagée. Mieux vaut sans doute, comme le proposent de nombreux chercheurs, avoir différents degrés de transcription, dans lesquels figurent des informations différentes. Mais une transcription selon l'orthographe traditionnelle, sans aménagement, est dans l'incapacité de traduire de nombreux faits présents dans l'oral, et peut-être pertinents pour l'analyse. Savoir ce qui est « perdu », et ce qui reste, est aussi essentiel pour le chercheur.

On a aussi insisté ici sur un cas particulier, l'amorce de morphème, dont la fréquence d'apparition à l'oral est très élevée. Marquée de façon non ambiguë dans les transcriptions, son interprétation par le transcripteur relève dans de nombreux cas d'une analyse contextuelle de l'énoncé. Choisir l'une ou l'autre transcription pour les cas indécidables, quand la multitranscription surcharge la lecture, a inévitablement des incidences sur l'interprétation du texte transcrit. Ce simple fait, par sa fréquence d'apparition, et parce qu'il touche une zone sensible de l'énoncé, un endroit où le locuteur achoppe, a des répercussions non négligeables dans l'analyse de la planification de l'énonciation, par exemple, mais aussi en termes d'analyse morphosyntaxique, de statistique lexicale, etc.

L'établissement de conventions de transcription demandent donc une réflexion préalable, dans laquelle les enjeux seront théorisés. En effet,

Our data may have a future if we give them the attention
they deserve. (Ochs 1979 : 72)

CHAPITRE 2

LE CORPUS

Un corpus est d'abord et avant tout représentatif de lui-même. (Maurice Gross, communication personnelle)

1 VOUS AVEZ DIT *CORPUS* ?

Le terme *corpus* est utilisé dans la littérature avec des sens parfois très éloignés les uns des autres, bien qu'il soit généralement fait référence à du *texte*¹. Ainsi, alors que la linguistique de corpus est l'objet depuis quelques années d'un intérêt croissant en francophonie², suivant en cela le mouvement initié dans le monde anglo-saxon, les linguistes travaillant sur corpus n'ont parfois que peu de terrain d'entente, non seulement dans les objectifs qu'ils poursuivent, mais également dans les données qu'ils rassemblent dans des *corpus*³.

Par ailleurs, les linguistes ne sont pas les seuls à travailler sur des corpus : la sociopolitique, l'ethnométhodologie, les études littéraires ou encore la lexicométrie se fondent également sur des corpus. Ceux-ci peuvent être constitués de texte suivi, de phrases isolées (Bestgen *et al.* 2004) voire de syntagmes (Ploux 2005) ou de lexèmes (Dister et Moreau 2006). Il peut également s'agir de données mixtes, comme chez Kastberg Sjöblom (2006) dont le corpus est le dictionnaire franco-suédois *Norstedts stora svensk-franska ordbok*. Le corpus peut encore être monolingue ou multilingue et, dans ce dernier cas, aligné ou non. Il est documenté ou non, enrichi d'annotations (voir chap. 6) ou livré dans sa forme « brute ».

Si l'on n'a pas attendu l'apparition des ordinateurs pour constituer des corpus (voir Kennedy 1998 : 13-19), les analyser et, notamment, se livrer à

¹ Notons que ce n'est pas toujours le cas. Ainsi, un corpus en langue des signes n'est pas constitué d'un matériau texte mais bien d'un ensemble de gestes.

² Pour preuve, notamment, les rencontres organisées chaque année à Lorient sous le titre « Linguistique de corpus », les *JADT (Journées d'analyse des données textuelles)*, qui rassemblent linguistes, mais aussi statisticiens ou analystes du discours, certains numéros thématiques de revues ou encore la naissance de revues spécialisées (cf. la revue *Corpus* à l'Université de Nice).

³ Ces divergences se traduisent dans le pluriel adopté par Habert *et al.* (1997) pour le titre de leur livre : *Les linguistiques de corpus*. « La convergence apparente des intérêts ne doit pas masquer les divergences théoriques et pratiques. » (Habert 2000 : 17)

des calculs statistiques sur ceux-ci (la tradition des concordanciers chez les latinistes)⁴, à l'heure actuelle, corpus et traitement informatique semblent étroitement liés. En effet, le développement des outils informatiques a permis d'accroître considérablement les capacités de stockage et le volume des données conservées. Des logiciels se sont développés, qui permettent non seulement de stocker mais aussi de gérer facilement de très gros *corpus électroniques (machine-readable)* : indexation, comptage ou encore présentation d'occurrences en contexte sont des tâches relativement bien maîtrisées par les outils actuels. De plus, la numérisation et les facilités d'échanges permettent actuellement de constituer et de diffuser d'immenses corpus, à un coût très faible. Multitrans, une société de traduction, vante ainsi sur son site web les mérites de son gestionnaire de corpus :

Ainsi, un corpus contenant plusieurs millions de mots peut être créé en l'espace de moins d'une heure. (Gestionnaire de corpus MultiTrans, http://www.multicorpora.ca/corpusbuilderandsearch_f.html, consulté le 11 janvier 2006)

Ainsi, alors qu'il y a encore peu de temps, un corpus d'un million de mots constituait une ressource qui semblait énorme, il est actuellement banal de travailler sur plusieurs dizaines, voire centaines de millions de mots⁵.

2 CORPUS AU SENS LARGE

Le corpus est donc une réalité mouvante, fluctuante. Pour certains, dans une acception large, *corpus* équivaut simplement à *données textuelles*. Un corpus est alors envisagé comme un ensemble de textes, d'une taille généralement considérable. À cet égard, le web peut apparaître comme un corpus idéal, qui se renouvelle chaque jour et dont la quantité phénoménale de textes n'a d'égal que la variété de ses scripteurs : sites gouvernementaux, textes de revues scientifiques de tous domaines, sites de recettes de cuisine, pages personnelles (les fameux blogs), etc. La toile est ainsi vue comme un vaste réservoir de la langue contemporaine, en dépit de la qualité des données, souvent remise en cause. On n'en finirait pas de citer les recherches actuelles, aussi diverses dans leurs objectifs que dans leurs méthodes, qui portent sur le web comme corpus. D'une part, on utilise « le web comme corpus » (Grefenstette (1999) pour la traduction de noms composés ; Gala

⁴ Dans le même ordre d'idées, on peut citer le recensement de Gougenheim et de ses collègues dans l'entreprise à destination de l'enseignement du français dans *Le français Fondamental*. De même, l'école de Prague s'est livrée à des études quantitatives sur des corpus (voir Kennedy 1998 : 10).

⁵ Nous ne parlons pas ici de corpus transcrits de l'oral, dont on a vu au chapitre 1 la difficulté de constitution.

(2003) pour la désambiguïsation du rattachement prépositionnel ; etc.). D'autre part, des outils sont développés pour affiner ces recherches (outils de « web crawling » ou « butinage du web ») : Webaffix (Tanguy et Hathout 2002) est un outil d'acquisition morphologique dérivationnelle à partir du web ; etc. Des rencontres scientifiques sont organisées, sur le thème du web comme ressource linguistique⁶ (Kilgariff et Grefenstette 2003).

Outre la masse impressionnante de données qu'il représente, c'est principalement le critère de la *disponibilité* de celles-ci qui fait du web un corpus privilégié. En effet, les données sont non seulement volumineuses, mais également gratuites⁷ et très facilement accessibles. Dans le même ordre d'idées, c'est cette disponibilité qui surdétermine le choix de journaux comme corpus dans un grand nombre de recherches. Par exemple, la diffusion sur CD-Rom des archives annuelles des journaux *Le Monde* (France) ou *Le Soir* (Belgique francophone) – pour un prix dérisoire dans ce dernier cas – en fait des ressources textuelles de prédilection pour les chercheurs qui travaillent sur du texte. C'est ici la disponibilité des données et non leur caractère particulier (textes de presse, qui seraient choisis pour eux-mêmes en tant que pratique textuelle particulière) qui constitue le critère prépondérant de sélection du corpus, renforçant la prédominance que l'on connaissait déjà de la langue normée écrite comme objet d'étude⁸.

Pour les utilisateurs des corpus ainsi définis, seul compte donc le volume des données, rassemblées de manière opportuniste (Kennedy 1998 : 61) : peu importe le type de productions langagières que celles-ci caractérisent.

3 CORPUS AU SENS RESTREINT

À côté de cette définition très générale du corpus, on trouve des définitions beaucoup plus strictes, dans lesquelles la notion a été théorisée et où le corpus est plus qu'un simple assemblage de textes. John Sinclair, l'un des pères de la linguistique de corpus anglo-saxonne, en donne une définition particulièrement restrictive :

A corpus is a collection of naturally-occurring language text, chosen to characterize a state or variety of a language. (Sinclair 1997 : 171)

⁶ Journée ATALA du 11 mars 2006 *Le Web comme ressource pour le TAL ; Web as Corpus Workshop*, deuxième édition, lors de la conférence EACL de juin 2006 ; etc.

⁷ Ce qui ne règle pas toujours la question des droits d'accès.

⁸ Avec les dérapages que cela peut provoquer chez certains où le français = le français du *Monde*.

A collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language. (Sinclair 1996)

Cette définition restrictive du corpus est à la base, notamment, de la constitution de corpus représentatifs (cf. § 4.4).

4 DIFFÉRENTS TYPES DE CORPUS

En lien avec une définition plus ou moins restrictive donnée au corpus, on peut dégager plusieurs grands types de corpus.

4.1 Le corpus clos

Appelé aussi corpus *fermé* ou *statique*, le corpus clos correspond au « sample corpus » tel que le définit Sinclair (1997: 171) :

(...) a finite collections of texts, often chosen with great care and studied in detail. Once a sample corpus is established, it is not added to or changed in any way.

Le corpus clos, constitué une fois pour toutes, est un corpus immuable, de taille finie, déterminée à l'avance lors de sa constitution.

L'une des raisons pour lesquelles un corpus est clos est qu'il constitue par la force des choses un *corpus complet* (*full-text corpus* de Kennedy (1998)). Parmi les corpus clos, on trouve notamment les corpus constitués de la totalité des œuvres d'un auteur. Si celui-ci est mort, sauf cas rarissimes où l'on retrouverait des textes inédits, le corpus des œuvres est définitivement fermé. Le corpus complet vise à l'*exhaustivité* qui, dans le cas des œuvres d'un auteur, peut être atteinte.

4.2 Le corpus ouvert

À l'opposé du corpus clos, le corpus ouvert n'est pas figé, constitué une fois pour toutes, mais il accueille, au fil du temps, de nouveaux textes. Ceux-ci peuvent s'ajouter de façon non prédictible. C'est bien entendu le cas du web. Mais c'est aussi le cas, par exemple, du *corpus de la littérature belge*, qui, s'il vise à l'exhaustivité, fait référence à l'ensemble des œuvres produites par les écrivains belges, et s'enrichit à chaque nouvelle publication d'un auteur belge. Dans le cas du web, l'enrichissement se fait au hasard, dans celui de la littérature belge francophone, pas. Une définition préalable de ce que l'on entend par *littérature belge* est nécessaire, qui permet de la distinguer de la littérature française, de même qu'une caractérisation de l'auteur belge s'avèrera indispensable (né en Belgique ? et/ou qui vit en

Belgique ? et/ou qui publie en Belgique ? Etc.). Dans ce cas concret, ce qui est défini, ce sont les critères d'entrée de l'item dans le corpus : qu'est-ce qui définit une œuvre belge ? Dès lors que l'on dispose des critères théoriques, la définition du corpus est satisfaisante même si celui-ci n'est pas décrit exhaustivement ou concrètement rassemblé (on a une définition en compréhension et non en extension). On possède les coordonnées qui permettent potentiellement de le constituer. Non prédictible ne signifie pas ici au hasard.

En linguistique, un corpus peut être ouvert afin de permettre au chercheur d'ajouter des données. Par exemple, afin de rééquilibrer le volume des données déjà présentes dans le corpus selon un paramètre particulier.

4.3 Le corpus dynamique

Le corpus dynamique est une sous-catégorie de corpus ouvert. On doit cette notion à Antoinette Renouf (1994) où le corpus est vu comme un flux de données textuelles. La chercheuse a ainsi analysé des apparitions ou des disparitions de mots, des évolutions de sens, de nouvelles combinaisons de mots, etc. dans le *Times*, de 1989 à 1993 (projet AVIATOR). Utiliser le web comme un corpus dynamique est également l'idée centrale du logiciel WebCorp⁹ (Renouf et Kehoe 2002) de l'Université de Birmingham.

C'est la même idée qui a donné naissance à un corpus ouvert, qualifié de *monitor corpus* (*corpus de suivi*), à l'équipe de Sinclair à l'Université de Birmingham, autour de l'entreprise lexicographique du Cobuild :

which re-uses language text that has been prepared in machine-readable form for others purposes – for typesetters of newspapers, magazines, books, and, increasingly, word-sprocessors ; and the spoken language mainly for legal or bureaucratic reasons. (Sinclair 1996 : 172)

Le corpus devient de plus en plus gros et fonctionne comme un corpus de veille, un observatoire des changements linguistiques.

Si la qualité des données ne doit pas être remise en cause par cet enrichissement croissant, en termes quantitatifs, l'échantillonnage est beaucoup plus malaisé et moins rigoureux que dans un corpus fermé (McEnery et Wilson 2001 [1996] : 32). À cet égard, le *monitor corpus* est plus une collection de textes qu'un véritable corpus au sens de la définition que nous reprenions de Sinclair (4.1).

⁹ Voir <http://www.webcorp.org.uk/>.

S'il est également non statique, et qu'il s'incrémente de nouvelles données textuelles, le corpus dynamique peut aussi offrir la particularité que certaines données disparaissent. C'est le principe du logiciel *GlossaNet*¹⁰, développé par Cédric Fairon à l'Université catholique de Louvain. En effet, *GlossaNet* analyse quotidiennement les sites en ligne de plus de 100 quotidiens, mais avec la particularité que seule la dernière édition des journaux sert de corpus (Fairon et Courtois 1998, Dister et Fairon 2004). Celui-ci se renouvelle donc chaque jour, alors que la requête des utilisateurs de Glossanet n'est introduite qu'une seule fois dans le système. L'idée sous-jacente pourrait être : « On ne se baigne jamais deux fois dans le même fleuve. »

4.4 Le corpus de référence

Sinclair (1996) définit ainsi le corpus de référence :

A reference corpus is one that is designed to provide comprehensive information about a language. It aims to be large enough to represent all the relevant varieties of the language, and the characteristic vocabulary, so that it can be used as a basis for reliable grammars, dictionaries, thesauri and other language reference materials. The model for selection usually defines a number of parameters that provide for the inclusion of as many sociolinguistic variables as possible and prescribes the proportions of each text type that are selected. A large reference corpus may have a hierarchically ordered structure of components and subcorpora.

Si les corpus de référence sont une tradition bien établie dans le domaine anglo-saxon – le *BNC*, le *Brown* ou encore le *LOB* en sont des exemples parfaits (cf. chap. 6) –, le monde francophone, à notre connaissance, ne connaît qu'une seule tentative revendiquée comme telle de corpus de référence : le *Corpus de référence du français parlé* (cf. § 4.4.1)¹¹.

En effet, la constitution de gros corpus en francophonie se fait en général de manière opportuniste : afin de garantir un volume important de données,

¹⁰ <http://glossa.fltr.ucl.ac.be/>.

¹¹ On mentionnera aussi le projet Ciel-f, pour *Corpus International et Équilibré de la Langue Française*, qui à l'heure où nous écrivons ces lignes n'est pas encore clairement défini. L'idée revendiquée par ses concepteurs de l'Université de Freiburg serait de faire pour le domaine francophone un corpus comparable à l'*ICE Corpus* de l'anglais (cf. § 4.4.3 et § 4.4.5).

Voir http://www.hpcl.uni-freiburg.de/projekte/kailuweit/copy3_of_projekt.c.

on rassemble, dans un recyclage plus ou moins explicite, des textes qui existent par ailleurs¹².

Benoît Habert (1999 : 14-15) propose la fiche signalétique de ce que serait un corpus de référence du français contemporain (vu comme un corpus clé dans le sens où il « donnerait les clés du français actuel »)¹³. Il comprendrait 20 millions de mots étiquetés, dont 10 % relèveraient de productions orales. Les textes, échantillonnés, seraient postérieurs à 1980, afin de refléter un état de langue contemporain. Ceux-ci seraient rigoureusement documentés par des métadonnées et représenteraient six emplois : la presse nationale et régionale (4 millions de mots pour chacune) ; le français non hexagonal (5 millions de mots) ; le français non standard (2,5 millions de mots) ; l'oral spontané et non spontané (1 million de mots pour chacun) ; le français scientifique et technique (2,5 millions de mots). Cette entreprise d'un corpus clé pour le français a été lancée en 1999 à l'initiative de la DGLF (Direction Générale à la Langue Française). À notre connaissance, elle n'a pas abouti. Sans doute parce que

l'essentiel [des six] emplois ne figure pas dans les ressources existantes : la volonté de les représenter suppose une acquisition de données primaires non négligeable. (Habert 1999 : 14)

Le corpus de référence, parce qu'il est considéré comme *représentatif*, permet la comparaison entre lui-même et un autre corpus afin de mesurer des écarts. Il fonctionne ainsi comme un *corpus témoin*, dans le sens de « groupe témoin » en médecine par exemple. De même, parce qu'il a pour vocation de servir de base à de multiples recherches, la diversité des résultats obtenus résulte non pas d'une variation dans les données mais bien d'hypothèses de recherches ou de méthodologies différentes (McEnery et Wilson 2001 [1996] : 32).

Parce que chacun des textes qui le compose est documenté, il constitue un *réservoir à corpus idéal* :

Pour qu'un corpus se prête à de telles extractions et réorganisations, il importe que chacune des unités élémentaires qui le constituent soient « autonomisable » : on doit posséder suffisamment d'informations fines sur elle pour pouvoir l'extraire du corpus d'ensemble et l'assembler avec d'autres

¹² Voir à ce sujet Habert (2000 : 18-22) dont les titres des chapitres sont à cet égard évocateurs : *Misère de la philologie... et ...philologie de la misère*.

¹³ Il y a également derrière ce terme de *clé* l'aspect stratégique de l'entreprise, qui consiste à positionner le français par rapport aux autres langues, notamment l'anglais, dans l'établissement de données électroniques de référence.

éléments du même corpus ou d'autres corpus sans perdre ces renseignements qui sont indispensables pour interpréter les contrastes et les convergences manifestés dans le nouveau corpus rassemblé (Habert *et al.* 1998 : 37).

À cet égard, la banque de données VALIBEL constitue un réservoir à corpus idéal (cf. § 4.5).

4.4.1 Le Corpus de référence du français parlé

Le *Corpus de référence du français parlé (CRFP)* a été constitué à partir de 1998 par l'équipe Corpus de l'Université de Provence, dirigée par Claire Blanche-Benveniste, pour être ensuite achevé par l'équipe DELIC (Description Linguistique Informatisée sur Corpus) qui lui a succédé. Le CRFP se veut

(...) un témoignage de la langue française parlée aujourd'hui dans les principales villes de l'hexagone. Il s'agissait avant tout de recueillir des données représentatives d'un français parlé que nous pourrions qualifier d'« usage général et courant » (...) (DELIC 2004 : 11)

Ainsi que son nom l'indique et tel qu'il est explicitement présenté par ses auteurs, il a été conçu pour être représentatif du français parlé contemporain **dans les grandes villes de France**.

Le corpus contient 134 enregistrements, qui totalisent près de 37 heures de parole. Cela correspond approximativement à 440 000 mots. Seule une partie des enregistrements est transcrite (23 heures de parole). Trois situations de parole sont représentées, en proportion inégale : parole privée, parole professionnelle et parole publique. La répartition géographique privilégie la région du sud-est de la France.

Si l'article qui décrit le *CRFP* (DELIC 2004) fournit des données chiffrées sur les enregistrements qui composent le corpus, en termes de répartition géographique, de données sur les locuteurs, le type de parole, etc., il n'explicite en rien les principes d'échantillonnage qui ont prévalu à sa constitution¹⁴.

¹⁴ La note 2 de l'article (DELIC 2004) informe sans plus d'indication ni de référence à une publication : « Les critères d'échantillonnage adoptés pour la constitution de ce corpus ont été élaborés par Mireille Bilger. »

4.4.2 La représentativité

Sans pertinence pour la compilation d'un corpus selon l'acception large de « assemblage de textes », la question de la *représentativité des données* est centrale pour les concepteurs d'un corpus de référence tel que défini par Sinclair.

En fait, l'intérêt de la représentativité est que, si elle est atteinte, il est possible de généraliser les résultats obtenus sur le corpus à un ensemble plus large.

Le corpus est alors conçu comme un échantillon représentatif dans lequel se retrouvent toutes les caractéristiques présentes dans l'ensemble dont il est issu. La représentativité permet de lutter contre les généralisations abusives. Ces généralisations sont liées à deux types d'erreurs statistiques dénoncées par Biber (1993a) : « l'incertitude » (*random error*) et la « déformation » (*bias error*). On parle d'incertitude quand le corpus considéré comme échantillon est trop petit (cf. § 4.4.4) ; la déformation consiste à utiliser un échantillon dont les caractéristiques ne sont pas les mêmes que celles de la population que cet échantillon est censé représenter (nous parlions ci-dessus du français du *Monde* pour le français « tout court »).

Cette question de la représentativité du corpus est loin d'être triviale. La question principale qui se pose est de savoir de quoi le corpus est représentatif¹⁵. Quelle variété de langue le corpus représente-t-il ? Un corpus se veut-il représentatif de la langue dans son ensemble ? D'un sous-ensemble de celle-ci ? Comment se définit ce sous-ensemble ? Biber (1993a), dans un article consacré à la représentativité dans la conception des corpus, insiste sur la nécessité de définir aussi précisément que possible la population que l'on veut étudier, et qui sera représentée par un sous-ensemble. Cette étape est essentielle et doit se faire avant n'importe quelle autre afin de garantir la validité des résultats que l'on va obtenir. Il ne s'agit pas d'avoir une vague idée de ce que l'on va représenter ; la population doit avoir des contours aussi précis que possible.

4.4.3 L'échantillonnage

Si le corpus est un échantillon d'un ensemble plus large, il est en général lui-même composé d'échantillons. Ceux-ci sont soit des textes entiers, soit des parties de textes, en général de taille assez limitée. Cette dernière pratique est bien instituée dans le monde anglo-saxon : le *BNC* (cf. chap. 6) et l'*ICE Corpus* ont été constitués de la sorte.

¹⁵ Cf. l'article de Habert (2000) : « Des corpus représentatifs : de quoi, pour quoi, comment ? »

The International Corpus of English (ICE) est composé de dix-huit variétés nationales ou régionales, dont l'anglais est la première ou seconde langue officielle (Greenbaum 1992). Chaque variété représente un million de mots, échantillonnés en 500 textes de 2000 mots chacun. Les données sont strictement contemporaines (entre 1990 et 1994).

L'idée est que chaque corpus existe indépendamment, mais le fait de les regrouper ajoute une plus-value au corpus.

L'intérêt de procéder par échantillons est que ceux-ci sont de taille équivalente et permettent plus facilement des comparaisons entre les variétés visées.

4.4.4 La taille

La question de la *taille* du corpus (et des échantillons qui le composent) est étroitement liée à celle de la représentativité. Ainsi, plus le corpus est gros, plus il serait représentatif de la population qu'il vise à étudier. C'est ainsi que certains justifient l'utilisation du web, justification argumentée par la variété des données (même si celle-ci est souvent difficilement caractérisable, pour les productions plus informelles, en ce qui concerne notamment le scripteur).

Les partisans de cette optique des gros corpus partagent l'adage « more data is better data » (Pery-Woodley 1995).

Benoît Habert (2000) résume la conviction de ces chercheurs :

(...) l'élargissement mécanique des données mémorisables (les centaines de millions de mots actuelles deviendront à terme des milliards) produit inévitablement un échantillon de plus en plus représentatif de la langue traitée. Si l'on n'arrive pas à cerner précisément les caractéristiques de l'ensemble des productions langagières, il ne reste qu'à englober le maximum d'énoncés possibles. À terme, la nécessité de choisir finirait par s'estomper. (Habert 2000)

Mais selon certains chercheurs, la taille d'un corpus ne garantit sa représentativité (McEnery et Wilson 2001 [1996] : 78). Pour les traits fréquents, un petit corpus serait représentatif, alors que pour d'autres traits, la taille de l'échantillon doit être plus grande (Biber 2000).

À l'heure actuelle, les chercheurs n'ont pas abouti à un consensus en ce qui concerne la taille des échantillons (ou du corpus dans sa totalité). Biber (1988), pour son étude des genres, a estimé que des échantillons de 1000 mots étaient représentatifs et donc suffisants.



4.4.5 L'équilibre

Une autre question liée à la représentativité est celle de *l'équilibre* des différents échantillons entre eux. Un corpus représentatif est en effet un *corpus équilibré (balanced)*. Ainsi, si tous les genres¹⁶ qui composent la population générale sont représentés, ils doivent l'être non pas en quantité égale mais de manière proportionnelle par rapport à cette population. Ainsi, par exemple, si la littérature du XVI^e siècle se compose plus de poésie que de roman, une proportion équivalente de ces deux genres devra figurer dans le corpus représentatif de la littérature du XVI^e siècle.

En fait, la question de l'équilibre est un problème épineux, particulièrement en ce qui concerne la langue considérée globalement :

No one knows what proportion of the words produced in a language on any given day are spoken or written. Individually, speech makes up a greater proportion than does writing of the language most of us receive or produce on a typical day. However, a written text (say, in a newspaper article) may be read by 10 million people, whereas a spoken dialogue involving the purchase of a pair of shoes may never be heard by any person other than the two original interlocutors. Similarly, a broadcast conversation on radio or television will reach many more ears than a commercial encounter involving just a costumer and a salesperson. (Kennedy 1998: 63)

Comme le note Kennedy, la question de l'équilibre du corpus est liée à celle du point de vue. En effet, outre les types de textes (le « genre » exploré par Biber), on peut aborder les textes en fonction de leurs conditions de réception et/ou de production. Il s'agit là de « trois visions distinctes de la représentativité » (Habert 2000 : 23).

Dans le corpus *ICE*, qui regroupe des variétés d'anglais, depuis l'anglais des États-Unis jusqu'à celui des îles Fidji en passant par le Ghana, cette question de la répartition a été délicate à régler. En effet, s'il s'imposait pour certaines de ces variétés la prise en compte de formes de communication récentes comme le courrier électronique par exemple, cela n'avait aucune pertinence pour d'autres variétés (Nelson 1996a : 28).

¹⁶ La question des genres est une question épineuse. Abordée par Biber (1988) sur base de critères internes, elle a abouti chez cet auteur à une classification en 23 genres. Dans ce travail, nous ne l'avons guère abordée, étant donné les données disponibles dans la banque de données VALIBEL (cf. § 4.6).

Au final, les concepteurs ont donc réparti chaque sous-corpus en 300 textes oraux et 200 textes écrits. Nous donnons ci-dessous la représentation schématique des catégories de textes oraux, reprise de Nelson (1996a : 29).

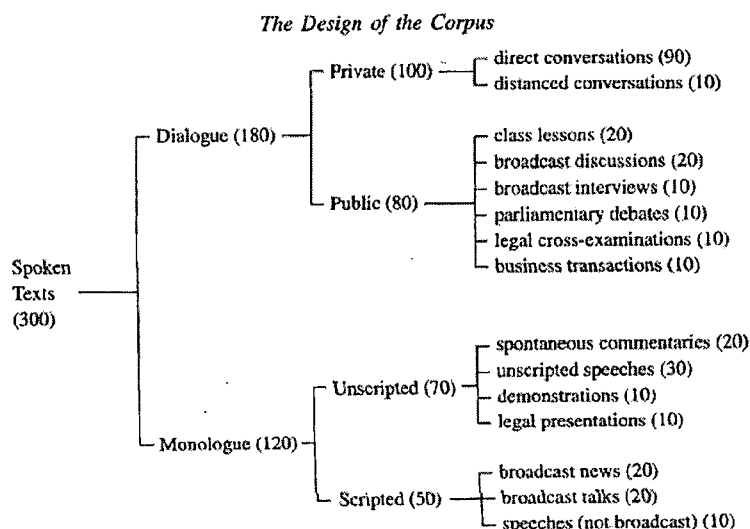


Figure 1. Représentation schématique des catégories de textes oraux

On le voit dans le tableau ci-dessus, l'équilibre est une question d'appréciation. Comme le dit Kennedy (1998 : 62) :

The notions of representativeness and balance are, of course, in the final analysis, matters of judgment and can only be approximate.

4.5 La banque de données VALIBEL

La banque de données VALIBEL n'est pas conçue comme un méga-corpus clos mais, ainsi que son nom l'indique, comme une base de données textuelles. Il s'agit donc d'un corpus ouvert, tel que nous venons de le définir. Au fil des recherches menées par les membres de l'équipe ou par des étudiants dans le cadre de leur mémoire s'ajoutent des enregistrements de productions orales, ainsi que les transcriptions de ces sources sonores.

Dans la banque de données VALIBEL, le mot *corpus* est entendu comme « ensemble de textes (oraux et écrits) collectés dans le cadre d'une étude particulière ». La banque de données se compose actuellement de 24 corpus¹⁷, définis par une thématique de recherche particulière. Chacun de ces

¹⁷ Pour une radioscopie complète de la banque de données, voir le site de VALIBEL à l'adresse suivante : <http://valibel.fltr.ucl.ac.be/>.

corpus se compose lui-même de différents enregistrements, plus ou moins nombreux selon les cas (un seul dans le corpus *SOU*, 49 dans le corpus *ACC*).

Malgré la succession des chercheurs au sein de l'équipe, une cohérence et une systématique exemplaires ont prévalu lors de l'ajout de nouvelles données :

- tous les enregistrements ont été conservés (et sont maintenant numérisés), afin de permettre des études de type phonétique/phonologique, études rendues impossibles par le seul recours à une transcription en orthographe standard ;
- chaque corpus est documenté¹⁸ : objectifs de la recherche, thématique, nom des chercheurs, date de constitution, nombre d'enregistrements qui composent le corpus, nombre de locuteurs, etc.
- chaque enregistrement est accompagné d'une fiche signalétique : date, lieu, qualité sonore, durée, cadre général de l'enregistrement (disposition des participants dans la pièce, état d'esprit de ceux-ci si cela est pertinent, etc.), responsable de l'enregistrement, type d'interaction (entretien sollicité, cours, etc.), nombre de locuteurs, relation entre eux, consentement des participants pour des exploitations ultérieures, etc. Les informations concernant la transcription sont aussi disponibles : nom du transcripateur, du réviseur, dates de ces deux étapes, nombre de mots, alignement ou non au son, etc. ;
- une fiche d'identification est attachée à chaque locuteur : outre les données telles que l'âge, le sexe, le niveau d'études, la localisation géographique, l'état civil, la profession ou encore la mobilité géographique (lieux de résidence successifs), des informations linguistiques sur les locuteurs sont précisées (langues pratiquées ou connues, et ce également chez les parents) ;
- des conventions explicites, établies dès le départ, ont été appliquées à l'ensemble des transcriptions.

Si le centre VALIBEL fait figure d'observatoire du français en Belgique, la banque de données qu'il gère ne peut se prévaloir de constituer un corpus représentatif du français parlé en Belgique. En effet, aucune idée de *représentativité*, telle que nous l'avons définie plus haut, n'a guidé la sélection des enregistrements présents dans la banque de données. Ainsi

¹⁸ Voir les propositions de la *TEI* (*Text Encoding Initiative*, TEI 2002) pour les en-têtes et cartouches destinés à documenter un corpus et ses composants <http://www.tei-c.org.uk/>. Voir aussi Habert *et al.* (1998).

donc, la plupart des productions orales relèvent du « genre » de l'entrevue ; elles ont en effet été enregistrées dans le cadre d'enquêtes sociolinguistiques menées avec des questionnaires semi-dirigés (cf. Francard *et al.* 1993). Les situations de parole publique et de parole professionnelle sont quasiment absentes de la banque de données¹⁹, les enregistrements relevant de conversations spontanées sont beaucoup plus rares mais aussi plus récents et il n'y a pas d'enregistrement d'oral planifié (dans le sens de oral préalablement écrit et ensuite lu, ou « récité » : conférence, pièce de théâtre, récitation, etc.).

4.6 Constitution de notre corpus

D'emblée, l'idée d'étiqueter la totalité de la banque de données textuelles orales de VALIBEL a été écartée. En effet, il nous semble essentiel, notamment dans la perspective de la diffusion du corpus auprès d'autres chercheurs, que l'annotation ne comporte pas d'erreurs et, pour ce faire, qu'elle soit validée par un expert humain. Nous avons donc fait un choix parmi cette masse considérable de données, afin de n'en conserver qu'une partie qu'il est humainement envisageable de relire et de corriger manuellement.

C'est en ce sens de choix, si l'on suit Habert, que nous pouvons parler pour notre assemblage de textes d'un corpus :

C'est l'opération de choix raisonné parmi les composants disponibles qui crée un corpus. (Habert 2000 : 13)

Deux options s'offraient théoriquement à nous pour la composition du corpus. La première correspond à un échantillonnage du corpus, à la manière dont le *Brown Corpus* a été conçu pour l'anglais américain (cf. chap. 6, § 3.1) : on choisit un nombre important de courts extraits. Cette méthode d'échantillonnage nous semble revêtir un inconvénient majeur. Les textes ainsi « saucissonnés » ne sont plus exploitables, par exemple, dans le cadre de l'analyse conversationnelle, qui est l'un des domaines de recherche à VALIBEL.

L'étiquetage morphosyntaxique n'étant pas une fin en soi mais conçu comme une base fiable pour des analyses ultérieures, il nous a donc semblé essentiel de conserver les textes dans leur intégralité, afin que ceux-ci puissent être utilisés dans des exploitations futures aussi variées que

¹⁹ Notons néanmoins qu'à côté des enregistrements et des transcriptions faits par les membres de l'équipe, VALIBEL archive les transcriptions de l'émission radiophonique *Matin Première* disponibles sur le site de la RTBF. En ce sens, il existe des transcriptions de parole publique planifiée, mais selon les conventions de l'orthographe standard.

possible. Nous avons donc opté pour la solution qui consiste à ne sélectionner que certains textes parmi tous ceux qui composent la banque de données VALIBEL, mais à garder ces textes complets.

Nous présentons maintenant les critères qui ont prévalu lors de la sélection des textes qui composent notre corpus. Nous donnons également des données chiffrées qui concernent la répartition des différents critères choisis sur l'ensemble du corpus.

Si, comme nous venons de le voir, les enregistrements qui composent la banque de données VALIBEL ne sont pas représentatifs des productions langagières que l'on peut trouver en Belgique francophone, nous avons néanmoins tenté d'effectuer une sélection raisonnée des textes afin de constituer notre corpus. Celui-ci n'a pas la prétention d'être représentatif, loin s'en faut, mais nous souhaitons obtenir une variété la plus large possible compte tenu des ressources dont nous disposions. Nous verrons que certaines catégories sont sur-représentées, mais cela ne doit pas occulter la grande variété des données ici représentées.

Nous allons détailler les textes qui composent notre corpus – ainsi que les locuteurs – car, comme le notent Habert *et al.* (1998 : 156) :

Sans une documentation jointe, un corpus est mort-né.
L'un des dangers de la facilité actuelle à rassembler des
textes électroniques est que précisément les objectifs du
regroupement ainsi que ceux des annotations effectuées
ne soient pas enregistrés : le corpus cesse d'être
utilisable dès que se perd la mémoire de ces choix.

Les critères de sélection sont de quatre ordres : une répartition raisonnée des textes parmi les sous-corpus constituant la banque de données VALIBEL, la qualité sonore des enregistrements, la situation de parole ainsi que des paramètres liés aux locuteurs. Nous détaillons maintenant ces choix.

4.6.1 Sous-corpus et enregistrements

Ayant décidé de ne sélectionner que des enregistrements entiers, il fallait dès lors en limiter le nombre. Nous avons donc d'abord procédé à une sélection des sous-corpus, avant de choisir parmi ces sous-corpus certains textes, en nombre relativement égal. Certains sous-corpus ont été écartés, pour des raisons diverses, liées notamment aux critères que nous définissons ci-dessous. Parmi les textes retenus pour chaque sous-corpus, nous avons tenté de varier autant que faire se peut les critères concernant les locuteurs (cf. § 4.6.4).

4.6.2 Qualité sonore des enregistrements

Le premier élément de sélection a permis d'éliminer d'emblée certains enregistrements : il s'agit de la qualité sonore de ceux-ci²⁰. En effet, les recherches les plus récentes menées au centre VALIBEL ont pris une orientation différente (tout en étant la continuité des recherches antérieures) en se centrant véritablement sur le matériau sonore (Simon 2004, Hambye 2005). Prosodie, phonétique et phonologie sont ainsi des axes de recherche privilégiés actuellement. Il nous semblait donc important, dans notre travail sur un corpus destiné à l'ensemble des membres de l'équipe, de sélectionner des données qui seraient exploitables également au niveau du signal sonore. En effet, notre souhait est d'obtenir un texte non seulement enrichi par l'étiquetage morphosyntaxique mais également aligné au son. Cette double approche, qui donne accès au corpus par deux entrées possibles, la morphosyntaxe et le signal, est à notre connaissance inédite²¹.

4.6.3 Situation de parole

Comme nous l'avons dit dans la présentation de la banque de données VALIBEL, les types de parole (parole privée, publique, professionnelle, récit, argumentation, etc.) et les types d'interaction (monologue, dialogue, entretien, etc.) sont très peu variés. Nous avons sélectionné deux situations de communication, qui relèvent toutes deux de la parole privée : l'entretien semi-directif, représenté par 51 entrevues réalisées dans le cadre d'enquêtes sociolinguistiques, et la conversation informelle entre amis ou membres d'une même famille, dans 9 enregistrements²².

Si la situation de l'entrevue est plus formelle (les locuteurs sont d'ailleurs appelés d'une part *intervieweur* et de d'autre part *informateur*), les caractéristiques de l'oral en termes de non-planification de la parole sont très

²⁰ Une étape indispensable du travail a donc été la numérisation des données et la vérification de leur qualité sonore par l'équipe de VALIBEL. Les enregistrements, qui étaient tous conservés sur mini-disc, sont maintenant disponibles sur support informatique.

²¹ Un outil qui permette la gestion en parallèle de texte et de son alignés a également été choisi : ProsoDB, développé à l'Université de Freiburg par Peter Gilles. Voir <http://fips.igl.uni-freiburg.de/prosodb-test/>. ProsoDB a fait l'objet d'un développement par le CENTAL et le Centre de recherche VALIBEL, afin d'adjoindre aux données sonores et textuelles les informations sur les locuteurs, les situations d'enregistrements, etc. Le système complet porte maintenant le nom de [moca] (Kervers *et al.* 2006).

²² Nous avons en fait pris tous les enregistrements de conversations informelles présents dans la banque de données.

Notons que la distinction que nous faisons ici entre type de parole et type d'interaction demanderait à être davantage explorée.

largement convergentes, voire identiques à de nombreux points de vue. Pour ces deux situations de communication, nous parlons indifféremment d'oral spontané. Oral spontané est ici entendu par opposition à la parole qui oralise un texte préalablement écrit (*read speech* vs *unprepared speech*) ; ce texte peut être lu (oral lu, dans la présentation des journaux télévisés, les conférences, etc.), mais il peut aussi être « récité », comme c'est le cas dans les poésies, les représentations théâtrales, etc., ou à tout le moins préparé à l'avance. Dans ces différents cas, il s'agit d'oral planifié.

Des divergences entre ces deux grands types de productions, oral spontané et oral planifié, existent, que l'on peut synthétiser comme suit (Listerri 1996) :

- alors que la planification du texte écrit donne – en général – à l'oral un aspect fluide et lisse, sans accroc, la parole spontanée comporte nombre de disfluences (nous aurons l'occasion d'y revenir au chapitre suivant), celles-là même qui compliquent considérablement la tâche de transcription, par exemple, ou qui diminuent les performances des systèmes de reconnaissance de la parole (Adda-Decker *et al.* 2003) ;
- à l'oral spontané, la notion de phrase est sans pertinence (cf. chap. 1 et 4), alors que l'écrit oralisé garde les structures qui sont celles de l'écrit ;
- les interruptions, les chevauchements de parole ou encore les conversations en parallèle sont des phénomènes courants dans l'oral spontané, qui n'apparaissent que marginalement dans les polylogues de l'écrit destinés à être oralisés.

Nous venons de mentionner le caractère plus formel de l'entrevue, où les rôles sont clairement identifiés. Il faut noter que dans certaines entrevues, l'intervieweur et l'informateur sont familiers²³, se tutoient, etc., ce qui rend à la conversation un caractère moins formel. De plus, dans certains entretiens semi-dirigés (le corpus Chalon notamment), on a de nombreux récits de vie. Dans ces cas, les frontières établies selon le critère conversation libre vs entretien semi-dirigé ne sont plus si nettes.

²³ Nous pouvons citer les enregistrements magPM1r ou encore accCP1r où l'informateur est le frère de l'enquêtrice. On voit donc que les limites définies entre les types d'entrevues peuvent ne pas être totalement étanches, et l'on comprend l'intérêt de documenter ainsi chaque enregistrement, chaque texte et les conditions qui ont présidé à leur constitution (prise de contact entre les locuteurs, conditions d'enregistrement, normes de transcription, etc.).

Ce type d'information figure dans la fiche d'identification attachée à l'enregistrement. De même, par exemple, le fait que le locuteur paraisse tendu, pressé ou encore fatigué est consigné.

Ajoutons encore que les enregistrements n'ont pas été faits dans une situation contrainte de type laboratoire, mais en situation réelle (dans les écoles, au domicile des personnes, sur leur lieu de travail, etc.).

Par ailleurs, aussi bien dans les entrevues que dans les conversations familières, les locuteurs savent qu'ils sont enregistrés²⁴, avec les effets que l'on ne peut négliger liés au paradoxe de l'observateur (Labov 1973).

4.6.4 Les locuteurs²⁵

Si les deux critères qui précèdent procèdent principalement à une sélection par défaut, l'échantillonnage des locuteurs²⁶ s'est fait en fonction de leur profil selon les quatre paramètres suivants : le sexe, l'âge, le degré de scolarité et la localisation géographique.

Nous avons tenté de combiner ces paramètres entre eux, afin d'obtenir des profils variés, en tenant compte également des sous-corpus dans lesquels les locuteurs interviennent. Idéalement, si l'on voulait avoir un locuteur qui résulte de chaque combinaison possible de ces 4 paramètres, en considérant qu'il y a 2 variantes pour le sexe, 4 pour l'âge, 7 pour le niveau d'études et 7 également pour la localisation géographique, nous aurions dû rassembler 2 x 4 x 7 x 7 locuteurs, soit 392 locuteurs. Compte tenu des contraintes pratiques, et des données disponibles dans VALIBEL²⁷, cela était totalement inenvisageable.

Dans les 60 enregistrements que nous avons retenus, 90 locuteurs interviennent, locuteurs qui se ventilent comme suit dans les catégories ayant servi de base à la sélection.

²⁴ Un seule exception : l'entrevue souVC1r.

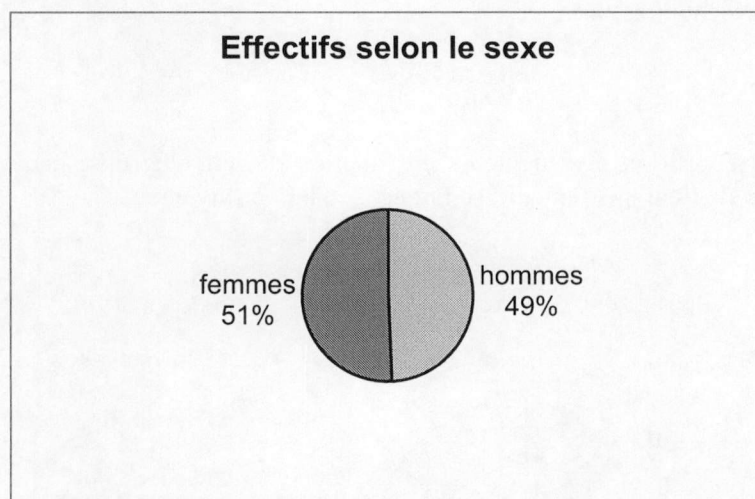
²⁵ Le tableau 1 en annexe reprend toutes les données présentées ici.

²⁶ Les chiffres que nous donnons se basent sur les codes qui sont attribués aux locuteurs. Ainsi, dans un sous-corpus, un locuteur participe à plusieurs enregistrements garde néanmoins le même code. C'est typiquement le cas de l'intervieweur. Par contre, si ce locuteur intervient également dans un autre sous-corpus de la banque de données VALIBEL, le code qui lui est attribué va changer ; il compte alors pour deux locuteurs différents. Sauf mention explicite, nous ne séparons pas ici les enquêteurs des autres locuteurs. Ils apparaissent donc au même titre que les locuteurs dans les chiffres que nous donnons.

²⁷ À titre d'exemple, dans le corpus des rhétos (dernière année de l'enseignement secondaire), les locuteurs de la province du Luxembourg se répartissent de la manière suivante : les 6 locuteurs masculins sont dans l'enseignement général, tandis que les 12 locutrices sont dans l'enseignement technique ou professionnel. Sans raffiner nécessairement à ce point les paramètres pris en considération, il n'était tout simplement pas possible d'avoir pour les 4 tranches d'âge considérées un locuteur de chaque sexe n'ayant que le niveau de scolarisation de l'enseignement primaire.

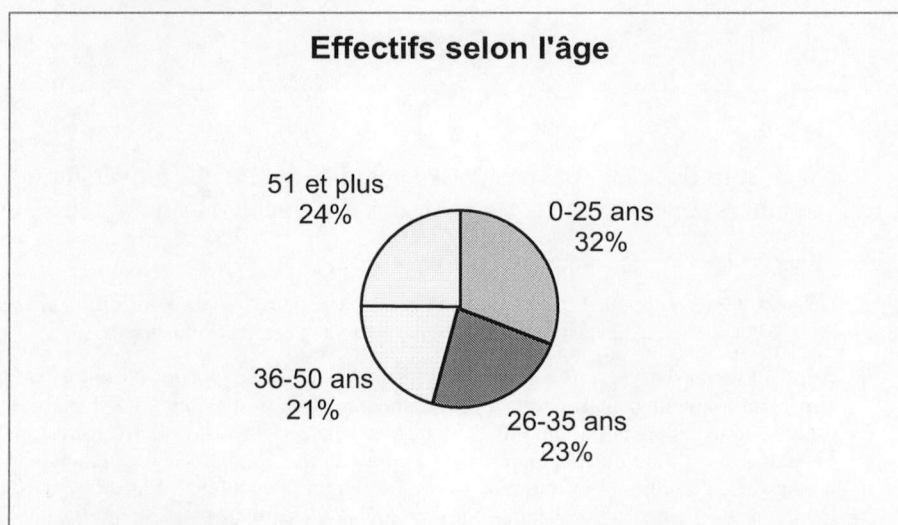
4.6.4.1 *Le sexe*

Selon le critère classique de la sociolinguistique qui concerne le sexe du locuteur, nous avons tenté de faire une répartition égale entre hommes et femmes. Notre corpus fait intervenir 44 femmes et 46 hommes.



4.6.4.2 *L'âge*

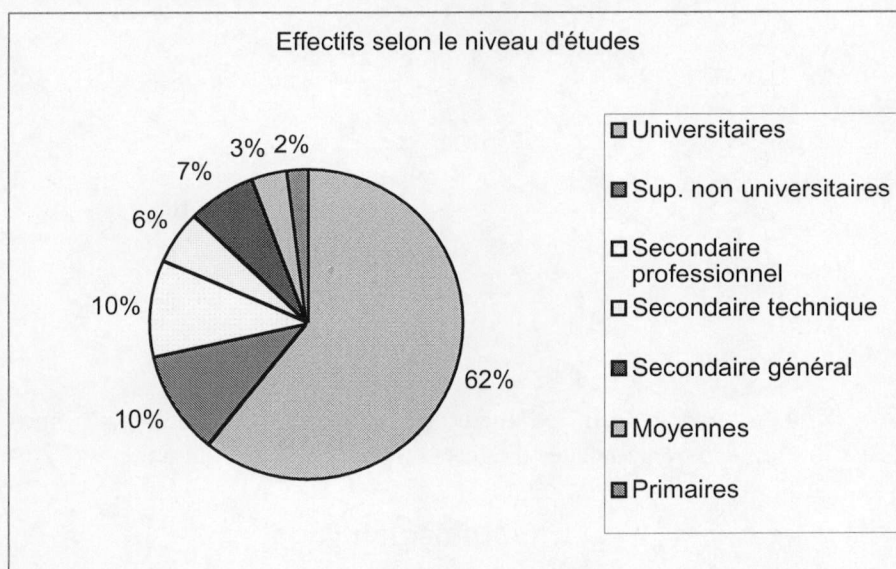
Nous avons réparti les locuteurs selon quatre tranches d'âge. La distribution dans ces tranches d'âge se fait de la manière suivante :



Le plus grand nombre de locuteurs se situe dans la première tranche d'âge, celle des 0-25 ans. On a trois locuteurs très jeunes (1, 5 et 8 ans) dans l'entrevue souVC1r, qui est une conversation informelle durant un apéritif²⁸. Hormis ces cas, les autres locuteurs sont de jeunes adultes : la plupart du temps, soit ils interviennent dans le corpus sur l'insécurité linguistique des rhéto, soit il s'agit des enquêteurs eux-mêmes (étudiants ou doctorants). Le choix de comptabiliser l'intervieweur parmi les locuteurs produit en partie la sur-représentation observée.

4.6.4.3 Le degré de scolarité

La répartition des locuteurs en fonction de leur degré de scolarité au moment de l'enregistrement²⁹ se fait de la manière suivante :



On le voit, dans les sept catégories possibles, celle des universitaires est très largement sur-représentée. Deux raisons principales en sont la cause :

²⁸ On peut s'interroger sur l'intérêt de coder un locuteur aussi jeune : en fait, il intervient dans le cours de la conversation, qui est plusieurs fois interrompue à cause de lui.

²⁹ Nous n'avons pas séparé le niveau d'études selon que celles-ci sont terminées ou non. Ainsi, un locuteur peut figurer ici en secondaire général soit parce qu'il n'a pas fait d'études universitaires ou non universitaires, soit parce qu'il est effectivement toujours en secondaire (alors qu'il s'apprête peut-être à entrer à l'université). Dans les chiffres donnés ici, nous avons néanmoins exclu les 3 jeunes locuteurs de souVC1r. Les 2 locuteurs du primaire sont donc bien des locuteurs adultes n'ayant que ce diplôme d'école primaire (accPH1 et chaBR1).

-
- les enquêteurs sont tous universitaires, étudiants ou chercheurs ;
 - les thématiques de recherche ont induit d'une certaine manière le profil des locuteurs en termes de niveau de scolarisation. Ainsi, par exemple, les vastes enquêtes sociolinguistiques sur l'insécurité linguistique se composent de cinq sous-corpus qui visaient à étudier le phénomène chez cinq populations cibles : les journalistes de la presse écrite d'une part et de la presse audiovisuelle d'autre part, les décideurs économiques (*grosso modo* des cadres), les hommes politiques et les étudiants de dernière année de l'enseignement secondaire. Hormis dans cette dernière catégorie, les locuteurs des quatre premiers sous-corpus cités ont pratiquement tous un diplôme universitaire, ou du supérieur non universitaire. Ainsi, parmi les 50 locuteurs du corpus des cadres (ilc), 42 ont un diplôme universitaire, 5 un diplôme du supérieur et 3 possèdent un diplôme de l'enseignement secondaire général. De la même manière dans l'enquête qui concerne les journalistes de la presse écrite, 96 % d'entre eux (sur un total de 50 locuteurs) ont un diplôme universitaire ou du supérieur. La recherche menée sur la norme de prononciation et les accents régionaux (sous-corpus *nor*) interrogeait des grammairiens, des linguistes, des professeurs de français ou de diction, tous avec un niveau d'études élevé.

4.6.4.4 La localisation géographique

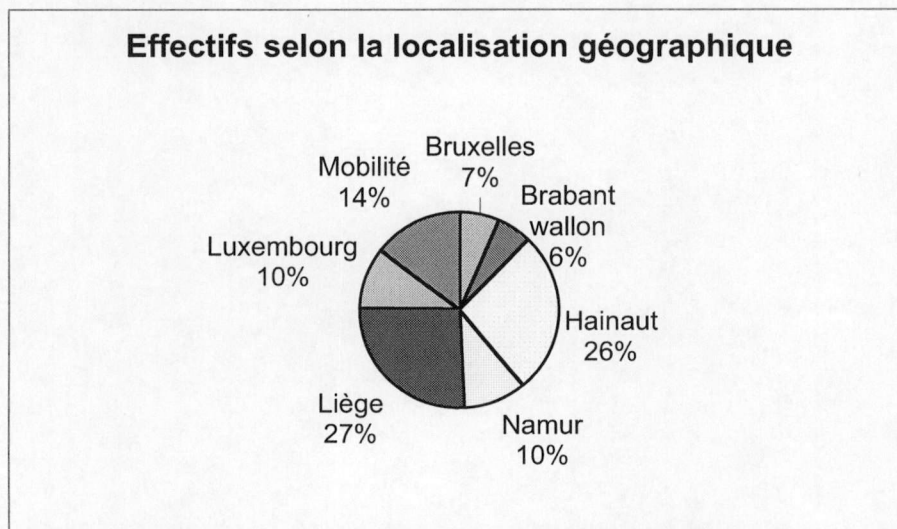
La localisation géographique comporte sept variables, dont six correspondent aux provinces francophones de Belgique que l'on peut voir sur la carte ci-dessous : Bruxelles, Brabant wallon, Hainaut, Liège, Luxembourg, Namur.

uniquement comme locutrice liégeoise. Les locuteurs de ce type entrent donc dans une catégorie que nous appelons « mobilité ».

Ce critère concernant les locuteurs a été essentiel pour effectuer un premier tri parmi les enregistrements retenus dans chacun des sous-corpus de la banque de données. Nous souhaitions en effet que les enregistrements sélectionnés dans chacun des sous-corpus soit distribué au mieux entre les zones géographiques. Cela n'a pas toujours été possible. Ainsi par exemple, le *Corpus Chalon (cha)*, constitué pour étudier la productivité lexicale des corpus oraux, ne donne la parole qu'à des locuteurs liégeois ; le *Corpus Accent (acc)* a été réalisé dans le Hainaut ; le corpus *ilj* (insécurité linguistique des journalistes de la presse audiovisuelle) n'a aucun locuteur du Luxembourg ; sur les 108 entrevues du corpus *ilr* (insécurité linguistique des rhétos), pas une seule n'a été réalisée dans une école du Brabant wallon.

Nous aurions pu rééquilibrer les données en choisissant dans les autres sous-corpus des locuteurs qui auraient eu les caractéristiques qui nous manquaient, mais nous avons décidé de privilégier la diversité à l'intérieur de chacun des autres sous-corpus, lorsque celle-ci était présente au départ.

La répartition selon la localisation géographique des 90 locuteurs est la suivante :



La province du Brabant wallon³² est la moins représentée, et cela déjà dans l'ensemble de la banque de données VALIBEL. Suit ensuite Bruxelles,

³² Cela en accord avec la maxime qui dit que le cordonnier est le plus mal chaussé. L'Université catholique de Louvain se situe en effet dans le Brabant wallon.

dont le *Corpus Marolles* (*mar*), constitué d'entretiens réalisés dans le quartier populaire des Marolles n'a pas été retenu, pour des raisons liées notamment à la fréquence élevée dans le texte du dialecte bruxellois³³. C'est également pour des raisons de changements de code trop nombreux (*code switching*) que le *Corpus Diglossie* (*dig*) a été écarté, et avec lui les locuteurs de la province du Luxembourg.

Dans la mesure du possible, nous avons tenté de faire varier à l'intérieur de chacun des sous-corpus ces quatre paramètres de sexe, âge, niveau d'étude et localisation géographique. Comme nous l'avons montré, certaines combinaisons étaient impossibles (pas de locuteurs non liégeois dans le *Corpus Chalon* (*cha*) par exemple).

³³ Mais également pour des questions d'autorisation à réutiliser les données dans un autre cadre que celui de l'enquête réalisée à l'époque.

4.6.5 Tableau synoptique des textes qui composent le corpus

Code du corpus	Intitulé	thématique de recherche	année d'enregistrement	nbre de textes
nor	norme	Norme(s) de prononciation et accents régionaux. Réflexions et enquêtes sur un aspect du français parlé en Belgique francophone et ses incidences dans l'enseignement. (mémoire de licence, UCL)	1987	6
acc	accent	Les locuteurs et l'accent régional. Enquête sociolinguistique dans la région de Charleroi. (mémoire de licence, UCL)	1989	5
mag	magasin	La variation lexicale (les emplettes) (enregistrement réalisé à Bastogne)	1990	1
cha	Chalon	La productivité lexicale des corpus oraux	1992	5
ile	insécurité linguistique de la presse écrite	Représentations linguistiques des journalistes belges francophones de la presse écrite	1993	5
ilj	insécurité linguistique des journalistes	Représentations linguistiques des journalistes belges francophones de la presse audiovisuelle	1993	6
ilc	insécurité linguistique des cadres	Les décideurs économiques francophones face au français	1994	5
ilp	insécurité linguistique des politiciens	L'insécurité linguistique en Belgique francophone. Enquêtes sur les représentations linguistiques de la classe politique. (mémoire de licence, UCL)	1992-1997	7
ilr	insécurité linguistique des rhétores	Les jeunes Belges francophones et l'insécurité linguistique.	1991	11
sou	souper	Ellipse et cohésion d'après Halliday et Hasan. Application à un corpus oral spontané. (mémoire de licence, UCL)	1988	1
fam	famille	Conversations familiales.	2003	3
bla		Conversations libres enregistrées dans le cadre du projet PFC (Phonologie du Français Contemporain).	2003-2004	5

Les 60 textes qui composent le corpus font intervenir 90 locuteurs et totalisent 443 047 mots. Cela correspond *grosso modo* à 40 heures d'enregistrement.

4.6.6 Répartition des textes

La durée moyenne d'un enregistrement est de 40 minutes ; l'enregistrement le plus bref est de 7 minutes (chaSR1r), le plus long de 86 minutes (norHJ1r). En ce qui concerne les transcriptions, celles-ci varient de 1 286 mots (l'enregistrement de 7 minutes) à 14 189 mots (celui de 86 minutes), avec une moyenne de 7 384 mots par texte.

Le tableau suivant monte la distribution des textes en fonction du nombre de mots. On a dégagé six catégories, le découpage se faisant par tranche de 3000 mots. La colonne de gauche indique, pour chacune de ces catégories, le pourcentage de textes par rapport au nombre total de 60 textes ; la colonne de droite indique le pourcentage de mots concernés par les textes de la catégorie, par rapport au nombre total de mots du corpus.

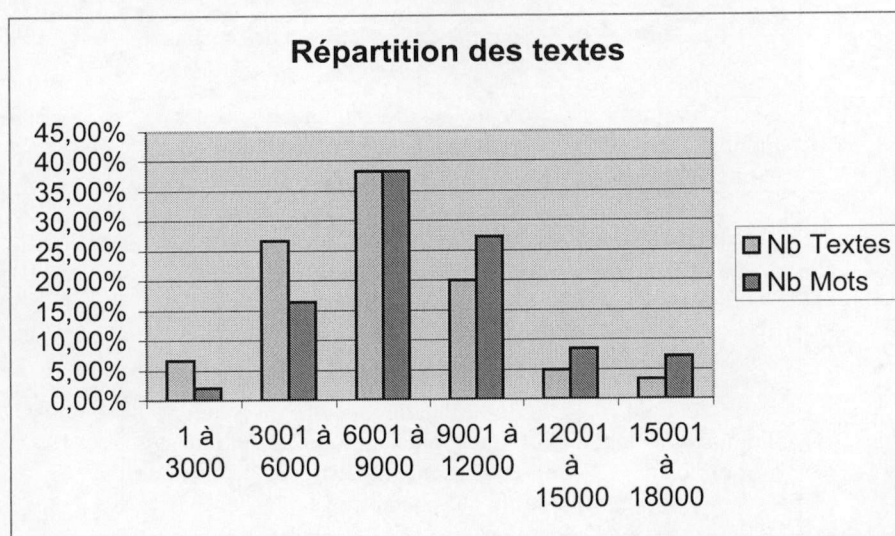


Figure 7. Distribution des textes selon le nombre de mots

Ainsi, on peut voir que le corpus contient plus de textes très brefs (de 1 à 3000 mots, 4 textes, soit 6,67 % de l'ensemble des textes) que très longs (de 15 001 à 18 000 mots, 2 textes, soit seulement 3,33 % des textes), mais la masse de données textuelles représentée par les textes très longs (7,23 %) dépasse largement celle des textes très courts (seulement 2,04 %). En ce qui concerne les textes dont la longueur est proche de la longueur moyenne des textes du corpus (les 7 480 mots calculés ci-dessus), on voit qu'ils représentent non seulement la majorité en nombre de textes concernés, mais

également en nombre de mots (dans les deux cas, un peu plus de 38 % des effectifs).

4.6.6.1 *Nombre de locuteurs par enregistrement*

Les enregistrements sont majoritairement des dialogues, comme le montrent les chiffres du tableau ci-dessous :

Nombre de locuteurs	Nombre de textes
2	54
3	4
4	1
7	1

Tableau 3. *Nombre de locuteurs par enregistrement*

Cette prédominance des dialogues est due au nombre élevé d'entrevues semi-dirigées reprises dans notre corpus.

L'enregistrement qui fait intervenir 7 locuteurs est souVC1r, mais on a vu que certains de ceux-ci interviennent très peu dans la conversation.

4.6.6.2 *Le poids des locuteurs*

Les données chiffrées que nous venons de présenter concernent la répartition des locuteurs dans l'ensemble des entrevues choisies, chaque locuteur ayant un poids égal. Or, certains locuteurs interviennent considérablement plus que d'autres en ce qui concerne leur temps de parole et, partant, le nombre de mots dans les transcriptions. Ainsi par exemple, accCT0, l'intervieweuse du Corpus Accent (*acc*), comptabilise le plus grand nombre de mots dans les transcriptions : 18 185 mots pour les 5 enregistrements dans lesquels elle intervient. Elle compte pourtant dans les données ci-dessus comme une seule locutrice qui a les caractéristiques suivantes : femme, de la tranche d'âge 0-25 ans, localisée dans le Hainaut et ayant un diplôme universitaire. De la même manière, ilrMF0, intervieweuse des élèves de secondaire, totalise 16 185 mots dans les 11 entrevues sélectionnées. Sous un autre code, c'est la même intervieweuse que dans le corpus *Cadres* (ilcMF0)³⁴, où elle totalise 5270 mots. Sur l'ensemble de

³⁴ Ayant un autre code, elle compte donc pour deux locutrices différentes dans les données de répartition reprises dans les tableaux.

notre corpus, cette locutrice comptabilise donc 21 455 mots. À l'inverse, la transcription de la locutrice chaSR1 ayant les caractéristiques « femme, de la tranche d'âge 51 et plus, localisée à Liège avec un diplôme de moyennes » ne comporte que 964 mots soit plus de 10 fois moins. Ainsi, on constate que si on répartissait les locuteurs en fonction du nombre de mots et non plus de leur code dans les transcriptions, on aurait des données différentes de celles présentées dans les tableaux proposés ci-dessus tant pour la localisation géographique que pour le niveau d'études, etc.

Le corpus comprend, en plus des 90 locuteurs détaillés, une dizaine de locuteurs dont nous n'avons pas parlé dans les données ci-dessus. Ils interviennent en effet de manière très brève et ponctuelle, et nous ne possédons pas d'information permettant de les caractériser comme les autres locuteurs. En général, il s'agit de personnes entrant dans la pièce et ignorant qu'un enregistrement est en train de se faire, d'appels audibles via un interphone, etc. Ces locuteurs reçoivent le code du sous-corpus ainsi que des initiales XX (pour « inconnu ») et un chiffre permettant de les distinguer les uns des autres à l'intérieur du même sous-corpus. Ils sont pour nous considérés comme des locuteurs marginaux, et les exemples que nous donnerons dans ce travail ne leur seront jamais attribués.

Si l'on considère chaque locuteur intervenant dans une entrevue comme un locuteur différent (accCT0 est donc 5 locutrices distinctes) et que l'on exclut les locuteurs sans identifications (géographique, socio-professionnelle, etc.), on a 133 locuteurs dans le corpus ; la moyenne du nombre de mots par locuteur est de 3331 mots³⁵.

Séparant ainsi les locuteurs, le nombre le plus important de mots de accCT0 dans une entrevue est de 4527 ; elle arrive en 40^e place si l'on classe les locuteurs par fréquence de mots décroissante. Selon ce classement, on constate que dans les entrevues, les locuteurs occupant le rôle d'intervieweur apparaissent pratiquement tous dans le bas du tableau. Quand on s'attache à chaque entrevue individuellement, on constate une répartition inégale de la parole, où l'un des locuteurs a un usage prépondérant de la parole. On observe une constante : le nombre de mots de l'informateur est toujours supérieur au nombre de mots de l'intervieweur. Rien d'étonnant à cela. Seuls les mauvais journalistes parlent plus que les personnes qu'ils interviewent. De manière générale, c'est dans le corpus *Chalon* (cha) que la parole est la plus inégalement répartie, en termes de nombre de mots, entre intervieweuse et informateur, allant, dans l'entrevue chaBR1r, jusqu'à 99,37 % de la parole en faveur de l'interviewé.

³⁵ Si l'on exclut les 5 locuteurs (caractérisés sociologiquement) ayant un nombre de mots inférieur à 100, on obtient alors une moyenne de 3500 mots par locuteur.

Le tableau présenté en annexe 2 donne pour chaque texte du corpus la répartition du nombre de mots par locuteur.

4.6.6.3 *Les tours de parole*

Le nombre de tours de parole³⁶ par entrevue est très variable puisqu'il va de 1392 tours dans famES1r (conversation informelle de trois personnes lors d'un souper) à 80 tours pour l'entrevue la plus brève déjà citée (chaSR1r). La moyenne est de 390 tours de parole par texte.

Les textes les plus longs en nombre de mots ne contiennent pas nécessairement le plus grand nombre de tours de parole et inversement. Ainsi, par exemple, le troisième texte en termes de nombre de mots (iljPF1r, 13 008 mots) ne contient que 321 tours de parole (31^e rang). À l'inverse, la conversation souVC1r a un nombre de tours de parole élevé (921, rang 3) alors qu'il se place au 31^e rang en termes de nombre de mots. Différents facteurs peuvent expliquer ces répartitions, et pour souVC1r, on avancera le nombre élevé des participants à la conversation qui multiplient les tours de parole.

La longueur moyenne des tours de parole par entrevue est elle aussi variable : elle va de 70,15 mots par tour (chaCL1r) à 7,07 mots (souVC1r) avec une moyenne de 19 mots par tour de parole. Toutes les conversations les plus informelles du corpus sont en-dessous de cette moyenne, hormis blaAT11 sur lequel nous allons revenir (cf. annexe 3). Parmi les quatre premiers textes où le nombre de mots par tour est le plus élevé, trois appartiennent au corpus *Chalon*, dont nous avons dit ci-dessus que l'intervieweuse laissait abondamment parler l'informateur, qui est amené à raconter des souvenirs de son enfance. Ceci est confirmé quand on observe la moyenne non plus simplement par entrevue, mais par locuteur.

En effet, la moyenne du nombre de mots établie selon chaque locuteur offre un éparpillement encore plus grand des données : de 127,43 mots à

³⁶ Nous donnons ici à « tour de parole » un autre sens que celui traditionnellement employé dans les conventions de VALIBEL (voir chap. 1), puisque nous comptons comme un tour de parole à part entière les interventions internes au tour de parole. Ces chevauchements internes, qui ne nécessitent pas un retour à la ligne dans les conventions de transcription (et ne sont donc pas considérés comme des tours de parole à part entière dans la mesure où le locuteur qui parlait ne « perd » pas la parole), ont été ici comptabilisés comme des tours de parole dans le sens où il s'agit d'une intervention d'un locuteur et que, par exemple en termes de nombre de mots, ces paroles lui sont effectivement attribuées, et non au locuteur dans le tour de parole (au sens VALIBEL) duquel elles sont insérées. On ne s'étonnera donc pas de voir un nombre de tours de parole non égal entre les locuteurs dans les dialogues.

1,53 mots par locuteur³⁷, avec au rang 1 le locuteur chaCL1 (cf. annexe 4). ChaBR1 occupe quant à lui le rang 3 avec 98,91 mots en moyenne par tour. Les résultats obtenus sur la totalité des entrevues sont donc encore accentués quand on ne regarde que les locuteurs individuellement.

Cet éparpillement des données nous permet de mettre en évidence un phénomène qui apparaît très clairement dans le tableau de l'annexe 4 :

1. Tous les locuteurs ayant le rôle d'intervieweur³⁸ dans les entrevues semi-dirigées ont une moyenne de nombre de mots par tour de parole relativement faible ; ils apparaissent tous dans le bas du tableau.
2. Inversément, les locuteurs faisant figure d'interviewés ont un nombre de mots par tour de parole supérieur à la moyenne.

Cela montre clairement que le genre de l'entrevue respecte un même schéma, relativement stable : l'interviewé s'exprime longuement, là où les prises de parole de l'intervieweur sont relativement brèves. Certains tours de parole plus longs, notamment lorsque le locuteur pose des questions, n'infléchissent pas cette tendance générale. Cette moyenne très basse est due au nombre élevé de chevauchements internes aux tours de parole de l'informateur, dans lesquels l'intervieweur a des répliques très brèves, sorte de marques d'écoute qui consistent principalement à encourager le locuteur à poursuivre son discours (les *régulateurs verbaux* de Gaulmyn 1987, la *régulation* de Cosnier 1987 le *feedback* de Fontaney 1987 ou encore les *continuers* (continueurs) de Schegloff 1981)³⁹.

3. Dans les conversations informelles, certains locuteurs obtiennent également un nombre de mots par tour de parole en dessous de la moyenne. Les trois exceptions notables concernent blaAT1 (38,31 mots/tour), blaJV1 (29,34 mots/tour) et blaJD1 (25,03 mots/tour). Ces résultats peuvent s'expliquer en référence aux conditions de récolte des données. En effet, s'il s'agit bien de conversations informelles entre personnes qui se connaissent très bien⁴⁰, les

³⁷ 1,53 pour l'un des cinq locuteurs (caractérisés sociologiquement) ayant un nombre de mots inférieur à 100. En excluant ce locuteur des comptes, la moyenne la plus basse de nombre de mots par tour de parole chez un locuteur est de 4,33.

³⁸ Facilement repérables au chiffre 0 à la fin de leur code locuteur.

³⁹ Ces interventions courtes de l'enquêteur dans le tour de parole de l'informateur sont relativement fréquentes, et toujours plus nombreuses que les interventions de l'informateur dans le tour de parole de l'enquêteur

⁴⁰ Les jeunes locuteurs de blaAT11 et blaJV11 sont des amis de longue date, et blaJD1 connaît son interlocutrice depuis plus de 50 ans.

locuteurs qui semblent faire ici exception ont en fait été sélectionnés dans le cadre du projet PFC (Phonologie du Français Contemporain), en fonction de leur profil (âge, niveau d'études, etc.). Le but était de recueillir des données langagières ainsi caractérisées, et représentées par ces locuteurs, lors d'une conversation libre⁴¹. Les autres personnes qui participent à la conversation sont donc là, d'une certaine manière, pour les faire parler. Elles ont un rôle analogue, *mutatis mutandis*, à celui d'interviewer. Elles obtiennent quant à elles un nombre de mots par tour de parole très faible, tout comme les enquêteurs des entrevues semi-dirigées (par exemple, blaCD1 dans le texte blaJD11, avec une moyenne de 5,85 mots par tour de parole). Cela dit, d'autres enregistrements faits dans le cadre de PFC donnent une répartition plus équitable en termes de mots, de nombre de tours de parole et de longueur des tours de parole entre les participants à la conversation (blaBD1 et blaGD1 dans blaGD11, par exemple)⁴².

Ainsi, dans les entrevues, la répartition est clairement en faveur d'une plus grande prise de parole par les personnes interviewées. Il serait d'ailleurs paradoxal que l'interviewer parle plus que son informateur. On s'attend logiquement à ce que les prises de parole ne soient pas équilibrées. Cela implique également d'autres caractéristiques : on peut par exemple s'attendre à ce que les interruptions soient moins fréquentes que dans une conversation spontanée entre locuteurs « égaux ».

4.7 Les formes du corpus

Nous avons présenté ci-dessus des données sur le corpus, en prenant soin de distinguer les locuteurs et les textes qui le composent. Nous voudrions maintenant avoir une vision plus globale, en considérant le corpus dans sa totalité et les formes les plus fréquentes qui le composent.

⁴¹ Par opposition à la conversation guidée qui consistait à répondre à une série de questions et à lire une liste de mots et un texte (Hambye 2005).

⁴² On voit ici clairement l'importance de tenir compte des conditions dans lesquelles les données sont recueillies, et comment ces conditions permettent de formuler des hypothèses pour l'analyse.

Nous classons néanmoins ces textes dans les conversations informelles, car non seulement les rôles des locuteurs y sont nettement moins distribués (dans l'esprit de celui qui sert d'informateur en tout cas) que pour les entrevues semi-dirigées, mais de plus le sujet de la conversation est quant à lui totalement libre.

Nous appelons *token*, comme il est traditionnel de le faire, une chaîne de caractères située entre deux séparateurs (blanc, apostrophe, tiret). Un token ainsi défini correspond donc à ce que nous appelions *mots* ci-dessus.

Dans notre corpus, nous avons 14 689 chaînes de caractères différentes. Parmi celles-ci, 6551 sont des hapax, c'est-à-dire qu'elles n'apparaissent qu'une fois dans le corpus. Ce chiffre constitue 44,6 % de l'ensemble des formes. Il n'est guère étonnant, car il a été montré depuis longtemps que dans tout corpus, on a un petit nombre de formes très fréquentes, un grand nombre de formes peu fréquentes et entre les deux, un groupe intermédiaire (voir aussi chap. 3.2).

Parmi ces hapax, on a des amorces de morphèmes (cf. chap. 3.4), des noms propres ou encore des sigles. Mais la plupart des formes ne sont pas des mots rares, précieux ou vieillis, mais semblent bien appartenir au vocabulaire de base du français. On a ainsi des verbes (*sortez, essayons, assistant, dépendent, fabriquez, tenus, examiner*), des adjectifs (*mentale, véritable, profitable, immédiate*), des noms (*euthanasie, célébrité, chauffage, discothèques, instrument*), ou encore des adverbes (*bizarrement, dangereusement*). En fait, en ce qui concerne les formes elles-mêmes (et non les lemmes ou la catégorie grammaticale), on ne peut guère tirer de conclusion sur un corpus tel que le nôtre, car la masse de données dont on aurait besoin pour des études statistiques pertinentes, en termes de lexique, doit être beaucoup plus importante (Francard *et al.* 2002). Ainsi par exemple, on ne dira pas, suite à la liste que nous donnons ci-dessus, que *chauffage* est un mot rare dans le français parlé en Belgique francophone⁴³.

Nous venons d'aborder les mots rares, voyons maintenant quelles sont les formes les plus fréquentes du corpus. Nous présentons, par ordre de fréquence décroissant, les 100 tokens les plus fréquents de notre corpus. Il s'agit donc bien ici simplement de chaînes de caractères. Celles-ci peuvent donc correspondre à plusieurs lemmes différents : *est* pour *être* ou pour *est*, et dans ce cas, qu'il s'agisse du nom ou de l'adjectif ; *le* pour le déterminant ou pour le pronom ; etc. Notons encore que comme l'apostrophe est considérée comme un séparateur, elle n'apparaît pas à côté des formes élidées. On a donc dans ce tableau *c, l, d*, etc., qui correspondent en fait à *c', l' et d'*, et non aux lettres. En effet, la liste des tokens donnée ici distingue les formes en fonction de la casse : une même chaîne de caractères, selon qu'elle comporte des majuscules ou des minuscules, est donc séparée dans les comptages.

Le premier tableau est celui qui correspond aux formes de notre corpus ; pour comparaison, nous présentons dans un deuxième tableau les données

⁴³ L'équipe VALIBEL cite fréquemment le cas du mot *plasticine* (*pâte à modeler*), absent des 4 millions de mots de la banque de données textuelles, mais connu et utilisé par tous les petits belges dès l'école maternelle.

obtenues par Jean Véronis sur un corpus de français parlé d'un million de mots (*Corpaix*, version mai 2000)⁴⁴.

13331	est	3777	une	1837	quoi	1200	sont
10699	je	3589	ce	1806	parce	1175	ben
10698	de	3578	y	1783	moi	1175	s
9923	c	3569	vous	1769	si	1119	peu
9423	euh	3346	mm	1749	alors	1046	gens
8548	que	3250	j	1730	du	977	se
8419	et	3216	dans	1707	m	943	ont
7555	il	3182	ne	1687	comme	942	parler
7471	pas	2830	bien	1598	ah	928	suis
7220	ça	2748	non	1576	français	924	parle
7201	oui	2600	ou	1573	enfin	913	dit
7055	un	2543	tout	1560	par	887	où
6826	à	2541	tu	1507	me	885	va
6751	le	2522	pour	1495	dire	877	faire
6426	on	2314	hein	1472	bon	847	sur
5842	la	2274	accent	1466	au	819	wallon
5395	des	2157	ai	1449	peut	803	ouais
5384	les	2150	plus	1376	être	798	beaucoup
5317	a	2094	n	1336	crois	786	nous
5307	mais	2074	quand	1328	était	763	puis
5118	qui	2008	même	1319	avec	740	avait
4766	l	1952	fait	1282	sais	736	faut
4674	en	1948	ils	1271	très	730	deux
4310	qu	1934	donc	1211	elle	719	exemple
4018	d	1869	là	1204	aussi	709	moins

Tableau 4. Formes de notre corpus classées par fréquence décroissante

⁴⁴ <http://www.up.univ-mrs.fr/~veronis>.

27 376	est	7873	ce	3850	moi	2560	être
26 507	de	7795	d'	3833	avait	2503	quoi
21 465	c'	7186	oui	3772	même	2448	dit
20 879	euh	6837	tu	3745	nous	2415	ont
20 212	il	6820	j'	3708	bien	2250	n'
20 162	et	6660	dans	3586	par	2224	ne
17 785	que	6446	pour	3570	si	2177	sais
17 191	le	6311	vous	3477	ou	2145	va
16 761	on	6157	ils	3337	se	2104	aussi
16 512	la	5172	fait	3330	hein	2015	voilà
16 026	à	5145	tout	3253	avec	1928	faut
15 114	je	5082	était	3171	dire	1861	beaucoup
14 891	pas	5070	plus	3125	très	1856	lui
14825	les	5060	là	3100	ouais	1840	deux
14022	a	5044	alors	2936	ah	1822	après
13533	un	4722	elle	2936	ben	1757	m'
12723	ça	4657	donc	2895	sont	1653	mon
12635	des	4577	parce	2846	s'	1637	suis
11530	en	4564	du	2803	puis	1594	où
11441	qui	4443	non	2795	peu	1521	hum
10825	l'	4354	ai	2786	faire	1462	mh
9728	y	4250	au	2762	sur	1442	gens
9503	qu'	quand	4238	2750	me	1409	cette
8501	une	comme	3974	2721	peut	1373	leur
8288	mais	bon	3969	2581	enfin	1353	mm

Tableau 5. Formes de Corpaix classées par fréquence décroissante

Dans notre corpus, la forme la plus fréquente est *est*, suivie ensuite de *je* et de *de*. On peut s'étonner de ce classement : en effet, en général, la forme la plus fréquente en français, dans les corpus écrits standard, est le *de*. (cf. Dister 2000b sur une année du *Monde*, dorénavant *LM94*). Elle arrive ici seulement en troisième position. La présence du *je* en deuxième position étonne néanmoins, même si on peut l'expliquer par le fait que les locuteurs

s'expriment à la 1^{re} personne, donnent leur avis, parlent d'eux, etc. On notera une divergence avec les données de *Corpaix* où le *je* apparaît « seulement » en 12^e position. Un autre pronom sujet, *tu*, est également présent dans notre classement en 38^e position, avec 1507 occurrences. On peut ici penser qu'il s'agit effectivement du pronom et non du participe passé du verbe *taire*. Si l'on peut faire cette hypothèse, c'est également en connaissant notre corpus. En effet, Watrin (2006), dans un corpus économique, n'a que de rares occurrences de *tu*, et aucune ne concerne le pronom.

Parmi ces 100 formes les plus fréquentes, on constate que peu d'entre elles sont des mots lexicaux. On a potentiellement : *est, je, pas, bien, tout, accent, fait, français, dire, peut, être, croit, était, sais, sont, gens, ont, parler, parle, dit, va, faire, wallon, avait, faut, deux, exemple* et *moins*. Nous écrivons *potentiellement* puisqu'à ce stade où l'on se concentre uniquement sur la graphie, ces formes sont ambiguës. Ainsi, si l'on considère dans cette liste *je* comme un mot lexical, c'est à cause de son analyse possible en un substantif (le *je* de la psychanalyse). Or, il est vraisemblable qu'en corpus (notre corpus en tout cas), c'est l'analyse pronom qui sera la bonne pour cette forme dans pratiquement la totalité des occurrences. Intervient donc ici une distinction que nous reprendrons plus loin (chap. 5 et 6) sur l'ambiguïté. En fait, dans la liste ci-dessus, le premier substantif non ambigu est *accent*. Ici aussi, on peut s'étonner d'un score aussi élevé pour un mot somme toute peu courant. Sa fréquence d'apparition est évidemment liée à notre corpus, principalement aux enquêtes sociolinguistiques sur la perception de l'accent et sur l'insécurité linguistique. On voit donc la nécessité impérieuse de documenter le corpus que l'on utilise, afin de ne pas faire de généralisation abusive : les données que nous fournissons concernent donc bien notre corpus, avec ses particularités propres en termes de types de parole et d'interaction, de sujet abordé, de profils de locuteurs, etc.

De la même manière, ce sont les thèmes des enquêtes qui expliquent la présence dans notre classement de *français* et *wallon* ou encore des verbes de parole *dire, parler, parle* et *dit*. Pour comparaison, parmi ces formes, seul *dire* est présent dans les 100 formes les plus fréquentes de *Corpaix*.

En fait, si les mots lexicaux présents dans le classement des formes les plus fréquentes dépendent des thèmes abordés et diffèrent d'un corpus à l'autre⁴⁵, on observe une constante en ce qui concerne les mots grammaticaux : quel que soit le corpus pris en considération (*LM94, Corpaix* ou notre corpus), ils occupent la majorité des premiers rangs. Ainsi, parmi

⁴⁵ Pour comparaison, le corpus composé de discours du roi Baudouin (de son mariage en 1961 à sa mort en 1993) comprend un seul substantif parmi les 50 tokens les plus fréquents : *pays*. Viennent ensuite *valeurs, esprit, respect, cœur, courage* (Deroubaix et Gobin 2000). On voit la thématique et l'idéologie ici abordées, dans des discours souvent faits à Noël ou lors de la fête nationale.

les 100 formes ici, on a pratiquement 70 % de mots grammaticaux et d'adverbes. Ce pourcentage augmente à 92 % si on ne tient compte que des 25 mots les plus fréquents.

En ce qui concerne la particularité de notre corpus, liée cette fois au fait qu'il s'agit d'un corpus textuel **oral**, outre la présence des pronoms *je* et *tu*, on notera la présence d'onomatopées et d'interjections absentes des listes de fréquences des corpus écrits : *euh* (en 5^e position), *mm*, *hein*, *m*, *ah* et *ben*. Ces formes sont évidemment absentes de *LM94*. On y ajoutera également le *quoi et bon* qui peuvent être ponctuants à l'oral (cf. chap. 5).

Nous verrons au chapitre 6 les implications de cette forte fréquence de mots grammaticaux sur la levée d'ambiguïtés.

4.8 Conclusions

Nous venons de présenter, peut-être un peu longuement, notre corpus. Cette étape nous paraît pourtant essentielle, afin de ne pas travailler sur des données « abstraites ».

En effet, il nous semblait important de décrire les textes qui composent notre corpus, d'autant plus que ces textes sont l'œuvre de locuteurs que nous avons la chance de pouvoir caractériser. Tout au long de ce travail, nous donnerons des exemples attestés, référencés grâce aux codes en vigueur dans la banque de données VALIBEL.

De plus, si l'étiquetage morphosyntaxique peut être vu comme une tâche mécanique, l'étude que nous allons faire de certaines disfluences renverra parfois à des locuteurs qui ont des caractéristiques particulières et ne sont pas des locuteurs « idéaux ».

Ainsi, si nous n'avons pas constitué un corpus représentatif (« Même la plus belle femme du monde ne peut donner que ce qu'elle a ») au sens où nous l'avons défini ici, nous l'avons documenté, ce qui lui permet de pouvoir être réutilisé par d'autres chercheurs.

CHAPITRE 3. SECTION 1

LES « DISFLUENCES » DE L'ORAL

iljDV1 (...) j'ai l'impression parfois que
les le voca/ le vocabulaire enfin les les /
(toux) / le langage est confus (...)
[iljDV1r]

ilrMS1 je sais pas / parler sans accent
pour moi c'est sans // sans // sans
bafouiller sans / sans sans se tromper de
mots quoi sans sans sans que la la langue
fourche quoi sans [ilrMS1r]

1 INTRODUCTION

Les études sur la langue parlée ont permis de dégager des phénomènes propres à l'oral, qu'on regroupe souvent sous l'appellation générale de *disfluences*¹. On entend par là un certain nombre de traits propres à la production de la langue parlée, d'« achoppements » dans la linéarité de l'énoncé, de marques du discours en cours d'élaboration ; ces phénomènes sont inhérents aux productions orales, même si leur fréquence semble dépendante du « type » d'oral (planifié ou non, cf. § 4.1 et § 4.6.3).

Nous consacrons un chapitre important de ce travail à l'étude de certaines de ces disfluences, parce qu'elles ont d'emblée été reconnues comme l'une des pierres d'achoppement majeures des traitements informatisés ou semi-informatisés de la langue orale. En effet, qu'il s'agisse du niveau de la transcription² (automatisée ou non) (Adda-Decker *et al.* 2003

¹ « Si la *fluence* est un terme en usage dans la langue française, celui de *disfluence* ne l'est pas, du moins pas encore. À voir la fréquence de son emploi au moins dans le monde linguistique, il y a fort à parier qu'il franchira vite les portes des dictionnaires. » (Pallaud 2004 : 83)

² La transcription automatique vise à obtenir de manière totalement automatisée des transcriptions exactes et de qualité, mais débarrassées des disfluences de l'oral.

et 2004), de la reconnaissance de la parole, du dialogue homme-machine (Bear *et al.* 1992) ou encore de l'analyse morphosyntaxique ou syntaxique (Hindle 1983), les disfluences semblent bien constituer l'écueil majeur des recherches du domaine TAL sur le parlé.

Connaissant un intérêt récent auprès des linguistes étudiant l'oral³, les disfluences ont principalement fait l'objet d'études de la part des psycholinguistes (les pionniers Maclay et Osgood 1959, Levelt 1983), des psychanalystes (on sait l'intérêt de Freud pour le lapsus, où l'attention est portée davantage sur le lexique) ou encore des psycho-cliniciens. C'est l'aspect « raté » qui a d'abord intéressé, ainsi que les mécanismes d'encodage et de décodage des énoncés (cf. Duez 2001 pour un historique de ces notions).

L'intérêt relativement tardif des linguistes – dans les années 1950, d'abord dans le monde anglo-saxon⁴ – s'explique sans doute par le fait que les recherches sur l'oral, abordé principalement d'un point de vue phonétique ou phonologique, se sont centrées sur la parole de laboratoire⁵, délaissant la parole en contexte, la parole spontanée, caractérisée justement par les disfluences.

Depuis 1999, un colloque⁶ rassemble tous les deux ans des chercheurs sur l'oral, intéressés par la production ou la perception du phénomène ou encore par les défis qu'il pose aux technologies de la parole.

Certains auteurs (Benzitoun *et al.* 2004 par exemple) réfutent le terme de *disfluence*, qui implique notamment une comparaison implicite avec l'écrit : si l'oral est disfluent, c'est par rapport à un écrit qui, évidemment, ne l'est pas. La disfluence de certains énoncés est donc opposée à la fluence de certains autres, avec les jugements de valeur que cela implique bien souvent. Comme le constate Habert (2005b : 57) :

(...) on manque ainsi de termes positifs pour décrire les régulations de l'oral, parfois fâcheusement dénommées *disfluences* par transfert de l'anglais *disfluencies*.

³ La thèse de Sandrine Henry en cours sur les répétitions, le travail de Berthille Pallaud (1999, 2002b et 2003) sur les amorces, la thèse de Maria Candéa (2000a) et les travaux de Danielle Duez (1991, 1995, 1997 et 2001) sur la pause, pour ne citer que le domaine francophone.

⁴ Voir Duez (1991) pour les principales études jusqu'en 1987.

⁵ Nous l'avons dit : c'est cela qui explique également les nombreuses études sur les pauses silencieuses à côté de celles réservées aux pauses pleines. Candéa (2000b) rappelle également l'intérêt des traités de déclamation pour la pause silencieuse.

⁶ Le dernier en date, *Diss05*, s'est tenu à Aix en septembre 2005. On citera encore la journée de l'ATALA qui s'est déroulée en avril 2005 et intitulée *Hésitations, disfluences, répétitions, faux départs : quel ordre dans le désordre ?*

D'autre part, il s'agit de faire le départ entre ce que l'on englobe dans les disfluences, et qui est propre à la planification de tout énoncé oral chez tout locuteur, et ce qui relève plus particulièrement d'une pathologie du langage⁷.

On sait de plus l'effet désastreux provoqué par la lecture des transcriptions qui consignent ces disfluences chez des lecteurs non habitués à l'exercice : c'est qu'il ne peut s'agir que de locuteurs ayant de graves problèmes d'expression ou maîtrisant mal la langue. Or, comme le dit Pallaud (2004) :

Ces achoppements ne définissent pas une catégorie d'énoncés ; tout au plus, constate-t-on des variations intra- et inter- individuelles. On est donc dans la variante ; les linguistes et psycholinguistes ont, alors, préféré le terme de disfluence à celui de dysfluence qu'ils réservent au domaine spécialisé des pathologies du langage. (...) Le terme de dysfluence est en effet dévolu à tout ce qui appartient au domaine de la pathologie du langage.

Dans la littérature, les auteurs assimilent fréquemment les disfluences à des manifestations d'hésitation de la part du locuteur. C'est d'ailleurs le terme *hésitation* qui a prévalu longtemps dans la plupart des études sur le sujet, autant dans le monde anglo-saxon que francophone. Selon Candéa (2000a : 13 et 2000b : 2), c'est à l'étude de Maclay et Osgood (1959) que l'on doit l'emploi du terme *hesitation phenomena*, « au détriment des termes de type "disturbances" ou "disfluencies" ».

Ces auteurs ont établi une typologie des phénomènes d'hésitation qui a été fréquemment reprise dans les études postérieures. Selon eux, l'hésitation se manifeste dans la parole des locuteurs par les pauses remplies, les syllabes allongées, les faux départs (repris ou non), les répétitions (non sémantiques, cf. chap. 3.2) et les pauses silencieuses non syntaxiques (cf. chap. 3.4).

On assiste pourtant aujourd'hui à un retour en grâce du terme *disfluence* (cf. le colloque *DISS* cité ci-dessus). Dans la littérature anglo-saxonne, on rencontre également parfois le terme *nonfluencies* (Hindle 1983, Neslon 1996).

De notre point de vue, le terme *hésitation* a ceci de gênant qu'il fait référence à un processus cognitif. Ainsi, malgré les critiques que l'on peut émettre sur le choix du terme *disfluence*, nous le préférons à celui de *hésitation*. En effet, nous ne nous intéresserons pas ici aux fonctions des disfluences en termes cognitifs ou selon leurs éventuelles fonctions

⁷ Cf. exergue de la section 2 de ce chapitre consacrée aux répétitions, où est mis sur le compte du bégaiement pathologique des répétitions tout à fait banales chez tout locuteur (hormis le *cacacapable* qui lui est bien dysfluent).

énonciatives – qu'elles soient des marques du travail de formulation (Morel et Danon-Boileau 1998), signe d'hésitation, signe d'un malaise ou d'une difficulté d'encodage, recherche du mot juste dans la mémoire du locuteur, marque signifiant la volonté de poursuivre l'énoncé, etc. – mais à leur fréquence et leur régularité en corpus. Notre but est en effet de cerner le phénomène du point de vue de la morphosyntaxe et d'en voir les éventuelles régularités afin de les formaliser dans la perspective d'une analyse automatisée ou semi-automatisée.

Cela étant dit, nous avons la chance d'avoir des données dans lesquelles les locuteurs sont identifiés en tant que tels, et auxquels un certain nombre de paramètres sociolinguistiques et sociologiques sont attribués (cf. chap. 2). Lorsque des divergences sensibles dans nos données apparaîtront, nous tenterons de voir si une explication en termes de profil de locuteur peut être avancée.

En ce qui concerne les disfluences, la littérature fournit une terminologie qui est loin d'être unifiée⁸, pour des phénomènes qui ne le sont pas non plus, tant du point de vue formel que fonctionnel. On a ainsi dans la littérature en français : *phénomènes de performance propres à la langue parlée* (Valli et Véronis 1999), *inattendus structuraux* (Goulian et al. 2002), *trébuchements* (Pallaud 2004), *achoppements* (Pallaud 2003, Henry 2005), *accidents de parole* (Zellner 1992), *défaillances, réparations* (Martinie 1999), *marques du travail de formulation* (Morel et Danon-Boileau 1998), *amorces* (Jeanjean 1984, Pallaud 2002b), *borborygme* (Brès et Gardes-Madray 1991), *bafouillage*, *bégaillage* (Zellner 1992), *bribes* (Blanche-Benvebiste et al. 1979, Cappeau 1998), *répétitions, autocorrections, auto-interruptions* (Coste 1986), *lapsus* (Pallaud 1999), *reprises* (Coste 1986), *ruptures de construction, pauses, télescopage syntaxique* (Fiala et Boutet 1986), *inachèvement* (Grunig), *phénomènes dits d'hésitation* (Candéa 2000a, Candéa et al. 2005), *cafouillage, turbulences, scories, accidents, ratés* (Jeanjean 1984), *ratages* (Brès et Gardes-Madray 1991), *répétitions-hésitations* (Pallaud 2004), *révisions* (Adda-Decker et al. 2004, Bénard 2005), *erreurs*, etc.

Il n'y a donc pas de consensus terminologique, mais une large variété de phénomènes et d'appellations diverses qui se recoupent et se chevauchent parfois. Ainsi par exemple, chez Oostdijk (2003), le terme *hésitation*, qui apparaît pourtant comme le terme générique, couvre seulement (pour l'anglais) les *uh* et *uhm*, les répétitions et les amorces de morphèmes complétées immédiatement. Lorsque celles-ci sont complétées avec décalage, elles entrent dans les auto-corrections.

⁸ Voir la thèse de Candéa (2000a) pour un tour d'horizon de la terminologie des « disfluences » qu'elle prend en compte.

En ce qui nous concerne, dès lors que l'ambiguïté entre le pathologique et le non pathologique est levée, que l'on distingue **dysfluence**-pathologie et **disfluence**, cette dernière caractérisée comme propre à toute production du parler ordinaire⁹ et non comme une transgression (transgression par rapport à la norme, transgression par rapport à l'écrit, voir Pallaud 2004), le terme *disfluence* nous semble pouvoir être envisagé comme relativement neutre.

Il a néanmoins le désavantage d'englober, comme nous venons de le dire, des phénomènes variés, formellement et fonctionnellement très différents¹⁰.

2 MODÉLISATION DES DISFLUENCES

De manière relativement classique désormais, la littérature consacrée au sujet voit la disfluence comme un endroit où le déroulement linéaire de l'énoncé est brisé, parce qu'il y a piétinement en un point de l'axe syntagmatique : « le déroulement syntagmatique est brisé » (Blanche-Benveniste *et al.* 1990).

Shriberg (1994 : 7-9), à la suite notamment de Levelt (1983), a modélisé la séquence disfluente en la décomposant en quatre éléments distincts, qui correspondent à trois régions :

- ***reparandum*** : le *reparandum* (*RM*) est la partie produite par le locuteur qui ne sera pas conservée et sera remplacée ultérieurement au profit du *repair* ;
- ***interrupting point*** : le *point d'interruption* (*IP*) est le moment de l'énoncé qui coïncide avec la fin du *reparandum*. Ce point d'interruption est vide ;
- ***interregnum***¹¹ : l'*interregnum* (*IM*) est la région qui commence à la fin du *reparandum* et s'achève au début du *repair*. L'*interregnum* peut ou non contenir un terme d'édition (*editing term*), c'est-à-dire une pause silencieuse, une pause remplie, etc., ou plusieurs autres tentatives de formulation inachevées également;

⁹ « Que faire de ces accidents [ratés de la communication] ? Les mettre au compte d'une mauvaise maîtrise de la performance, d'un dysfonctionnement de la parole ? Les expliquer en termes de pathologie et, au mieux, les renvoyer à l'orthophonie ? Nous ne le pensons pas. C'est l'absence totale de ratage – mais est-ce possible – qui serait un dysfonctionnement : « Il parle comme un livre ! ». (Bres et Gardes-Madray 1991 : 93)

¹⁰ Guénou (2005, 2006) parle de *disfluences* en faisant des diverses réalisations un traitement unique, sans sous-catégorisation. Selon Duez (2001 : 114), le terme *disfluence* englobe classiquement les lapsus et les erreurs.

¹¹ L'*interregnum* correspond à l'*editing phase* de Levelt.

- **repair** : le *repair* (RR) indique le retour à la « fluence », par la réparation, la correction du *reparandum*.

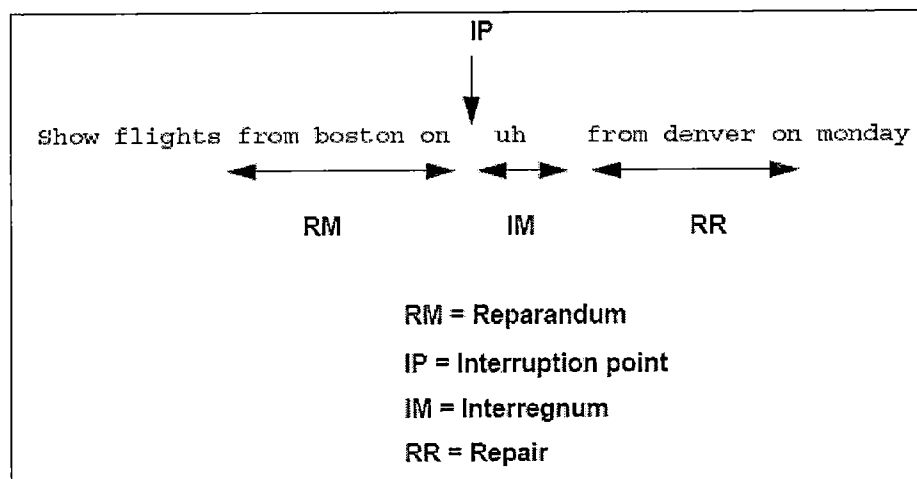


Figure 8. Modélisation de la séquence disfluente (Shriberg 1994 : 8)

Depuis longtemps, on a tenté de montrer certaines régularités qui affectent les disfluences, notamment pour l'anglais (Blankenship et Kay 1964, Cook 1971). L'équipe groupée autour de Claire Blanche-Benveniste, le Groupe Aixois de Recherche en Syntaxe (GARS), s'attache elle aussi, depuis ses débuts, à mettre en évidence les régularités des modes de production de l'oral – y compris les disfluences –, et à les intégrer dans une description unifiée de la langue.

Mais comme le fait remarquer Guénot (2005) :

Toutes ces études, qui décrivent l'organisation interne des disfluences (...) les présentent comme un phénomène unique, avec des caractéristiques régulières (l'entassement paradigmatique, l'absence de fonction syntaxique et de fonction sémantique, les espaces internes), et des caractéristiques plus spécifiques à certains cas (bribes *vs.* amorces, invachèvement *vs.* complétion *vs.* modification, composition de l'interregnum). Cependant elles n'indiquent pas comment l'on différencie une disfluence d'une autre construction, ni comment l'on doit les traiter lors de l'analyse d'un énoncé.

3 LES DISFLUENCES ET LE TAL

Nous l'avons dit, les disfluences constituent un vrai problème dans l'analyse automatisée de l'oral (Adda-Decker *et al.* 2003, Antoine *et al.*

2003, Bénard 2005, Benzitoun 2004, Benzitoun *et al.* 2004, Garside 1995, Guénot 2005, Nivre et Grönqvist 2001, Oostdijk 2003, Valli et Véronis 1999, etc.). En effet, ces particularités du langage parlé réduisent considérablement les performances des outils construits pour l'écrit standard, quand elles ne les rendent pas tout simplement inutilisables. Se pose dès lors de manière cruciale l'étude approfondie des disfluences et autres particularités de l'oral, afin de les décrire et de permettre leur modélisation.

Les descriptions linguistiques, parfois très fines, qui ont été faites des disfluences (cf. notre bibliographie) n'ont, à notre connaissance, jamais connu de formalisation directe dans le cadre de l'étiquetage morphosyntaxique du français parlé¹². La seule tentative en ce sens a été faite par Valli et Véronis (1999).

Dans le cadre de l'analyse syntaxique, on obtient des résultats très médiocres (Oostdijk 2003 pour l'anglais) avec des outils existants.

Guénot¹³ (2005) résume les options prises en analyse syntaxique automatique concernant les disfluences, options qui dépendent étroitement de l'approche et de la tâche considérées¹⁴ :

1. supprimer les disfluences en remplaçant le texte par du « standard » (Oostdijk 2003) ;
2. ignorer les disfluences ;
3. regrouper les disfluences en un groupe (Antoine *et al.* 2003).

Chacune de ces trois options est liée au choix de l'outil d'annotation que l'on compte utiliser. Ainsi, il est envisageable soit d'adapter des outils existants, soit de développer de véritables outils *ad hoc* pour l'analyse automatique de l'oral, selon les options prises dans le traitement des disfluences.

Dans leur expérience, Valli et Véronis (1999) adaptent le logiciel Cordial, développé pour l'écrit, afin d'étiqueter morphosyntaxiquement des textes transcrits de l'oral. Une phase de pré-édition prépare le texte pour

¹² Par exemple, Sandrine Henry étudie les répétitions afin, dit-elle explicitement, de les prendre en compte dans le cadre de l'étiquetage, mais sa thèse (en cours) ne va pas au-delà de l'étape de description et d'établissement d'une typologie.

¹³ Notons que Guénot rassemble sous le terme générique de *disfluences* différents phénomènes qu'elle ne distingue pas les uns des autres et auxquels elle accorde des propriétés semblables. Nous ne pouvons nous contenter ici d'une telle généralisation sous cet hyperonyme.

¹⁴ Voir l'article (Guénot 2005) pour les références illustratives de ces 3 options, qui dépassent le cadre de l'étiquetage morphosyntaxique.

l'analyseur et une phase de post-édition rétablit ensuite le texte original. Les auteurs obtiennent de bons résultats¹⁵ :

les étiquettes fournies par Cordial sont à 97,9 % acceptables, performance égale ou supérieure à la plupart de celles publiées pour l'écrit. (Valli et Véronis 1999 : 126)

Valli et Véronis ont néanmoins des étiquettes erronées, notamment pour les répétitions. Si nous avons décidé de ne pas utiliser Cordial, c'est que ce logiciel fonctionne comme une boîte noire sur laquelle le linguiste n'a pas de prise. De plus, nous souhaitons mettre au point un système qui ne se contente pas d'étiquettes *acceptables*, et qui ne fournit pas des étiquettes erronées (cf. chapitre 6).

Nous verrons dans ce travail comment nous proposons de traiter les disfluences, après les avoir analysées de manière approfondie. En effet, si les études antérieures sont éclairantes à plus d'un titre, elles ne peuvent totalement être satisfaisantes en ce qui concerne notre corpus : les descriptions ont été faites sans lien direct avec un objectif final aussi clairement identifié que le nôtre.

De plus, les études ont été faites sur des corpus qui divergent du nôtre, parce que les conventions de transcription adoptées ne sont pas les mêmes : ainsi, des résultats obtenus sur des études acoustiques fines de la pause, par exemple, ne peuvent être appliqués sans vérification préalable à nos corpus où la notation de la pause est subjective (cf. chap. 4). Par exemple, le *British National Corpus* compte 10 millions de mots étiquetés pour l'oral, mais les transcriptions sont ponctuées (cf. chap. 6).

De plus, force est de constater que sur les disfluences, on manque de données quantitatives complètes, et, à notre connaissance, aucune étude n'aborde l'ensemble des disfluences. L'étude la plus englobante est sans doute celle de Candéa (2000a), mais elle est faite sur un corpus relativement limité et dont les locuteurs sont des enfants en classe de français. C'est en effet que les dénombrements de ce type sont très fastidieux, et que certains sont difficilement automatisables : le repérage du faux-départ n'est guère aisé, et encore faut-il s'entendre sur ce que l'on appelle faux-départ (cf. les problèmes terminologiques que nous évoquions).

Nous verrons au cours de ce travail la position que nous adoptons, qui dépend des disfluences envisagées.

¹⁵ Les auteurs affirment adopter une position « généreuse » dans l'évaluation de leurs résultats.

4 FRÉQUENCE DES DISFLUENCES

4.1 Selon le type d'oral et la situation de communication¹⁶

On l'a dit, les disfluences sont typiques de tout parler ordinaire. Il semble néanmoins que la planification préalable des énoncés par les locuteurs produise des discours dans lesquels les disfluences sont moins nombreuses que lorsque la parole est spontanée. De même, les monologues contiendraient moins de disfluences que les dialogues, où le chevauchement de parole serait un facteur favorisant les disfluences. Ce sont les résultats auxquels aboutit Oostdijk (2003 : 62) lors de la normalisation de disfluences dans un extrait du corpus oral d'anglais britannique ICE-GB, lui-même extrait de l'*International Corpus of English* (Greenbaum 1996). La répartition est la suivante, en fonction des trois catégories de textes mentionnées ci-dessus :

Catégorie de textes	Nombre de mots	Nombre de normalisations effectuées par 100 mots
dialogue	360 000	18,9
monologue, non écrit	140 000	17,4
monologue, écrit	100 000	4,7

Tableau 6. Nombre de normalisations selon la catégorie de textes

La différence la plus nette se situe bien au niveau de la préparation du discours, entre le non-planifié et le planifié, celui-ci correspondant à du « scripted speech », c'est-à-dire de l'écrit oralisé (notamment dans des émissions télévisées, des journaux parlés, etc.). La fréquence des disfluences semble donc bien corrélée au degré de préparation du discours, et à l'interaction entre les locuteurs.

Grosjean et Deschamps ont comparé les variables temporelles du français chez des locuteurs intervenant dans des situations de parole différentes : d'une part des interviews radiophoniques dans lesquelles les sujets s'expriment sur leurs centres d'activités ou d'intérêts (1972) et, d'autre part, des descriptions de dessins humoristiques (1973). Les auteurs

¹⁶ Duez (1991) s'est intéressée à l'utilisation stylistique de la pause silencieuse dans le discours des hommes politiques. Son principal centre d'intérêt est la pause silencieuse, qu'elle envisage dans son interaction avec des facteurs extralinguistiques et notamment le type de parole (discours politique, interview politique ou encore interview amicale). On regrettera néanmoins l'intérêt relativement marginal qu'elle accorde aux marques de disfluence prises en compte (seulement si ces marques sont dans le voisinage direct d'une pause silencieuse).

constatent deux fois plus de répétitions¹⁷ chez des locuteurs se livrant à la tâche de description que lors des interviews. Ils expliquent cette divergence par le fait que dans ce dernier cas, la tâche est moins contraignante pour le locuteur et demande un effort cognitif moindre. Selon eux, c'est le degré de contrainte qui favorise l'émergence de ce qu'ils appellent des *marques d'hésitation*.

De la même manière, ils observent que les pauses sonores (syllabes allongées et *euh*) sont également deux fois plus nombreuses et les faux départs quatre fois plus nombreux dans les descriptions.

Dans le corpus étudié par Adda-Decker *et al.* (2003), composé d'extraits d'entrevues télévisées où une personnalité est interrogée par plusieurs journalistes (une heure de parole), les auteurs notent que les chevauchements de parole favorisent les disfluences. Ces chevauchements se produisent notamment lorsque le journaliste qui mène le débat coupe la parole des personnes interviewées, le discours étant vu par les auteurs comme « une situation de lutte pour la parole ». On regrette néanmoins que cette intuition ne soit pas quantifiée, aucune donnée ne distinguant les disfluences *dans* et *hors* parole superposée.

Le type de corpus sur lequel nous travaillons se situe dans la première catégorie établie par Oostdijk (2003), celle des dialogues favorisant les disfluences.

Si l'on se réfère à Grosjean et Deschamps et que l'on classe nos données selon le degré de contrainte pour le locuteur, celles-ci se répartissent différemment dans les entrevues semi-dirigées : en effet, dans certains sous-corpus, les locuteurs s'expriment sur des sujets qui ne leur sont pas familiers (corpus *acc* (accent) sur la perception de l'accent par exemple), tandis que d'autres sont invités à raconter leurs souvenirs (certes sur un sujet précis, le corpus *cha* (Chalon)). Ainsi, si l'on est bien dans du dialogue, le degré de contrainte est sans doute différent. À l'opposé, on pourrait penser que les locuteurs qui s'expriment dans des conversations amicales ont un discours peu contraint, qui serait donc moins riche en disfluences.

Nous devons de plus introduire un bémol en ce qui concerne les textes des entrevues semi-dirigées, puisque si ceux-ci participent bien d'une conversation (la plupart du temps, effectivement des dialogues), l'écoute de l'intervieweur est une écoute attentive, que l'on pourrait qualifier de bienveillante, et l'échange verbal ne peut être envisagé comme une « lutte

¹⁷ En ce qui concerne la distribution des répétitions selon ce qui est répété (syllabe, mot outil, mot lexical, etc.), les auteurs ne remarquent pas de divergence entre les deux types de discours, la répartition des sous-catégories de séquences étant approximativement la même.

pour la parole », comme c'est le cas dans le corpus étudié par Adda-Decker *et al.* (2003).

Comme le fait remarquer Duez (2001 : 117) :

Le lien entre l'anxiété et les différents phénomènes
d'hésitation paraît extrêmement complexe.

Ainsi, Cook (1969 et 1971, cité par Duez 2001) montre que le nombre d'occurrences de pauses remplies ne semble en aucun cas lié à l'anxiété du locuteur¹⁸.

Nous verrons dans la suite que nous avons nous-même du mal à établir un lien entre fréquences des disfluences et paramètres liés aux locuteurs ou à la situation de parole.

4.2 Selon le type de disfluences

Les différents types de disfluences que nous envisageons obtiennent des fréquences très variables en corpus, mais on observe une certaine forme de régularité.

En effet, dans les études antérieures que nous avons consultées sur le français, on constate une même hiérarchisation des disfluences selon leur fréquence relative. Chez Grosjean et Deschamps (1972), qui étudient les variables temporelles du français sur un échantillon de parole de 30 locuteurs, les *euuh* sont les plus fréquents, viennent ensuite les allongements vocaliques, les répétitions et enfin les faux départs.

La répartition qu'ils obtiennent (p. 150) est la suivante :

Type de pauses	%
Pauses non sonores	59,28
pauses de respiration	(44,60)
pause de non-respiration	(14,68)
Pauses sonores	28,11
pauses remplies	(15,93)
syllabes allongées	(12,19)
Répétitions	6,84
Faux départs	5,77

Grosjean et Deschamps considèrent ici un tout qui comprend également les pauses non sonores, car ils englobent les répétitions et les faux départs

¹⁸ Voir Duez (2001) pour un résumé des études antérieures sur les liens entre anxiété/émotion et disfluence.

dans la catégorie générale des pauses, vues comme des phénomènes marquant l'hésitation. Si l'on reprend les données ci-dessus pour ne garder que les marques qui nous intéressent et permettent la comparaison, nous obtenons le tableau suivant :

marque	<i>eah</i>	allongement vocalique	répétition	faux départ
%	39,12	29,93	16,79	14,16

Tableau 7. Répartition des types de disfluences chez Grosjean et Deschamps (1972)

La hiérarchie obtenue par Maria Candéa (2000a), qui envisage quatre types de disfluences dans sa thèse, est comparable à celle de Grosjean et Deschamps. Son corpus est pourtant très différent, puisqu'elle a analysé des récits en classe de français, auprès d'enfants âgés d'une dizaine d'années. Les résultats qu'elle obtient sont les suivants (pp. 37-38) :

marque	<i>eah</i>	allongement vocalique	répétition	autocorrection
n=	544	284	110	45
%	55,34	28,89	11,19	4,58

Tableau 8. Répartition des types de disfluences chez Candéa (2000a)

Les résultats de Grosjean et Deschamps et ceux de Candéa, s'ils sont distribués selon le même ordre de fréquence, se distinguent en ce qui concerne les pourcentages recueillis par les premières et dernières catégories de disfluences. Alors que Candéa obtient plus de la moitié de disfluences réalisées par un *eah*, inversement, le pourcentage obtenu chez Grosjean et Deschamps pour le faux départ est bien supérieur à celui de Candéa pour l'autocorrection. En effet, les premiers ne parlent pas d'*autocorrection immédiate* mais bien de *faux départ*, concept qui recouvre une gamme plus large de phénomènes, puisque les auteurs définissent le faux départ comme :

tout énoncé interrompu après une syllabe, un mot, un groupe de mots ou une phrase (ex. : « on est allé...il est allé en ville »). (Grosjean et Deschamps 1972 :133)

Alors que Candéa (2000a : 29) définit l'autocorrection immédiate de manière beaucoup plus restrictive :

dans une séquence parlée de type *X Y*, nous considérons que la partie *Y* est une autocorrection immédiate de *X*
- si *Y* est prononcé pour remplacer *X*, et

- si Y corrige *un seul et unique trait phonétique ou morphologique* de X

La même tendance générale de répartition des quatre types de disfluence est également obtenue chez Duez sur la parole d'hommes politiques (1991).

Une hypothèse qu'invalide Candéa, et qui peut être rapprochée du degré de contrainte, est le facteur émotivité. En effet, on pourrait penser que plus un locuteur est stressé, tendu, soumis à une émotion quelconque, plus il produit des disfluences dans son discours¹⁹. Or, l'auteure constate qu'

il n'y a pas de corrélation entre l'émotion croissante du locuteur et la fréquence des marques de travail de formulation. (Candéa 2000a : 84)

Nous ne pouvons guère effectuer de vérification de ce genre sur notre corpus, n'ayant pas un même locuteur soumis à des types de discours différents. Par contre, sur base des enregistrements effectués dans le cadre du projet PFC (Phonologie du Français Contemporain, Hambye 2005), où le locuteur s'exprime dans des conversations guidées et dans des conversations libres, soumises donc à des contraintes différentes, on pourrait valider ou infirmer cette hypothèse.

4.3 Selon les locuteurs

On peut se demander si cette régularité observée dans la distribution des différents types de disfluences est le fait d'une tendance générale du français ou si l'on constate une variation qui dépend des locuteurs. En effet, les chiffres donnés ci-dessus sont des résultats globaux où les locuteurs ne sont pas différenciés.

Comme le mentionnent Grosjean et Deschamps (1972 : 150),

(...) les sujets, qu'ils s'expriment rapidement ou lentement, qu'ils mettent beaucoup ou peu de pauses non sonores, ont tous besoins [sic] de faire un minimum de pauses sonores [c'est-à-dire de répétitions, de faux départs, de *euh* et d'allongement vocaliques].

Une idée reçue sur les disfluences est que les locuteurs professionnels – qui ont une très bonne maîtrise de la langue orale – seraient en quelque sorte à l'abri :

¹⁹ Le prévenu ne sait plus trop. Confesse qu'il avait pris « des médicaments ». S'énervé. Bafouille, bredouille, cafouille. (Joël Matriche, *Le Soir en ligne*, 31 janvier 2007)

La présence de ces variables [secondaires : pauses sonores, répétitions et faux départs] n'est pas automatiquement requise lorsque nous parlons ; elles n'interviennent presque jamais en lecture par exemple, et seulement partiellement chez certains locuteurs qui s'expriment avec beaucoup d'aisance. Ces variables correspondent à ce qu'on nomme « les phénomènes d'hésitation ». (Grosjean et Deschamps 1975 : 148)

Or, on constate, comme le mentionne Pallaud (2004) que

si certains locuteurs en [des disfluences] produisent plus que d'autres, personne n'est épargné (Pallaud 2004 : 83).

Le corpus analysé par Adda-Decker *et al.* (2003)²⁰ se compose de deux types de locuteurs : des journalistes et les personnalités interviewées par ceux-ci. Dans une analyse de correspondance, les auteurs ne constatent aucun regroupement possible, sur base de l'analyse des disfluences. Cependant, les auteurs remarquent que les répétitions sont nettement dominantes chez les journalistes, avec l'hypothèse que celles-ci manifestent chez eux une difficulté d'interrompre l'interviewé. Par ailleurs, certains locuteurs font plus de faux départs (restart) que de répétitions, sans que ce fait puisse être corrélé à un autre facteur.

Nous verrons dans la suite de ce travail ce que l'on peut tirer comme conclusion sur nos données.

4.4 Quatre types de disfluences

Les transcriptions que nous utilisons, nous l'avons dit, contiennent la trace de ces disfluences. Mais celles-ci ne sont pas notées en tant que telles, balisées par exemple par le transcripteur ou par l'expert qui valide la transcription. En fait, la tâche de transcription en elle-même est suffisamment lourde et accaparante pour ne pas encore la complexifier en ajoutant des consignes qui viseraient à noter ces phénomènes (cf. chap. 1).

Dans ce travail, nous regroupons sous le terme de *disfluence* :

- les répétitions ;
- les corrections directes, liées aux répétitions ;
- les amorces de morphèmes ;
- le morphème *euh*.

²⁰ Voir aussi <http://www.limsi.fr/Individu/habert/Projets/ATALA05Copte.html>.

Nous détaillerons ces quatre points dans la suite de ce chapitre. Nous n'aborderons pas ici la pause silencieuse, sauf lorsqu'elle cooccure avec un autre phénomène pris ici en considération. La pause silencieuse, dont on verra qu'elle participe d'une certaine forme de disfluence lorsqu'elle est non démarcative, fera l'objet d'un traitement plus approfondi dans le chapitre suivant, consacré au découpage du texte.

L'étude des quatre types de disfluences que nous allons maintenant aborder vise non seulement à décrire les données sur lesquelles nous travaillons, mais aussi à annoter ces disfluences dans le corpus et évaluer comment les traiter au mieux dans le cadre de l'étiquetage morphosyntaxique. Nous avons donc notre propre procédure d'analyse, liée en partie à l'objectif final de la recherche qui est l'étiquetage automatique. Nous allons nous livrer à une étude linguistique, mais qui se base sur des repérages qu'il est possible de faire de manière informatisée. Nous laisserons donc de côté, par la force des choses, un certain nombre d'indices qui pourraient s'avérer pertinents mais que nous sommes incapable de traiter à ce stade de la recherche.

4.5 Un exemple d'annotation de corpus oraux : le projet COPTE

Dans certains corpus oraux, les disfluences sont annotées manuellement. C'est le cas dans l'action transverse COPTE (Corpus parole / Texte et Évaluation) du département Communication Homme-Machine du LIMSI (Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur). Le projet s'intéresse particulièrement aux phénomènes liés à la parole spontanée. L'un des objectifs principaux de COPTE est de produire automatiquement des transcriptions enrichies, c'est-à-dire des transcriptions accompagnées d'informations sur les tours de parole, les entités nommées ou encore les disfluences (Paroubek 2005). COPTE appelle *disfluences*

tout ce qui n'est pas reconnu dans le langage de modèle
du système de reconnaissance automatique de la parole
du LIMSI (Bénard 2005 : 49).

Les études réalisées au LIMSI visent essentiellement à améliorer le système de reconnaissance de la parole (voir aussi Adda-Decker *et al.* 2003 et 2004, Vasilescu *et al.* 2005, Clerc-Renaud *et al.* 2004).

Le projet COPTE adopte les consignes d'annotation développées par le Linguistic Data Consortium (LDC)²¹, qui considère :

- les « fillers » :

²¹ Voir <http://www ldc.upenn.edu/Projects/MDE/>.

-
- FW : Filler Words, c'est-à-dire les remplisseurs ;
 - DM : Discourse Markers, les marqueurs du discours (*anyway, see, so, actually...*) ;
 - EET : Explicit Editing Terms, la reconnaissance explicite d'une disfluente par le locuteur lui-même (*I mean, sorry...*) ;
 - AS: Asides (apartés) (*I don't know...*) ;
 - PAR : PArentheticals (commentaire sur le même sujet).
- les « edit disfluencies » :
 - RP : RePetitions ;
 - RV : ReVisions (redémarrage partiel) ;
 - RS : ReStarts (la correction remplace ce qui précède ou en modifie le sens) ;
 - les « sentence like unit » (SU), unités assimilables à des phrases, unités de pensée cohérente.

Voici un exemple d'annotation pour COPTE proposé par Bénard (2005) :

(FP euh euh) (RP cert(ains) {truncation}) certains disent qu' (FP euh) , en ce moment - - je pense , par exemple , au député UDF (FP euh) François Daubert - - c'est pas le seul mais (DM enfin) (RV c'est) disons qu'il est très en pointe là-dessus , (RP certains disent qu'en ce moment) (FP euh) , il y a en France une pénétration accélérée de la Mafia

C'est le même type d'annotation que propose Strassel (2003), ou encore Meeter *et al.* (1995).

Dans le projet COPTE, les disfluences sont annotées manuellement. Pour éviter cette étape supplémentaire après la transcription, qui nécessite un codage fait par un expert humain, nous souhaitons nous concentrer sur des phénomènes qu'il est possible de repérer formellement, de manière automatisée.

Ce sont ceux-ci que nous allons maintenant détailler dans les quatre sections suivantes.

CHAPITRE 3. SECTION 2

LES RÉPÉTITIONS

- Ah, j'me rappelle que... que... quand t'étais pas plus haute que... que... que trois pommes ! a bafouillé l'un de ses authentiques cousins, Doogie Wade, et mince, t'étais cacacapable d'embobiner ton monde, déjà ! »

C'était un grand échalas d'une trentaine d'années auquel il manquait deux incisives depuis un certain samedi soir, sans qu'il puisse se rappeler où et comment c'était arrivé, et qui se mettait à bégayer dès qu'il devait former le son « K ».

(Moi, *Charlotte Simmons*, Tom Wolfe, Paris, Robert Laffont, p. 27)

1 QUELS TYPES DE RÉPÉTITIONS ?

En linguistique, le terme *répétition* est relativement englobant et permet de regrouper des phénomènes parfois très diversifiés, qui ne semblent avoir qu'un lien parfois très étroit les uns avec les autres, aussi bien au niveau formel qu'en ce qui concerne leur fonction (syntaxique, sémantique, pragmatique). Les étiquettes sont d'ailleurs multiples : *répétition-hésitation*, *hésitation*, *reprise*, *recommencement*, *réitération*, *séquence réitérée*, *réduplication*, *triplication*, *ressassement*, *ré-énonciation*, etc.

La rhétorique classique, déjà, faisait usage du concept de la répétition, terme générique qui comprenait des figures différentes. Comme le dit Durand (1970) :

La rhétorique classique connaît un grand nombre de figures de répétition, qui se distinguent par la substance de l'élément répété (son, mot, groupe de mots) ou par la position de cet élément dans la chaîne parlée (début ou fin de phrase, etc.) : assonance, rime, allitération, homéotéleute, anaphore, épanaphore, épistrophe, épanalepse, anadiplose, épanadiplose, concaténation, symploque.

Parmi toutes ces figures qui relèvent de la répétition, certaines concernent une identité de contenu, d'autres une identité de forme. Pour éviter une étape d'annotation manuelle, c'est aux répétitions strictement formelles de mots ou de groupes de mots que nous nous intéressons ici, laissant de côté, par exemple, les répétitions telles que celle-ci où le locuteur modifie le temps de son verbe :

blaBD1 et c'est avec qui qu'il est / qu'il serait euh
[blaGD1]

Nous n'envisagerons pas davantage un certain nombre d'occurrences qui participent de ce que Blanche-Benveniste (1985) appelle les répétitions-hésitations, et qui sont liées chez le locuteur à une recherche de dénomination :

blaNB1 ouais // un petit problème de fisc pour le moment j'attends (rires) attention / non c'est / c'est pas vraiment ça c'est / c'est notre comptable / qui qui / qui a fait une euh / une encéphalite à her/ à herpès une encéphalite herpétique (...) [blaNB1]

La locutrice, une médecin, cherche ici le terme exact : elle amorce tout d'abord le morphème *herpès* (*à her/*), le complète (*à herpès*), avant de reprendre une formulation qui lui semble mieux convenir, où *à herpès* apparaît sous sa forme adjectivale *herpétique*, terme relevant sans doute plus du jargon médical. S'il est évident que l'énoncé ci-dessus participe d'une certaine forme de répétition (la locutrice dit exactement la même chose, une fois sous forme nominale, la seconde sous forme adjectivale), nous ne retiendrons pourtant pas ce phénomène parmi ceux abordés ici¹.

De la même manière, nous ne retiendrons pas des formes de répétitions, pourtant très proches de celles que nous allons étudier, dans lesquelles le locuteur corrige l'énoncé pour ajouter une précision, comme l'illustrent les deux exemples suivants :

blaAT1 (...) ils sont pas du tout prêts à suivre des études euh de vraies études euh / plus tard quoi tu vois |- enfin <blaND1> m -| // certains veulent continuer euh c' est l'er/ c'est l'enfer quoi [blaAT1]

blaAT1 (...) quand je repense des fois euh / je me retrouvais place Saint-Lamb/ enfin je sais pas place Saint-Lambert à minuit euh / enfin tu vois où je reprenais le bus euh le dernier bus mais / tu vois une sale ambiance glauque enfin j/ / jamais je jamais enfin

¹ Nous parlons de la recherche de la dénomination. En ce qui concerne l'amorce de morphème (*à herp/*), nous renvoyons le lecteur à la section consacrée à ce sujet.

maintenant j'ai l'impression que j' étais inconsciente quoi
tu vois (...)[blaAT1l]

Dans les deux cas, la locutrice précise en répétant le substantif (et le déterminant, nous y reviendrons plus tard) qu'elle a déjà énoncé et en ajoutant un adjectif qui apporte une précision essentielle. Dans le premier exemple, il ne s'agit pas simplement de suivre des études mais de *vraies* études. Dans le second exemple, la locutrice explique que son inconscience est liée au fait qu'elle reprenait le bus tardivement place Saint-Lambert, et fait aggravant, pas n'importe quel bus : le *dernier* bus.

Si nous ne tenons pas compte d'exemples de ce type, c'est que nous reprenons à notre compte la définition de la répétition proposée par Candéa (2000a : 315) :

Par définition nous considérons que toute répétition forme un bloc dans la parole qui comporte au minimum deux éléments : un premier élément que nous appellerons le « *répétable* » et un deuxième élément, identique au premier, que nous appellerons le « *répété* ». Il va de soi qu'en théorie toute unité produite par la parole est en principe un *répétable* et ce n'est que la présence d'un *répété* immédiatement après qui fait que ce *répétable* va entrer effectivement dans la composition d'un bloc que nous appelons *a posteriori* une « répétition ».

Une répétition est donc une séquence totale qui comprend un répété et un ou plusieurs répétables.

Pour le dire autrement : **répétition = répétable + répété(s)**

La longueur du répétable est variable. Celui-ci peut être constitué soit d'un seul mot, soit de plusieurs mots, au sens typographique du terme. Nous ne descendrons pas en deçà du niveau du mot graphique, comme c'est le cas dans certaines études (Grosjean et Deschamps 1972 par exemple). Nous intégrons les répétitions de syllabes, s'il y a lieu, dans la partie sur les amorces de morphèmes.

Le répétable peut être repris une seule fois ou plusieurs fois. On parlera d'un mot répété une fois (répétition simple), s'il apparaît deux fois (répétable + répété), répété deux fois (répétition multiple) s'il apparaît trois fois (répétable + répété1 + répété2), etc.².

² Danon-Boileau et Morel (1998) font cette distinction, mais appellent notre *répétition simple* une *répétition double*.

Nous verrons dans l'analyse des données ci-dessous comment les répétitions se répartissent dans notre corpus en fonction de ces distinctions qui concernent la longueur / du répétable et le nombre *n* de répétés.

Sous ce schéma formel, se regroupent deux types de répétitions fondamentalement différentes quant à leur(s) fonction(s) langagière(s) : les répétitions « faits de langue » et les répétitions « faits de parole » (Henry 2005)³. Si ces dernières participent bien du travail de formulation propre à la planification de la parole et trouvent donc leur place parmi les disfluences, les premières sont quant à elles inscrites dans la grammaire de la langue, et se rencontrent tant dans les productions orales que dans les productions écrites les plus normées.

2 LES RÉPÉTITIONS « FAITS DE LANGUE »

Dans les répétitions « faits de langue », on distinguera également deux grands groupes : le premier, dans lequel la répétition relève d'une exigence de la syntaxe et est donc due au hasard de l'homonymie, et le second, dans lequel la répétition participe d'un effet oratoire. Nous ne donnons ici que quelques exemples en illustration, qui sont très loin de représenter l'ensemble des manifestations des répétitions « faits de langue ».

2.1 La répétition exigée par la syntaxe

On a en fait ici affaire à des cas de pseudo-répétition : la reprise d'un élément ne relève pas d'un choix du locuteur mais des exigences de la syntaxe française. Il en est ainsi dans la conjugaison des verbes réciproques et réfléchis : pour 1^{res} et 2^{es} personnes du pluriel, la réitération du pronom est prévue par la grammaire. On a ainsi :

ilcDE1 (...) si au contraire **vous vous** trouvez dans une situation où la qualité de la traduction est médiocre (...) [ilcDE1r]

ilpHJ1 (...) et puis alors euh le fait d'habiter Fourons euh nous avons été confrontés au problème très spécifique des Fourons / et euh les les mouvements de jeunesse de de de la région / euh étant donné je vais dire notre / euh no/ notre aptitude à / à organiser des manifestations et autres revendications nous ont demandé euh / si on ne pouvait pas s'occuper aussi de la cause fouronnaise / et **nous nous** sommes occupés

³ Nous gardons cette distinction « faits de langue » / « faits de parole » même si à l'évidence toutes les répétitions ayant un effet oratoire relèvent de la parole. L'opposition doit plutôt être comprise en termes de disfluent/non disfluent.

de la cause fouronnaise / et alors bon ben la politique a
été une suite logique de cette démarche [ilpHJ1r]

On trouve le même cas de figure avec la suite de la préposition *en* et du pronom préverbal *en*, comme dans :

iljPM1 (...) d'autre part on a de plus en pluS tendance
également à intégrer des inventions / des publicitaires
des gens de variétés / ça ce sont deux choses / euh //
et à simplifier euh les les / les tournures / tout **en en**
ajoutant d'autres donc c/ hein c'est un petit peu ça la
synthèse des deux (...) [iljPM1r]

Les répétitions ici relevées n'ont aucun effet oratoire. Elles sont dues aux hasards de l'évolution phonétique et, par exemple pour les verbes réfléchis, elles ne se rencontrent pas aux autres personnes (*je me rase, tu te rends compte, etc.*).

2.2 La répétition à effet oratoire

2.2.1 L'intensité

Parmi les effets oratoires de certaines répétitions⁴, on trouve fréquemment l'interprétation intensive : le locuteur, en réitérant un mot ou un groupe de mots, vise à insister particulièrement sur ceux-ci. C'est le cas des tournures quasi figées telles que *c'est pas joli joli, il est un peu fou fou, tiens tiens !*, etc., qui n'ont en fait rien de tautologique.

Dans l'exemple suivant, la locutrice, une journaliste, nuance le terme de *carrière*, qui pourrait sans doute donner lieu à une interprétation de grande expérience chez elle, en insistant sur le fait qu'elle ne fait que débiter dans le journalisme (les adjectifs *jeune* et *petite*) : l'adverbe *très* apparaît trois fois devant chacun de ces adjectifs ; la journaliste n'exerce pas depuis longtemps :

iljDV1 (...) quand je devais prendre la parole à un cours
pour présenter un travail / ben un petit vent de panique
quand même / euh me touchait / mais euh / au fur et à
mesure non parce que / au fur et à mesure / que que
j'avance dans dans ma carrière encore **très très très**
jeune **très très / très** petite euh et bien euh / non on a
l'habitude maintenant de de parler euh devant un micro
de parler devant des gens / donc ça m'effraie vraiment
plus du tout / plus du tout du tout du tout [iljDV1r]

⁴ Nous sommes donc ici plus dans des effets de style que dans la langue au sens strict du terme.

Toujours dans l'exemple précédent, le syntagme entier *plus du tout* est répété deux fois⁵, avec un effet d'insistance indéniable. On n'est pas ici dans la disfluence ou, pour le dire autrement, le bafouillage. Par la répétition, la locutrice est catégorique : elle n'est vraiment plus effrayée.

Sur le même principe, on trouve, toujours chez cette locutrice – qui semble affectionner particulièrement la tournure de style –, la répétition intensive de l'adverbe *tout le temps*, également répété deux fois.

iljDV1 (...) (silence) même en gardant beaucoup de rigueur on peut avoir un langage / très cool / encore une fois (rire) très (rire) essayer de trouver un synonyme en français / hein // (bruits) parce que ça ne sert à rien d'être austère **tout le temps tout le temps tout le temps** je pense qu' on doit aussi parler on doit on doit // on doit parler euh en TV et en radio comme on parle dans la vie de tous les jours mais sans sans exagérer (silence) [iljDV1r]

Comme dans les cas évidents des pseudo-répétitions ci-dessus, la reprise fait progresser le discours sur l'axe syntagmatique. En effet, à la différence des répétitions « faits de parole » que nous envisagerons plus loin, les répétitions présentées ici sont « envisagées comme enchainées et capitalisées sur la chaîne syntagmatique » (Constantin de Chanay 2005 : 1). La linéarité du discours n'est aucunement brisée, le sens se déduisant précisément de l'accumulation des sens du répétable et du (ou des) répété(s).

⁵ Selon la terminologie que nous adoptons, une séquence répétée 2 fois apparaît donc 3 fois dans l'énoncé.

2.2.2 La gradation

L'exemple suivant, tiré d'Hergé et repris à Constantin de Chanay (2005), illustre à merveille cette progression sur l'axe syntagmatique.

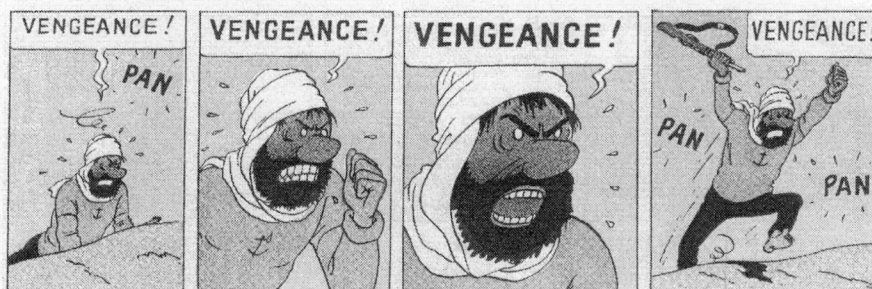


Figure 9. Hergé, *Le Crabe aux pinces d'or*, p. 37, 7-9

On le voit, la répétition du terme *vengeance* dans trois vignettes différentes contigües marque très clairement la progression de l'énoncé sur l'axe syntagmatique. Cette progression s'accompagne non seulement d'un jeu sur la typographie, par l'utilisation de lettres progressivement plus grandes et plus grasses, mais également d'une gradation dans le dessin.

2.2.3 La modification du sens

Dans le cas suivant, la répétition n'est tautologique qu'en apparence.

chaGG0 le lieu de naissance de vos parents / c'était
chaBPA1 euh c'était Seraing | - Seraing ouais
 <chaGG0> pour les deux ? - | pour les deux oui
chaGG0 pour les deux [chaBPA1r]

En effet, la répétition note la progression de la pensée du locuteur, le sens tel qu'il est en train de se construire. Alors que le répétable correspond à la première réponse donnée par le locuteur à la question de l'interviewer, après une pause pleine (*euh*), le locuteur confirme ensuite son assertion grâce au répété, accompagné d'un *ouais* ajoutant à la confirmation déjà apportée par la répétition.

On n'a donc pas affaire ici aux deux mêmes sens de *Seraing* : le répété donne une information sémantique différente que celle apportée par le répétable ; il confirme l'assertion avancée par le répétable.

En fait, ici, le locuteur semble évaluer, de manière positive, la dénomination qu'il est en train de construire (cf. Blanche-Benveniste 1985 : 122). C'est à ce processus de construction de la dénomination que ce type de

répétition nous permet d'accéder. Si les transcriptions étaient ponctuées, on aurait sans doute adopté le format suivant :

Euh, c'était à Seraing. Seraing, oui.

La modification de sens est encore plus nette dans l'exemple suivant, relativement fréquent à l'oral :

magPM1 avec de la salade / ça remet les idées en place après le repas // et puis alors un gâteau mais un **gâteau gâteau** pas un gâteau de glace [magPM1r]

Ici, tout le contexte de l'énoncé permet d'interpréter cette répétition comme une modification de sens de *gâteau*, modification de sens qui consiste en une restriction sémantique du champ d'application du mot *gâteau*. Pour la locutrice, il s'agit ici véritablement d'un gâteau (*un gâteau gâteau*) au sens strict, par opposition à ce que d'autres considèrent également comme faisant partie de la catégorie générale des gâteaux : le gâteau de glace.

2.2.4 La mise en évidence

Les exemples repris dans cette section participent d'une mise en évidence, même si la répétition relève, comme dans le cas des pseudo-répétitions, d'une contrainte de la syntaxe. Il s'agit des répétitions de pronom par dislocation, comme dans les exemples suivants :

Ne venez pas nous parler comme si nous étions une tache de boue sur vos chaussures ! Des générations de sorciers au sang pur, tous sans exception – je suis sûr que **vous, vous** ne pouvez en dire autant ! (*Harry Potter et le Prince de Sang-Mêlé*, J.K. Rowling, p. 232)

Bien entendu, **vous, vous** êtes catholique, n'est-ce pas ? (*Le Choix de Sophie*, William Styron : 414)

Et **vous, vous** me faites l'effet d'une de ces innombrables crapules de cette ville répugnante qui vous a engendrés, vous et vos pareils au langage ordurier ! déclama-t-il, prompt comme l'éclair à se lancer dans l'éternelle rhétorique de ses ancêtres. (*Le Choix de Sophie*, William Styron : 428)

Ces citations sont empruntées à l'écrit (mais à l'intérieur de dialogues), afin d'avoir des exemples clairement non ambigus. Il est intéressant de noter que le rôle démarcatif du pronom disloqué, déjà séparé du pronom sujet par

la virgule, est encore accentué par le marquage typographique⁶ : l'italique est présent dans les exemples relevés chez ces deux auteurs, publiés chez des éditeurs différents.

3 LES RÉPÉTITIONS « FAITS DE PAROLE »

Les répétitions sur lesquelles nous allons nous pencher maintenant ne se rencontrent qu'à l'oral, ou à l'écrit lorsque celui-ci tente d'imiter l'oral. Elles relèvent de la performance et participent des achoppements propres aux productions orales : à la différence des répétitions « faits de langue » pour lesquelles le discours poursuit son déroulement sur l'axe syntagmatique, on constate que les répétitions « faits de parole » « nous oblige[nt] en fait à “piétiner” sur le même emplacement syntaxique » (Blanche-Benveniste 1985 : 113) ; la linéarité du discours est interrompue et un entassement se fait sur un même point de l'axe syntagmatique. Voici un exemple typique de répétition « fait de parole » :

blaAD0 ça va il a l'air **de de** te reconnaître [blaNB1l]

Pour marquer visuellement cette stagnation en un point du déroulement de l'énoncé, la représentation en grilles (Blanche-Benveniste *et al.* 1979, Bilger 1982) adoptée par l'équipe du GARS est particulièrement éclairante :

blaAD0 ça va il a l'air de

de te reconnaître

Le GARS utilise la configuration en grilles notamment pour pallier les effets que produit sur l'interprétation des productions orales la linéarité de l'écrit traditionnel. Les grilles marquent visuellement le déroulement syntagmatique, représenté sur l'axe horizontal, et le déroulement paradigmatique, noté sur un axe vertical. La représentation en grilles note de manière superposée des segments qui occupent une même place syntaxique, sans que ceux-ci constituent nécessairement une répétition au sens où nous l'entendons ici (c'est-à-dire pas uniquement quand il y a piétinement). Ainsi, le déroulement paradigmatique qu'on rencontre, par exemple, dans une suite d'appositions ou dans une énumération sera noté dans une liste, à la verticale.

Le type de répétition de l'exemple ci-dessus n'a pas de fonction sémantique : le sens de l'énoncé ne change pas à cause de la répétition⁷ ; il

⁶ Nous soulignons en gras.

⁷ Nous entendons par là que l'accumulation ne change pas le sens linguistique de l'énoncé. On peut attribuer du sens à la répétition, en l'analysant comme c'est souvent le cas,

n'est pas inscrit dans la syntaxe de la langue. Il est donc imprévisible, même si nous allons voir qu'il obéit à certaines régularités.

4 UN PARTAGE DIFFICILE ENTRE DISFLUENCE ET FAIT DE LANGUE

Les exemples donnés jusqu'ici sont relativement clairs quant à leur statut : inscrits dans la grammaire de la langue, ayant un effet oratoire ou relevant d'un achoppement dans la linéarité de la parole. Dans notre corpus, de nombreux cas ne peuvent être catégorisés avec autant de certitude.

Reprenons le cas de la répétition de la forme *nous*, déjà abordé ci-dessus. Deux cas de figure peuvent se présenter :

- la répétition est disfluente. Dès lors que l'on n'est pas dans l'environnement d'un verbe potentiellement réciproque ou réfléchi, la reprise du pronom *nous* est clairement disfluente lorsque le verbe dans le contexte droit n'est pas conjugué à la 1^{re} personne du pluriel. Dans l'exemple suivant, c'est bien le pronom complément qui est répété ; le phénomène se produit deux fois à trois mots d'intervalle.

ilcDA1 et nous avons souvent des ennuis quand on se trompe -| et l'informatique **nous nous** conduit parfois à **nous nous** tromper [ilcDA1r]

- la répétition n'est pas disfluente et relève d'une répétition « fait de langue ». On est alors dans le cas des exemples cités plus haut, que nous avons repris de l'écrit. Voici un exemple de notre corpus, dans lequel le premier *nous* appartient au syntagme *chez nous*, où il est complément, alors que le deuxième *nous* qui le suit directement est le sujet du verbe *étions*.

accPH1 plus // par exemple chez **nous // nous** étions à beaucoup // on parlait (silence) wallon [accPH1]

Dans l'exemple suivant, le contexte de la répétition nous permet d'interpréter celle-ci comme une répétition intensive, dans laquelle le premier *nous* a un rôle démarcatif :

iljPF1 mais / oui bè par la force des choses puisque je vous dis que ma belle-soeur est professeur à l'université de Lausanne et qu'elle est linguiste / elle a elle-même un accent mais elle ne s'en cache absolument pas et beaucoup pluS que moi proportionnellement elle a l'accent suisse / mais il nous arrive régulièrement d'avoir

comme l'indice d'une hésitation, d'un trouble du locuteur, d'une difficulté d'encodage, etc. (cf. chap. 3.1).

des conversations sur le sujet / d'autant qu'elle le fait d'une manière scientifique et elle se lance dans des explications à n'en plus finir / sur pourquoi dites-vous ceci **nous nous** disons cela il faudrait dire ceci voilà c'est évidemment une relation je dirais presque (dans un rire) professionnelle |- puisque <iljGF0> mm -| / moi étant journaliste et elle linguiste évidemment [iljPF1r]

Mais dans la plupart des cas rencontrés, il est impossible de trancher. Ainsi, l'exemple suivant illustre bien cette difficulté, où il est impossible de savoir, à la seule lecture de la transcription, si l'on est devant une pseudo-répétition liée à l'organisation syntaxique de l'énoncé ou à une répétition disfluente :

accCP1 on n'a pas suffisamment de recul et comme **nous nous** n'avons pas / vécu / étant trop jeunes nous n'avons pas vécu la transition [accCP1r]

Orthographiée et ponctuée selon les règles de l'écrit, on aurait deux versions différentes selon que la répétition est ou non disfluente :

répétition liée à la syntaxe :

On n'a pas suffisamment de recul et comme **nous, nous** n'avons pas vécu – étant trop jeunes –, nous n'avons pas vécu la transition.

répétition disfluente, effacée dans l'écrit standard :

On n'a pas suffisamment de recul et comme **nous** n'avons pas vécu – étant trop jeunes –, nous n'avons pas vécu la transition.

On se trouve dans la même situation d'indétermination avec les répétitions de *ça*. Même devant des exemples de répétition multiple, le problème reste entier :

accFJ1 je sais bien qu'il y en a même qui exagèrent dans le Midi parce que **ça ça ça** fait bien / ils dénasalisent à tour de bras ils ajoutent des E / euh des (x) tant qu'on en veut mais enfin (rire) (...) [accFJ1r]

Si l'on est obligatoirement devant une répétition disfluente, deux possibilités subsistent néanmoins. S'agit-il d'un premier répétable *ça* et de deux répétés disfluents (une vraie répétition multiple), ou d'un premier *ça* disloqué, dont la fonction est celle de complément d'objet, suivi d'un répétable sujet du verbe *fait* et d'un seul répété (cas pour lequel on aurait une

répétition simple) ? Les deux analyses donnent les deux configurations en grilles suivantes⁸ :

accFJ1 je sais bien qu'il y en a même qui exagèrent
dans le Midi parce que **ça**

ça

ça fait bien / ils dénasalisent à
tour de bras ils ajoutent des E /
euh des (x) tant qu' on en veut
mais enfin (rire) (...) [accFJ1r]

accFJ1 je sais bien qu'il y en a même qui exagèrent
dans le Midi parce que **ça ça**

ça fait bien / ils dénasalisent
à tour de bras ils ajoutent des
E / euh des (x) tant qu' on en
veut mais enfin (rire) (...) [accFJ1r]

Prenons encore l'exemple suivant de répétition formelle de *ça*.

blaGD1 j'ai dit oui mais **ça ça** dépend [blaGD1l]

Les indices prosodiques sont ici très clairs. À l'écoute de l'enregistrement, on ne peut interpréter cette répétition que comme une répétition intensive. Pourtant, rien dans la transcription ne le laisse présager (on n'a, par exemple, aucune pause notée, comme on pourrait éventuellement s'y attendre entre le *ça* disloqué à gauche et le *ça* sujet de *dépend*).

Selon Henry (2005 : 83), la prosodie permet de trancher dans ces différents cas, pour lesquels les analyses morphosyntaxiques divergent. L'auteure ne semble pas suggérer que des ambiguïtés puissent subsister, même après le recours à la bande son. En ce qui nous concerne, nous n'avons pas eu recours aux enregistrements pour faire le départ entre les deux types de répétitions, et nous les considérons donc a priori dans une même analyse.

En fait, l'impossibilité de trancher dépasse le cadre de la lecture de l'oral transcrit. En effet, disfluences et figures de style sont parfois très proches. Entre les deux : la notion d'intention du locuteur.

Martinie (2001) remet en cause cette notion d'intention, se basant sur l'exemple suivant issu de l'oral :

⁸ La mise en grille n'est pas une « simple » transcription mais est déjà une analyse grammaticale, qui fait ressortir la structuration syntaxique de l'énoncé (cf. Blanche-Benveniste *et al.* 1990 : 214 et sv.).

j'achète des rasoirs – des rasoirs et du savon
(Clochards)

L'auteur se demande :

Ainsi, ce qui sous la plume d'un auteur illustre recevait le terme d'« anadiplose » sera peut-être plus facilement perçu comme une autocorrection dans la bouche du locuteur [de l'exemple cité], un clochard de la région parisienne, mais sur quelles bases objectives trancher ? Et même, à la rigueur, pourquoi trancher ? (Martinie 2001, 190-191)

On rencontre ce type de répétitions dans notre corpus, comme l'illustre l'énoncé suivant où cela se produit à deux reprises :

chaHI1 ah dans le temps oui hein / dans le temps
il y avait plus de monde que on a maintenant /
maintenant on fait les enterrements en voiture / pour
aller plus vite / et liquider sur deux heures (...) [chaHI1r]

Une transcription ponctuée, aux normes de l'écrit standard, prendrait la forme suivante, si l'on jugeait qu'il s'agit chez cette locutrice âgée et peu scolarisée d'un effet de style et non de répétitions inutiles voire disfluentes, relevant de sa mauvaise maîtrise du français :

chaHI1 Ah, dans le temps... Dans le temps, il y avait
plus de monde que maintenant. Maintenant, on fait les
enterrements en voiture, pour aller plus vite et liquider
sur deux heures (...)

Dans la suite de ce travail, nous évacuons cette notion d'intention du locuteur non seulement parce qu'elle échappe à nos instruments, mais parce qu'elle implique un jugement. Nous faisons donc un traitement analogue pour tous les types de répétitions.

5 LES RÉPÉTITIONS DANS LE CORPUS

5.1 Le repérage des répétitions dans le corpus

5.1.1 Méthode de repérage

Le repérage des répétitions dans le corpus s'est fait de manière totalement automatisée, en recherchant par programme un mot (ou une suite de mots, toujours au sens typographique du terme) répété immédiatement dans son contexte droit. La longueur du répétable (*l*), en termes de nombre de mots, n'a pas été prédéfinie, de même que le nombre de répétés (*n*).

Il arrive que la linéarité des paroles soit brisée dans la transcription. Ainsi, les choix de représentation des chevauchements dans le texte écrit matérialisent certaines répétitions de façon non contiguë. C'est le cas des chevauchements internes au tour de parole, comme le montre l'exemple suivant :

accCT0 ah bon |- **ça c'est** <accFJ1> hein |- **c'est** une bonne méthode [accFJ1r]

Même si une portion de ses paroles a été chevauchée par accFJ1 (ce qui permet d'avancer l'hypothèse que la répétition serait justement due à ce chevauchement de parole, cf. chap. 3.1), accCT0 a bien dit de façon continue et linéaire :

accCT0 ah bon ça **c'est c'est** une bonne [accFJ1r]

La disposition graphique découlant des conventions de transcription adoptées par VALIBEL n'empêche pas le programme de repérer ces répétitions.

Néanmoins, il est des répétitions qui ne suivent pas le schéma formel d'une contiguïté stricte entre répétable et répété dans la chaîne sonore. En effet, il arrive que le répété soit séparé du répétable par l'insertion d'un élément, de nature et/ou de fonction variées. On parlera de *répétition associée* quand un élément apparaît entre le répétable et le répété, par opposition à la *répétition directe* où répétable et répété sont contigus (Henry 2001). Les exemples suivants illustrent cette configuration de répétition associée :

ilrCP1 ça sent c'est un accent **assez euh assez** fort quoi / alors bon euh / l'accent belge bon c'est un accent comme un autre quoi [ilrCP1r]

ilrTC1 parce qu'il y a des il y a **des comment des** interviews où c'est [ilrTC1r]

Afin de voir quelles étaient les séquences possiblement présentes entre le répétable et le répété, nous avons extrait les segments insérés, de maximum quatre mots, entre deux mots ou deux suites de mots⁹ identiques.

⁹ Une telle méthode génère nettement plus de bruit que de résultats corrects. En effet, dans la phrase suivante, *durs* et *vulgaires* sont insérés entre deux séquences *ou plus* :

accCT0 (...) est-ce qu'on pourrait faire une distinction entre des accents que tu considérerais je ne sais pas moi / plus chantants ou au contraire plus plats / euh / **ou plus durs ou plus** vulgaires ou plus distingués plus sympathiques ou au contraire plus antipathiques (...) [accCT0, accCP1r]

Nous avons ainsi identifié 21 marques principales¹⁰ insérées entre le répétable et le répété, ou entre deux répétés : *ah, bè, ben, boh, bon, comment, disons, enfin, euh, ha, hein, hum, m, mh, mm*¹¹, *non, oh, oui, ouais, pf, pff*. À cette liste s'ajoutent encore l'amorce de morphème, la pause brève (/), la pause longue (/ /) et le silence (*silence*), ainsi que les indications métalinguistiques (*rire*), (*rires*), (*toux*), (*soupir*), (*x*), (*xx*) et (*xxx*).

Nous excluons de cette recherche automatique les répétitions qui répondent pourtant à notre critère formel mais qui sont trans-tour de parole, qu'il s'agisse d'une répétition qui est le fait d'un même locuteur (comme tous les cas que nous prenons en considération ici), soit de deux locuteurs différents. Les deux exemples suivants illustrent ces deux cas de figure :

blaAT1 (...) maintenant je suis moi j'ai l'impression
même si je changerai encore mais là j'étais vraiment **en**
pleine euh

blaND1 mm

blaAT1 en pleine transformation et en pleine
recherche tu vois de / de ce que je voulais être (...) [blaAT1]

accCP1 (...) à côté de ça il y a des Bruxellois qui ont un
un petit accent léger |- ils ne me dérangent pas /
<accCT0> **un petit accent** / local -| les Namurois ne
me dérangent pas les / les Liégeois ne me dérangent
pas (...) [accCP1r]

Dans le premier exemple, l'intervention coopérative de blaND1 qui consiste en un simple *mm* est mise dans un tour de parole à part entière parce que ne se superposant pas à la parole de blaAT1 (cf. chap. 1, § 7.3). blaAT1 reprend la parole en répétant ce qu'elle avait dit juste auparavant, et qui avait été en quelque sorte coupé (le syntagme nominal entamé et interrompu après le *euh*). La répétition est ici répartie sur deux tours de parole d'un seul locuteur. On voit ce qu'impliquent les conventions de transcription : si l'on avait décidé de faire des brèves interventions coopératives de l'interlocuteur des chevauchements internes (que ces interventions soient effectivement de

S'il s'agit effectivement bien formellement d'une répétition associée, celle-ci ne pose aucun problème à l'analyse automatique et ne nous intéresse aucunement ici.

Bear *et al.* (1992) recensent aussi de fausses structures selon le schéma *R insertion R* comme *flights daily flights* dans *show me flights daily flights*.

¹⁰ On avait aussi comme segments insérés : 6 occurrences de *je dirais*, 4 de *j'allais dire*, 2 de *comment ça s'appelle* et de *comment vais-je dire*, une de *comment on appelle maintenant*, *comment on dit*, *comment dirais-je* et *par exemple*.

¹¹ Ce *mm* pourrait déjà lui-même être considéré comme une répétition.

la parole superposée ou non), ces répétitions auraient été prises en considération ici.

Dans le deuxième exemple, la répétition est le fait de l'interlocuteur. Selon Mary-Annick Morel¹², ce type de répétition où un locuteur reprend les paroles de celui qui vient de parler est le signe d'un échange langagier « bien réussi ». Ici, l'intervieweuse accCT0, dans le tour de parole de accCP1, reprend textuellement ce qui vient d'être dit, comme pour l'encourager à poursuivre. Ce type de répétition ne sera pas non plus repris ici, puisque ces deux séquences seront envisagées séparément, appartenant à deux tours de paroles attribués comme il se doit à deux locuteurs différents¹³.

5.1.2 Conséquences de la méthode

Cette méthode de reconnaissance automatique a évidemment ses lacunes. En effet, certaines suites ainsi reconnues comme des répétitions ne le sont pas en termes d'analyse syntaxique. C'est le cas dans les deux exemples suivants, où le hasard de la cooccurrence fait reconnaître des séquences de mots identiques qui se suivent :

norKJ1 (...) je crois / que nous aurions sur le plan // sociolinguistique avantage à décrire // toutes les performances // euh / allant du dialecte / au français le français le plus no/ normé le plus proche de la norme (...) [norKJ1r]

magFM0 oui

magPM1 bon ce sera un vin rouge / de la Côte-de-Bourg

magFM0 mm

magPM1 comme apéritif un apéritif un mousseux à l'ananas // or je pense / on n'est pas encore tout à fait décidé pour le potage mais ce sera un potage aux asperges (...) [magPM1r]

Ces cas sont très proches de ceux vus plus haut des répétitions syntagmatiques, à la différence qu'on reconnaît ici de faux syntagmes, alors que plus haut on avait des syntagmes « correctement » constitués, répétés par nécessité syntaxique ou pour divers effets oratoires. Ici, la répétition isole

¹² Communication orale au colloque « Représentations du sens linguistique », Université Libre de Bruxelles, 3 novembre 2005.

¹³ À la différence du premier cas mentionné, l'analyse faite ne dépend pas des choix de transcription.

Nous ne reprenons pas le concept de « locuteur collectif » développé au GARS (Loufrani 1984).

erronément les séquences *français le* et *apéritif un* dans lesquelles les deux termes repérés ne sont pas syntaxiquement liés.

Notons d'emblée que ces phénomènes sont peu fréquents relativement au nombre d'occurrences de répétitions au sein du corpus et n'invalident donc pas les données chiffrées que nous allons présenter. De plus, si le repérage, tel que nous le faisons ici, reconnaît de mauvaises suites et est inévitablement source d'erreurs pour l'analyse syntaxique, il n'en est pas de même dans le cas de l'étiquetage morphosyntaxique. En effet, l'étiquetage qui consiste à assigner un lemme et une catégorie grammaticale ne sera pas remis en question par ces faux repérages. En effet, après balisage de la répétition (cf chap. 4), le texte devient :

magPM1 comme <r> apéritif un </r> apéritif un
mousseux à l'ananas // or je pense / on n'est pas encore
tout à fait décidé pour le potage mais ce sera un potage
aux asperges (...) [magPM1r]

Les séquences balisées seront ignorées des grammaires de désambiguïsation, et seule le dernier répété sera analysé. L'analyse donnée du dernier répété sera ensuite appliquée au répétable balisé, et on voit ici qu'il s'agit du même lemme et de la même classe grammaticale, bien que les termes appartiennent à deux syntagmes distincts.

Une deuxième faiblesse de l'automatisation est que, nous basant sur une identité formelle stricte du répétable et du répété, nous passons à côté de phénomènes de répétitions très proches de ceux qui nous intéressent, mais pour lesquels le répété n'est que partiellement identique au répétable. C'est le cas des exemples que nous avons donnés au début de cette section, où le verbe est repris avec un changement de temps, ou encore où la locutrice médecin se livre à une recherche de la dénomination exacte (cf. § 1). On pourrait multiplier les exemples de variations minimales, mais non reconnues par notre méthode se limitant à une identité formelle et à une stricte contiguïté (moyennant la possibilité d'insertions limitées).

Néanmoins, nous analyserons certaines variations dans les répétitions de mots grammaticaux, dans la section 3 de ce chapitre consacrée aux autocorrections.

La conséquence principale de l'automatisation de la reconnaissance des séquences répétées est que notre collecte regroupe, sous le terme général de *répétition*, des identités formelles de séquences qui sont aussi bien exigées par la syntaxe ou voulues par le locuteur pour des effets oratoires que des disfluences.

La particularité de notre démarche est que nous ne faisons donc pas le départ entre les répétitions disfluentes et celles qui ne le sont pas, à la différence de tous les auteurs consultés qui se sont intéressés à ce sujet. Ainsi, Candéa (2000a) repère manuellement les répétitions disfluentes, mais

sur un corpus de taille très réduite ; le nombre de répétitions analysé par Henry *et al.* (2004) est lui aussi limité ; Constantin de Chanay (2005) fait la distinction entre les deux types de répétitions, mais en ne sélectionnant que certains exemples ; dans sa thèse en cours, Henry (2001) a repéré automatiquement toutes les répétitions formelles sur un corpus d'un million de mots, pour manuellement isoler celles qui relèvent de faits de langue et les exclure de son analyse ; dans le projet COPTÉ (Bénard 2005), les disfluences sont annotées préalablement en tant que telles dans le texte ; Shriberg (1994) a annoté également manuellement 5000 disfluences dans un corpus ; etc.

En fait, deux raisons nous poussent à ne pas distinguer répétitions disfluentes et non disfluentes. Nous souhaitons construire un système d'étiquetage qui ne soit pas tributaire d'une première annotation faite sur le corpus, annotation qui consisterait en une première analyse linguistique de celui-ci par un expert humain¹⁴. Nous partons des transcriptions établies par le Centre de recherche VALIBEL, sans modifier celles-ci ni par l'ajout de nouvelles conventions lors de la transcription (on marquerait les répétitions disfluentes par une balise par exemple), ou ultérieure à cette étape de transcription. Nous avons d'ailleurs vu que dans de nombreux cas, le partage entre les deux types de répétitions, disfluente ou non, est difficile à faire, même en recourant au son.

C'est donc une raison pragmatique qui nous pousse à agir ainsi, mais pas uniquement. En effet, nous ne faisons pas le départ entre ce qui relève de la performance d'une part et de la compétence d'autre part, afin d'intégrer le tout dans une description unifiée.

Certains auteurs (Jeanjean 1984, qui analyse les « ratés » de la communication ; Blanche-Benveniste 1985) voient dans le phénomène de la répétition (disfluent ou intentionnel) un phénomène régulier. C'est donc la régularité de ce procédé général que nous voudrions étudier ici, qu'il relève de la disfluence ou non.

5.2 Résultats globaux : répartition des répétitions dans le corpus

5.2.1 Fréquence générale

Le repérage automatique tel que décrit ci-dessus nous a permis de récolter 12 192 séquences répétitions. Comme nous l'avons dit, celles-ci

¹⁴ Mais on pourrait imaginer que cette première annotation pour repérer les disfluences se base sur une préanalyse automatisée, comme on pourrait le concevoir éventuellement pour repérer un verbe pronominal conjugué à la 1^{re} personne du pluriel et précédé de la séquence *nous nous*.

peuvent être aussi bien des répétitions de langue que des répétitions de parole, ces dernières étant nettement majoritaires dans notre corpus.

Si l'on considère le phénomène de façon générale, sur l'ensemble du corpus, on a 3,19 répétitions tous les 100 mots. Le calcul a été effectué en divisant le nombre de mots du texte par le nombre de mots de la séquence répétition dont on a soustrait le nombre de mots du répétable. On a ainsi considéré que le dernier répété fait partie du texte et non de la séquence disfluente, suivant en cela le schéma proposé par Schriberg (1994)¹⁵ (voir chap. 3.1).

Henry (2005) obtient environ 17 répétitions tous les 1000 mots, soit une répétition toutes les 23 secondes. L'auteure ne donne pas d'indication sur la manière dont elle a effectué le comptage. On ne sait donc pas si chaque mot compte pour un ou si c'est la séquence formée par la répétition qu'elle considère comme une unité, par exemple. On est devant la même incertitude pour le décompte effectué par Henry *et al.* (2004).

Chez Candéa (2000a), la répétition est la troisième marque la plus fréquente (après les *euh* et les allongements).

En fait, nos résultats montrent une fréquence du phénomène beaucoup plus élevée que chez Henry *et al.* (2004) ou que chez Candéa (2000a), principalement à cause du fait que nous avons un grand nombre de répétitions qui ne sont sans doute pas prises en compte chez ces auteurs (outre les répétitions non disfluents, les formes répétées de *oui* et de *non* par exemple, dont on verra ci-dessous la fréquence particulièrement élevée dans notre corpus).

5.2.2 La répartition par locuteur¹⁶

Comme la plupart des phénomènes relevant de la disfluence, la répétition connaît une grande variation inter-locuteurs (Pallaud 2004).

Dans notre corpus, la répartition de la fréquence par locuteur¹⁷ va de 10,18 séquences répétitions tous les 100 mots à 0,10 séquence tous les 100 mots pour le locuteur répétant le moins souvent. La moyenne observée sur l'ensemble des locuteurs du corpus est de 3,19 %.

¹⁵ Compter également le dernier répété dans la séquence répétition pour la division aurait augmenté la fréquence du phénomène.

¹⁶ Le détail des données pour chaque locuteur est présenté dans l'annexe 5.

¹⁷ Nous avons retiré les données non significatives de ces résultats (notamment lorsque le locuteur comptabilise peu de mots).

La locutrice « la plus répétitive » – faisant une séquence répétition en moyenne tous les 10 mots – est l'intervieweuse de l'entrevue norHJ1r, la plus longue des entrevues du corpus. Quand on y regarde de près, on se rend compte que beaucoup des occurrences qui la concernent sont en fait des répétitions visant à encourager son interlocuteur, dans une intervention coopérative comme la suivante :

norHJ1 et j'ai renoncé |- (rire) <norFA0> ça ne m'étonne pas -| donc j'ai renoncé à // à chercher euh vraiment à à me fond/ à me rallier à la prononciation non et ça c'est un de mes souvenirs enfin / et là j'ai été j'ai été déçu

norFA0 bien sûr bien sûr

norHJ1 je ne sais pas si vous avez déjà fait u/ une expérience de ce genre mais enfin là j'ai |- été frappé [norHJ1r]

Néanmoins, toutes les répétitions sont loin d'être de ce type, puisque l'on a aussi des répétitions relevant clairement de la disfluente, comme dans :

norFA0 oui et si je vous demandais de / de me définir un peu la notion d'accent régional // comment est-ce que vous définiriez ça [norHJ1r]

De plus, dans les autres entrevues, la locutrice obtient des scores beaucoup plus faibles, parfois même en dessous de la moyenne relevée sur l'ensemble du corpus (2,8 % dans norGA1r 3,16 % dans norIN1r ; 4,39 % dans norPM1r). On est donc là devant ce que relevait Pallaud (2004) : face aux phénomènes qui relèvent de la disfluente, et ici plus particulièrement la répétition, on observe non seulement une variation inter-locuteur, mais également intralocuteur. On est pourtant ici toujours dans la même situation de parole (une étudiante en licence qui mène une entrevue semi-dirigée face à des « experts » de la langue), le degré de contrainte ne semble pas varier d'une entrevue à l'autre, et toutes les entrevues sont faites dans un laps de temps très bref. Il est donc difficile d'émettre une hypothèse quant à la variation ici observée. Une analyse plus approfondie, qui tiendrait compte de plus de paramètres, permettrait sans doute de faire des hypothèses sur les divergences relevées ici (voir Branigan *et al.* 1999, Lickley 2001).

En ce qui concerne les autres locuteurs du corpus ayant le rôle d'intervieweur, on constate qu'ils ont en général des taux de répétition assez faibles : par exemple, ilrMF0 va, dans ses 11 entrevues, de 2,73 % à 0,47 %, soit en deçà de la moyenne du corpus de 3,19 %.

Si l'on cherche à établir une corrélation entre les différentes données que nous possédons, on n'obtient aucun résultat significatif. On n'observe ainsi aucune corrélation entre la longueur moyenne des répliques d'un locuteur et son taux de répétitions : un locuteur ayant un nombre élevé de séquences

répétitions n'a pas nécessairement des répliques dont la longueur moyenne est relativement courte ou relativement longue, et inversement. Par ailleurs, le paramètre qui consiste à prendre en compte le degré d'expertise du locuteur comme professionnel de la parole donne des résultats peu probants : on a des locuteurs en dernière année de l'enseignement professionnel qui obtiennent un score très bas (ilrGS2 : 1,61 % ; ilrLD3 : 1,64 %) et des journalistes de la presse télévisuelle avec un taux de répétitions élevé (iljDV1 : 5,54 % ; iljBB1 : 3,71 %) ; un locuteur n'ayant fait que ses études primaires (chaBR1) obtient 2,97 répétitions tous les 100 mots là où un professeur d'université (norKJ1) atteint un taux de 3,75 %.

D'après les résultats que nous obtenons ici, force est de constater que tombe le mythe qui voudrait qu'un locuteur bafouille, cafouille ou se répète parce qu'il est peu sûr de lui et/ou a une mauvaise maîtrise de la langue.

5.2.3 Répétitions et parole superposée

Nous évoquions ci-dessus (chap. 3.1) l'hypothèse que les chevauchements de parole favorisent les répétitions. Cette hypothèse est déjà avancée par Bénard, pour les disfluences en général :

Les disfluences sont particulièrement nombreuses aux frontières de chevauchements de paroles et en leur sein.
(Bénard 2005 : 54)

On regrette que l'auteure ne donne pas de chiffres ; de plus, elle supprime purement de son étude les disfluences dans ce contexte.

Nous avons calculé le nombre de répétitions dans les chevauchements. Dans notre corpus, on recense 254 répétitions de cette espèce, sur 2549 chevauchements, soit une répétition sur 10 chevauchements.

Force est de constater que les répétitions sont plus nombreuses en dehors des chevauchements. Une étude approfondie pourrait montrer si elles ont une structure différente de celles hors chevauchements.

5.2.4 Nombre de répétés (n)

Si l'on observe nos données en fonction du nombre de répétés, la répartition dans le corpus est la suivante :

Nombre de répétés	Nombre d'occurrences	Pourcentage
1	10661	87,44 %
2	1217	9,98 %
3	241	1,98 %
4	44	0,36 %
5	18	0,15 %
6	6	0,05 %
7	3	0,02 %
8	0	0 %
9	1	0,01 %
10	1	0,01 %

Tableau 9. Répartition selon le nombre de répétés

Avec 87,44 % de l'ensemble, les répétitions à un seul répété sont largement majoritaires. Elles sont suivies par celles constituées de 2 répétés, les pourcentages d'occurrences allant décroissant – et devenant pratiquement insignifiants – jusqu'à 10 répétés (n=10). On peut s'étonner de trouver des séquences où le nombre de répétés est si élevé.

Pourtant, comme le fait remarquer Blanche-Benveniste, à propos de ces répétés parfois très nombreux dans une séquence :

On peut aller jusqu'à sept, sans que cela se remarque.
(Blanche-Benveniste 2000c : 19)¹⁸

Ainsi, si à la lecture de l'oral transcrit on est inévitablement attiré par ces marques qui ne correspondent pas à nos habitudes de lecteur, dont l'œil est aguerri à de l'écrit standard, notre oreille semble quant à elle ignorer totalement ces marques qui passent donc la plupart du temps inaperçues¹⁹ tant elles sont communes dans l'oral spontané.

¹⁸ Il existe des seuils au-delà desquels on passe dans la *dysfluence* pathologique (cf. Zellner 1992).

¹⁹ Sauf dans des exercices spécifiques mis en place lors de tests de perception comme ceux de Candéa (2000a) ou de Duez (1997) par exemple, où l'attention du sujet est explicitement attirée par ce phénomène.

En ce qui concerne notre œil, le correcteur orthographique de Microsoft attire d'ailleurs notre attention sur les formes répétées contigües en les soulignant en rouge, ce qui est souvent inutile.

Le répétable ayant le plus grand nombre de répétés est l’adverbe *non*. Voici la séquence répétition qui comprend 11 *non* d’affilée, sans aucune autre marque entre le répétable et le premier répété, ni entre les répétés eux-mêmes :

accCT0 et c’est grossier |- (xxx) ?

accPH1 non non non -| non non non non non non

non non c’est pas grossier / |- non non non

<accCT0> c’est ça -| non // c’est un wallon euh //

français quoi // que je veux dire ce n’est |- pas

accCT0 plus -| teinté du français [accPH1r]

Plus loin dans l’énoncé, on trouve une seconde répétition de *non* chez ce même locuteur, toujours sans pause, où $n=3$.

En fait, les répétés de longueur $n \geq 7$ concernent exclusivement les adverbes *oui* et *non*. Dans notre corpus, ces deux adverbes sont des répétables particulièrement fréquents, puisque le *oui* est l’objet de 824 répétitions (c’est d’ailleurs le répétable le plus fréquent de notre corpus) et le *non* de 296 répétitions²⁰. Cette forte fréquence est évidemment liée au fait qu’une grande partie de nos textes sont des entrevues semi-dirigées, dans lesquelles les locuteurs répondent à une série de questions, comme c’est le cas dans l’exemple repris ci-dessus. *Oui* et *non* sont donc des formes très fréquentes en corpus, et le lieu privilégié de répétitions. On est en droit de se demander si celles-ci relèvent véritablement du phénomène de la disfluence tel que nous l’avons décrit. En y regardant de plus près, s’il fallait trancher, nous enlèverions sans doute ces occurrences des cas de répétitions disfluentes. C’est sans doute là l’un des facteurs qui explique les écarts entre nos résultats et ceux des études antérieures (notamment Henry *et al.* 2004), en termes de fréquence du phénomène.

On peut faire l’hypothèse que si le nombre de répétés est élevé, la répétition n’est pas disfluente. Cela se vérifie pour les séquences où $n \geq 7$.

Mais pour les séquences où $n=6$, si la répétition semble bien intensive quand le répétable est *oui*, *non* ou *pep* (onomatopée), il en va autrement pour *de*. Dans ce cas, la répétition relève sans nul doute de la disfluence :

iljVM1 c’est un petit peu comme les personnes qui vont euh qui vont quitter euh par exemple l’Europe pour aller vivre au Canada après quelques mois ils vont commencer à parler avec l’accent canadien / c’est quelque chose qui / qui fait partie de (silence) de de de de // de de l’auditif ambiant dirais-je [iljVM1r]

²⁰ Ces chiffres indiquent le nombre de séquences répétitions où le répétable est *oui* ou *non*, indépendamment du nombre de répétés.

Le nombre de répétés, s'il doit être pris en compte, n'est donc pas une **preuve** évidente de la disfluence de la séquence.

Pour les répétitions où $n=5$, le répétable est là aussi *oui* (et sa variante *ouais*), *non*, *tut* (onomatopée), mais également les formes *et*, *des* et *ou*.

accGP1 ou ou ou ou ou ou -| de / p / // ou de la voix
c'est-à-dire que c'est difficile à percevoir du fait qu'ils
crient [accGP1r]

Dans ce cas, où le répétable est également un mot grammatical monosyllabique, la répétition est clairement disfluente.

C'est en fait moins le nombre de répétés seul que la catégorie grammaticale du répétable qui permet de formuler des hypothèses quant à la disfluence ou non de la répétition. Nous reviendrons ci-dessous sur le problème de la catégorie grammaticale.

5.2.5 Longueur du répétable (l)

La longueur du répétable (l), en termes de nombre de mots graphiques, est variable : elle va de $l=1$ à $l=8$. Les répétitions les plus fréquentes sont celles où le répétable est composé d'un seul mot ; elles représentent près de trois-quart des occurrences. Ici aussi, on observe une régularité indéniable : le nombre d'occurrences des séquences répétées va en décroissant selon que la longueur du répétable augmente.

longueur du répétable	nombre d'occurrences	pourcentage
1	9045	74,19 %
2	2141	17,56 %
3	694	5,69 %
4	217	1,78 %
5	58	0,48 %
6	24	0,20 %
7	10	0,08 %
8	3	0,02 %

Tableau 10. Répartition selon la longueur du répétable

Si l'on croise maintenant les deux paramètres de la longueur du répétable et du nombre de répétés, on obtient les deux tableaux suivants. Le tableau 11

donne les chiffres en fréquence absolue et le tableau 12 présente la répartition en pourcentages.

Longueur du répétable	Nombre de répétés										
	1	2	3	4	5	6	7	8	9	10	Total
1	7703	1049	225	39	18	6	3	0	1	1	9045
2	1983	139	15	4	0	0	0	0	0	0	2141
3	668	24	1	1	0	0	0	0	0	0	694
4	212	5	0	0	0	0	0	0	0	0	217
5	58	0	0	0	0	0	0	0	0	0	58
6	24	0	0	0	0	0	0	0	0	0	24
7	10	0	0	0	0	0	0	0	0	0	10
8	3	0	0	0	0	0	0	0	0	0	3
Total	10661	1217	241	44	18	6	3	0	1	1	12192

Tableau 11. Répartition selon la longueur du répétable et le nombre de répétés (valeur absolue)

Longueur du répétable	Nombre de répétés										
	1	2	3	4	5	6	7	8	9	10	Total
1	63,18 %	8,60 %	1,85 %	0,32 %	0,15 %	0,05 %	0,02 %	0,00 %	0,01 %	0,01 %	74,19 %
2	16,26 %	1,14 %	0,12 %	0,03 %	0	0	0	0	0	0	17,55 %
3	5,48 %	0,20 %	0,01 %	0,01 %	0	0	0	0	0	0	5,70 %
4	1,74 %	0,04 %	0	0	0	0	0	0	0	0	1,78 %
5	0,48 %	0	0	0	0	0	0	0	0	0	0,48 %
6	0,20 %	0	0	0	0	0	0	0	0	0	0,20 %
7	0,08 %	0	0	0	0	0	0	0	0	0	0,08 %
8	0,02 %	0	0	0	0	0	0	0	0	0	0,02 %
Total	87,44 %	9,98 %	1,98 %	0,36 %	0,15 %	0,05 %	0,02 %	0,00 %	0,01 %	0,01 %	

Tableau 12. Répartition selon la longueur du répétable et le nombre de répétés (pourcentage)

À la lecture de ce tableau, on peut faire trois remarques principales, qui reprennent partiellement ce qui a déjà pu être observé pour chaque paramètre séparément :

- en croisant nos deux paramètres, on constate que certaines configurations sont totalement absentes. Elles sont mêmes majoritaires sur l'ensemble du tableau puisque 57 des 80 possibilités ne sont pas actualisées dans notre corpus, qui comprend pourtant un nombre de séquences répétitions non négligeable ;
- les répétitions les plus fréquentes concernent majoritairement des répétables constitués d'un seul mot, répété une seule fois, puisque cette configuration représente 63,18 % du total des occurrences. Viennent ensuite les répétables composés de deux mots répétés également une seule fois avec 16,26 %. Ces séquences sont suivies par les répétables d'un mot répétés deux fois (8,60 %) puis par les répétables de trois mots répétés une seule fois (5,48 %). Ces quatre configurations constituent à elles seules 93,52 % de toutes nos occurrences de séquences répétitions ;
- les 18 autres configurations représentées dans le corpus totalisent chacune un nombre d'occurrences pratiquement insignifiant.

On le voit donc ici, apparaît dans ces données croisées une très grande régularité dans la forme que prennent les séquences répétitions. Nous allons voir que cette régularité concerne également les répétables.

5.2.6 Distribution des répétables

Dans notre corpus, 1757 formes différentes de répétables sont concernées par le phénomène de la répétition, indépendamment du nombre de répétés.

5.2.6.1 Formes les plus fréquentes

Les 30 formes les plus fréquentes dans notre corpus sont les suivantes, présentées par ordre de fréquence décroissant :

oui (824), *de* (732), *c'est* (613), *je* (424), *des* (382), *les* (335), *et* (334), *ça* (317), *qui* (309), *non* (296), *le* (291), *à* (281), *un* (265), *on* (253), *euh* (233), *dans* (204), *mais* (187), *la* (177), *en* (159), *que* (157), *il* (151), *vous* (142), *il y a* (140), *une* (121), *ou* (119), *très* (105), *ils* (88), *si* (78), *du* (73), *ouais* (71).

On retrouve dans cette liste les adverbes *oui*, *non* et *ouais* déjà mentionnés ci-dessus. Ce qui saute aux yeux à la lecture de cette liste est que tous les items sont des monosyllabes appartenant à la métacatégorie des mots

grammaticaux : pronom, déterminant, préposition, conjonction et adverbe. Si le *c'est* peut être analysé comme une suite PRO + V, on voit bien son fonctionnement comme une forme unique de présentatif. De la même manière, *il y a* n'est qu'une exception apparente à la catégorie des monosyllabes. Transcrit en une suite de trois mots graphiques, *il y a* est bien souvent prononcé par le locuteur en une seule syllabe [ja]²¹.

Ces 30 items les plus fréquents totalisent 7861 occurrences, soit près de 65 % des répétitions de notre corpus.

5.2.6.2 Les hapax

On a une majorité de répétables qui ne sont l'objet d'une répétition (quel que soit le nombre de répétés) qu'à une seule reprise dans le corpus.

En effet, sur les 1757 répétables différents, 1303 sont des hapax, ce qui représente 74,1 % des répétables recensés. Ce chiffre peut étonner, mais en fait il va dans le sens de la distribution des vocables en corpus, et ceci indépendamment du phénomène de la répétition, objet qui nous intéresse ici. En effet, comme nous l'avons déjà mentionné au chapitre 2, on a dans tout corpus 1) un ensemble composé d'un nombre restreint de formes qui apparaissent très fréquemment, 2) un second ensemble composé d'un nombre important de formes « rares », dans le sens où elles apparaissent une seule fois dans le corpus et 3) un 3^e ensemble qui comprend des vocables dont la fréquence d'apparition est entre ces deux groupes (cf. la loi de Zipf, 1949). La distribution des répétables ne semble donc pas déroger à cette règle générale, qu'on peut vérifier quel que soit le type de corpus, même si le pourcentage d'hapax est ici plus élevé que celui des hapax calculé sur l'ensemble des formes du corpus.

Les hapax concernent aussi bien des répétables longs que des répétables courts. Néanmoins, chaque catégorie est représentée en proportions variables : comme l'on pouvait s'y attendre, tous les répétables les plus longs sont des hapax ($l \geq 6$) ; ensuite, la répartition entre hapax et non-hapax varie selon la longueur du répétable, mais quelle que soit l , les hapax sont toujours plus nombreux (cf. tableau ci-dessous).

On a vu ci-dessus que les répétables les plus fréquents étaient des mots grammaticaux. Toutefois, on constate dans nos données que cette catégorie des mots grammaticaux n'est pas absente des hapax, qu'il s'agisse de prépositions (*devant*), d'adverbes (*absolument*, *aussi*), de déterminants (*ces*, *cet*, *tel*), d'onomatopées (*bah*, *bang*, *bouf*, *ouh*), de conjonctions (*quoique*, *puisque*) ou encore de pronoms (*cela*, *celui*, *ceux*, *eux*, *t'*, *te*).

²¹ Quelle que soit la prononciation, la transcription est *il y a* (voir le chap. 1, § 6.2.1 sur les trucs orthographiques).

Parmi les hapax où le répétable est de longueur $l=1$, les mots lexicaux sont largement majoritaires. En effet, ils représentent 84,4 % des occurrences de cette catégorie.

Néanmoins, il ne faudrait pas penser que les répétables qui sont des mots lexicaux seuls (c'est-à-dire sans déterminant d'aucune sorte) appartiennent à la catégorie des hapax. En effet, nous avons 21 répétables mots lexicaux seuls qui apparaissent dans des séquences répétitions à plus d'une reprise. Parmi eux : 2 adjectifs (*dur* et *normal*), un nom commun (*monsieur*), un nom propre (*Charleroi*), et 11 verbes (13 conjugués et 4 à l'infinitif). Pour ce qui est des verbes conjugués, on observe une constante dans la liste qui suit : *dis, doit, donne, es, font, parlent, vont, attends, avait, peut, ont, faut et sont*. En effet, tous ces verbes sont des verbes très fréquents en corpus, indépendamment de leur mise en place dans une séquence répétition. De plus, hormis *attends*, ils sont tous monosyllabiques. En cela, ils rejoignent d'une certaine manière les mots grammaticaux.

Le chiffre de 74,1 % des hapax est élevé, mais ramené à la fréquence effective des répétitions dans le corpus, et non à l'ensemble des formes de répétables différentes, ces hapax constituent seulement 10,69 % des occurrences.

Le tableau 13 synthétise ces données. La première colonne est la longueur l du répétable, la deuxième donne le nombre de répétables hapax, la troisième indique le nombre de répétables différents faisant au moins deux fois l'objet d'une répétition²², la quatrième colonne donne le pourcentage des occurrences d'hapax en corpus, la dernière colonne reprend la fréquence des occurrences de formes qui ne sont pas des hapax.

²² Les chiffres des colonnes 2 et 3 donnent le nombre de répétables différents (qui apparaissent une seule fois ou plusieurs fois) et non le nombre d'occurrences en corpus. Ces données ne font pas de distinction en fonction du nombre de répétitions.

longueur <i>l</i> du répétable	hapax n=	répétables apparaissant au moins 2 fois n=	Pourcentage des occurrences d'hapax en corpus	Pourcentage des occurrences de répétables apparaissant au moins 2 fois en corpus
1	218	177	1,79 %	72,40 %
2	482	210	3,95 %	13,61 %
3	358	49	2,94 %	2,76 %
4	152	17	1,25 %	0,53 %
5	56	1	0,46 %	0,01 %
6	24	0	0,20 %	0 %
7	10	0	0,08 %	0 %
8	3	0	0,02 %	0 %
TOTAL	1303	454	10,69 %	89,31 %

Tableau 13. Tableau synthétique des hapax

5.3 Forme de la séquence répétition : répétition directe, répétition associée, séquence mixte

Nous avons donné ci-dessus une analyse globale des répétitions, sans faire de distinction dans la structure de la séquence répétition selon que ses composants sont contigus (*répétition directe ou contigüe*) ou séparés par l'une des marques que nous avons prises en considération²³ (*répétition associée*)²⁴. La répétition peut également être cooccurrente à une autre marque disfluente (*séquence mixte*), qui est présente soit dans le contexte gauche de la répétition, soit dans son contexte droit, soit dans ces deux positions à la fois.

En effet, comme c'est souvent le cas pour les marques de disfluences, les répétitions sont fréquemment associées à une autre marque. Chez Grosjean et Deschamps (1972, étude sur des interviews radiophoniques), 49 % des

²³ Voir ci-dessus (§ 5.1.1) pour la liste de ces marques et la manière dont nous les avons répertoriées.

²⁴ La terminologie est celle de Candéa (2000a). Henry (2001) parle de répétition *directe* pour la répétition *contigüe* de Candéa.

répétitions sont soit suivies soit précédées d'autres types de pauses (c'est-à-dire, chez ces auteurs : un allongement vocalique, un *eu**h*, un faux départ ou une pause silencieuse). Notons que Grosjean et Deschamps comptabilisent parmi les répétitions les répétitions de syllabes (dont ils ont 3,67 % d'occurrences sur l'ensemble des séquences répétitions) que nous classons parmi les amorces de morphèmes. Les auteurs concluent :

Ceci montre que certaines coupures dans l'encodage dues à la sélection de l'énoncé ou à la recherche des termes, nécessitent parfois une suite de plusieurs formes de pauses avant que l'encodage ne puisse reprendre.

Soulignons enfin la corrélation du nombre de répétitions avec le nombre de pauses sonores ($r = 0,56$) et avec la [*sic*] nombre de faux départ [*sic*] ($r = 0,56$) ; celle-ci nous renforce dans notre hypothèse que lorsqu'un sujet hésite il utilise tous les moyens possibles de pauser qui sont à sa disposition, quelle que soit sa vitesse de parole. (Grosjean et Deschamps 1972 : 154)

Dans leur étude de 1973 sur des descriptions de dessins, les auteurs remarquent une forte augmentation du nombre de répétitions précédées ou suivies d'une autre marque, puisque ce cas de figure représente 70 % de leurs occurrences. Selon eux, cette augmentation du nombre de séquences mixtes est liée à l'effort que doit produire le locuteur, dans une activité langagière plus contraignante (ici décrire des dessins inconnus *vs* participer à une interview sur un sujet bien connu du locuteur).

Candéa (2000a), dans sa thèse sur les phénomènes dits « d'hésitation », constate que la répétition est la marque qui se combine le plus souvent avec une ou plusieurs autres marques. Dans son corpus, très réduit il est vrai (110 répétitions), seulement 12 % des répétitions sont des occurrences simples (c'est-à-dire sans autre marque du travail de formulation), et 8 % des répétitions contiennent une pause silencieuse entre le répétable et le répété. Il faut remarquer que Candéa prend en considération parmi les autres marques l'allongement final, absent de nos transcriptions. Celui-ci concerne 53,93 % des répétables de son corpus. Ainsi, si l'on avait pu tenir compte de cette marque particulièrement fréquente (toujours selon l'étude de Candéa, c'est la seconde marque de travail de formulation la plus fréquente, après le *eu**h*), le taux de séquences mixtes de notre corpus aurait été de toute évidence plus important.

5.3.1 Répétition directe *vs* répétition associée

Dans notre corpus, les répétitions strictement contigües (répétitions directes) sont les plus nombreuses, avec 63,12 %, de l'ensemble (7695 occurrences), contre 36,88 % de répétitions associées (4497). Le tableau 14

détaille ces résultats selon le nombre n de répétés, indépendamment de la longueur l du répétable. Nous avons laissé de côté les répétitions où $n > 4$, les effectifs étant relativement peu nombreux (cf. tableau ci-dessus : $n=5$ compte 18 occurrences, celles-ci allant en décroissant selon que n augmente). Le pourcentage des colonnes 3 et 4 (répétitions contigües et répétitions associées) est calculé sur l'ensemble des occurrences de chaque catégorie établie selon le nombre de répétés (et non sur l'ensemble des occurrences du corpus).

nombre n de répétés	nombre d'occurrences	pourcentage de répétitions directes	pourcentage de répétitions associées
1	10661	62,80 %	37,20 %
2	1217	65,65 %	34,35 %
3	241	61,41 %	39,49 %
4	44	65,91 %	34,09 %

Tableau 14. Répartition des répétitions directes et associées selon le nombre de répétés

Dans notre corpus global, les termes de la séquence répétition sont plus souvent contigus que séparés par l'insertion d'un élément. C'est également le cas quel que soit le nombre de répétés de la séquence, où l'avantage est en faveur des répétitions contigües, dans un rapport qui avoisine les 2/3 vs 1/3.

Le test du χ^2 (khi carré) effectué sur ces données montre que les différences de pourcentages du tableau ci-dessus ne sont pas statistiquement significatives, et donc que la proportion des répétitions directes et associées ne varie pas significativement d'une classe à l'autre (nombre de répétés). Pour le dire autrement, le nombre de répétés n'a donc aucune incidence sur le format de la répétition. On est donc en présence d'une répartition stable entre répétitions directes et répétitions associées et cette tendance générale se maintient indépendamment du nombre de répétés.

En ce qui concerne la position de la marque insérée, nous avons analysé les séquences où $n=2$ et où $n=3$ ²⁵. Le tableau 15 reprend la répartition des marques insérées à l'intérieur de la séquence répétition. Ce tableau est à lire de la manière suivante : R représente l'un des termes de la répétition (qu'il s'agisse du répétable ou du répété) et x symbolise l'élément inséré. La 3^e colonne RRR (ou $RRRR$ dans le cas de $n=3$) représente les répétitions

²⁵ Il n'y a évidemment pas de variation pour $n=1$, puisque la seule position possible pour l'insertion est entre le répétable et le répété.

directes. Les autres colonnes déclinent les différents types de répétitions associées en fonction de la place de l'insertion dans la séquence répétition.

	nombre d'occurrences	<i>RRR</i>	<i>RxRR</i>	<i>RRxR</i>	<i>RxRxR</i>				
n=2	1217	799	170	165	83				
n=3		<i>RRRR</i>	<i>RxRRR</i>	<i>RRxRR</i>	<i>RRRxR</i>	<i>RxRxRR</i>	<i>RxRRRxR</i>	<i>RRxRxR</i>	<i>RxRxRxR</i>
	241	148	25	22	21	6	6	8	5

Tableau 15. Insertions dans les répétitions

On voit que les insertions à un seul endroit sont largement plus fréquentes que les insertions qui se placent dans deux positions différentes de la séquence répétition (lorsqu'il y a deux x dans la configuration). On a en effet, à l'intérieur des répétitions associées, une écrasante majorité d'insertions à un seul endroit, respectivement 80,14 % quand n=2 et 73,12 % quand n=3.

Il ne semble pas y avoir de lieu privilégié quant à la place de cette insertion, la légère faveur pour la position entre le répétable et le premier répété n'étant pas significative.

5.3.2 Les insertions les plus fréquentes dans les répétitions associées

Le tableau 16 synthétise la fréquence d'occurrences des marques que nous avons prises en considération à l'intérieur d'une répétition, indépendamment de la longueur du répétable et du nombre de répétés. Elles sont classées par fréquence décroissante.



marque insérée	n=	pourcentage	marque insérée	n=	pourcentage
/	2389	44,30 %	<i>ah</i>	11	0,20 %
[[amorce]]	1094	20,29 %	<i>hum</i>	11	0,20 %
//	692	12,83 %	<i>ben</i>	10	0,19 %
<i>euh</i>	510	9,46 %	<i>oh</i>	10	0,19 %
<i>oui</i>	132	2,45 %	<i>pff</i>	10	0,19 %
(silence)	83	1,54 %	(x)	9	0,17 %
<i>non</i>	71	1,32 %	(toux)	7	0,13 %
<i>enfin</i>	70	1,30 %	(xx)	7	0,13 %
<i>hein</i>	66	1,22 %	(xxx)	7	0,11 %
(rire)	56	1,04 %	<i>bè</i>	4	0,07 %
<i>bon</i>	33	0,61 %	<i>mm</i>	4	0,07 %
<i>m</i>	32	0,59 %	(rires)	2	0,04 %
<i>pf</i>	22	0,41 %	(soupir)	2	0,04 %
<i>disons</i>	17	0,32 %	<i>bf</i>	2	0,04 %
<i>comment</i>	16	0,30 %	<i>boh</i>	2	0,04 %
<i>ouais</i>	13	0,24 %			

Tableau 16. Insertions dans les répétitions associées

La marque insérée de loin la plus fréquente est la pause brève (/), qui totalise 44,30 % des insertions dans les séquences répétitions. C'est aussi, parmi les marques reprises ici, la plus fréquente dans le corpus. On constate donc une forte interaction entre les répétitions et la pause, interaction que nous retrouverons également avec le *euh* (cf. chap. 3.5).

Nous avons voulu voir s'il y avait corrélation entre la fréquence des marques dans le corpus et leur fréquence d'apparition à l'intérieur d'une séquence répétition.

Nous avons soumis nos données au test de Pearson (test non paramétrique). Nous avons attribué un rang à chaque marque, le rang 1 étant dévolu à la marque la plus fréquente. Cette attribution se fait pour deux classements : d'une part selon la fréquence décroissante des marques dans les répétitions (*Rang in*), et d'autre part selon la fréquence décroissante des

marques dans un autre contexte (*Rang out*). Le graphique de dispersion des données est le suivant :

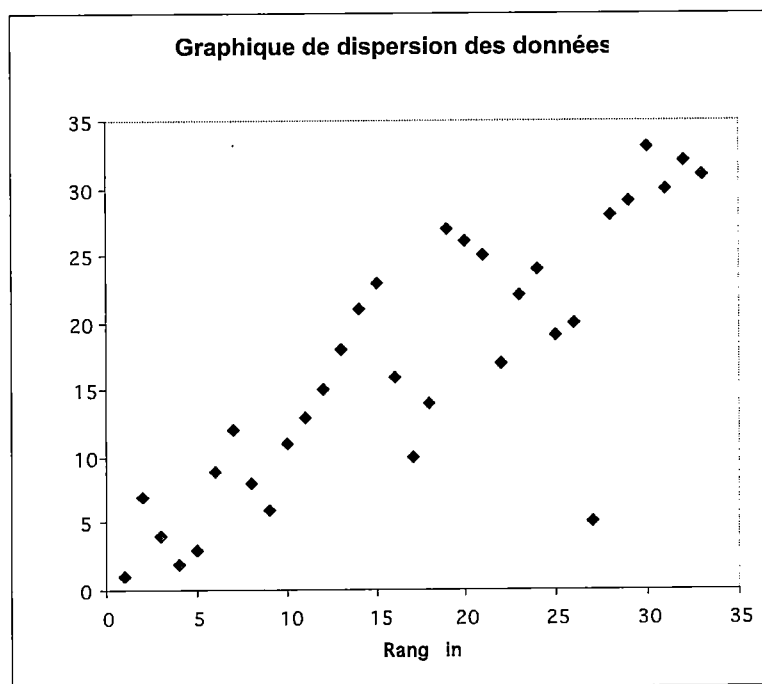


Figure 10. Graphique de dispersion des marques insérées

Le test de Pearson donne une valeur de 0,832²⁶. Cela signifie que les deux classements sont fortement corrélés. En d'autres termes, une marque à l'intérieur d'une séquence répétition n'est proportionnellement ni plus ni moins fréquente que dans un autre contexte. La probabilité que ce résultat soit dû au hasard est inférieure à 0,001.

La seule exception concerne la marque *mm* représentée par le point en bas à droite sur le graphique : on en a 4 occurrences dans une séquence répétition (rang 27) contre 3340 dans un autre contexte (rang 5). En fait, cette onomatopée semble réservée à l'approbation quasi silencieuse de la part de l'interlocuteur et n'a pas la même fonction discursive que les autres onomatopées que nous avons prises en considération.

En ce qui concerne la pause (pour ses trois degrés), nous verrons le traitement que nous proposons d'en faire au chapitre 4.

²⁶ Plus la valeur est proche de 1, plus la corrélation est forte. Une valeur à 1 signifie une corrélation directe et maximale entre les deux paramètres ; une valeur à -1 signifie que la corrélation est inverse ; pour une valeur à zéro, on n'a aucune corrélation.

De la même manière, la seconde marque la plus fréquente associée à la répétition est l'amorce de morphème, avec 1094 occurrences. Nous reviendrons sur cette interaction très forte entre ces deux phénomènes dans la section que nous consacrons à l'amorce (chap. 3.4).

Le *euh* totalise quant à lui 9,46 % des insertions dans une séquence répétition. Sans anticiper sur la section consacrée au *euh*, nous avons analysé 100 séquences sélectionnées au hasard dans lesquelles le *euh* apparaît à l'intérieur d'une répétition. Nous avons vérifié que les formats de la répétition étaient représentatifs des différents types que l'on rencontre dans le corpus : répétable de un ou plusieurs mots, répétés en nombre variable, insertion d'un *euh* seul, de plusieurs *euh* ou d'un *euh* et d'autres marques et enfin place variable de cette (ces) insertion(s). À l'analyse de nos données, on constate que 98 % des occurrences constituent un piétinement sur l'axe syntaxique, mais pas une rupture : le déroulement syntagmatique se poursuit normalement avec la reprise du *repair*. Le *euh* n'apparaît donc pas comme une marque dans l'énoncé qui modifierait le fonctionnement « régulier » de la répétition. Elle est une marque supplémentaire dans la séquence disfluente, caractérisée par un empilement de traces dans l'énoncé.

Les deux exceptions consistent en fait en un mauvais repérage de la répétition :

iljDV1 voilà désolée -| pour les euh les ah les euh
euh

iljGFO (rire) ne vous inquiétez pas (fin de l'entrevue)
[iljDV1r]

iljDV1 (...) par souci de simplifier les choses / euh on escamote un peu les / la vérité et la réalité / justement en essayant d'expliquer euh parfois euh ce que c'est une commission paritaire ce que c'est euh / c'est trop raccourci quoi // et alors le cinquième mot euh [iljDV1r]

Dans le premier exemple, le programme repère la séquence soulignée, qui va du premier *les* au troisième *les*, en passant par les insertions que nous avons autorisées : *euh* et *ah*. En fait, ici *euh* et *ah* sont utilisés métalinguistiquement par la locutrice, qui s'excuse d'avoir usé abondamment de ces formes dans la conversation. Ces usages autonymiques, qui nécessitent par ailleurs une recatégorisation grammaticale, sont impossibles à prévoir et à formaliser, et nous verrons dans les choix linguistiques que nous ne le considérons pas dans la procédure de levée d'ambiguïtés (cf. chap. 5).

Le second exemple, quant à lui, consiste en un repérage fautif de la répétition de *c'est*. Ici, le premier *c'est* constitue la fin du groupe syntaxique *ce que c'est*, alors que le deuxième *c'est* initie un autre groupe syntaxiquement autonome, *c'est trop raccourci quoi*. Face à cet exemple, on

pourrait émettre l'hypothèse que le *euh* suivi de la pause brève, à l'intérieur d'une séquence repérée comme une répétition, pourrait être un indicateur de mauvais repérage, séparant des constituants autonomes ou marquant une rupture de la construction. Or il n'en est rien, puisque dans les 100 exemples que nous avons repris ici, on a 46 autres cas où le *euh* est suivi d'une pause (brève ou longue) à l'intérieur d'une « bonne » répétition et où il n'entrave pas le fonctionnement régulier de la réparation. On échoue donc à trouver des marques formelles de rupture dans l'énoncé.

En ce qui concerne les autres formes/marques, elles ont des fréquences d'apparition très faibles puisque mises à part les quatre formes les plus fréquentes (les pauses brèves et longues, l'amorce et le *euh*), les pourcentages d'occurrences sont inférieurs à 2,5 %. Toutes ces occurrences additionnées, on n'arrive qu'à un total de 13,9 % des occurrences. Ce chiffre paraît très faible²⁷ en regard de l'intuition que l'on pouvait avoir du format d'une séquence répétition, où l'on aurait fréquemment des insertions entre répétable et répétés, insertions qui marquent l'hésitation (Blanche-Benveniste 1985). Ainsi, dans les deux exemples suivants, les termes *comment* et *disons* indiquent clairement que le locuteur est à la recherche de la dénomination exacte, qui lui permet d'exprimer au mieux sa pensée par le choix du terme le plus approprié.

accCT0 je ne sais pas |- c'est c'est <accGP1> (rire) -|
votre avis que je vous demande // donc une question de
|- prononciation ?

accGP1 oui -| // et puis le // comment le // peut-être le rythme le chantonnement // il y a un autre
terme pour / dire ça hein c' est pas chan/ c' est // c' est
pas un chantonnement mais c' est |- (xxx) [accGP1r]

ileDC1 (...) ils ont une langue assez claire / claire bien
bien découpée / un peu chantante sans sans en avoir
sans avoir la le le disons le le ton liégeois par exemple
hein alors par ailleurs euh la la proximité euh de la
France via donc la la Picardie / euh leur donne je dirais
une une faconde / sans une image // et une / une
richesse de vocabulaire / pour euh / supérieure /
légèrement supérieure (...) [ileDC1r]

On retrouve le même type de fonctionnement avec les adverbes *oui* et *non*, qui permettent au locuteur de confirmer ou de corriger ce qu'il est en train d'énoncer, qu'il s'agisse de l'idée de manière générale ou d'un lexème plus approprié. Ceci est illustré dans les exemples suivants :

²⁷ Cela dit, ce résultat faible en termes de pourcentage est dû au nombre très important de pauses silencieuses.

iljGF0 c'est Hannut c'est ça hein |- enfin près d'Hannut c'est ça

iljVM1 près d'Hannut oui entre - l oui entre Hannut et Waremme même c'est c'est Hannut c'est la commune de Geer en fait une grande commune qui s'appelle Geer G deux E R (silence) [iljVM1r]

norGA1 boh actuel je crois que ça c'est valable ça a toujours été valable c'est-à-dire que pour moi la démocratisation // par non par dans le sens que tout le monde devrait parler comme les gens du peuple mais que le plus grand nombre per/ possible de personnes / parlent bien (...) [norGA1r]

Cette utilisation du *oui* et du *non* par le locuteur, afin d'approuver ce qu'il vient de dire ou de le rectifier²⁸, se retrouve indépendamment de leur insertion à l'intérieur d'une séquence répétition.

5.3.3 Séquence mixte

Les séquences mixtes sont des séquences où la répétition cooccure avec une autre marque considérée comme disfluente. Les séquences mixtes reprennent donc toutes les séquences où une autre marque disfluente²⁹ précède ou suit la séquence répétition. On peut donc isoler deux groupes : les répétitions directes sans marque ni avant ni après et toutes les autres séquences répétitions qui cooccurrent avec une autre marque (présente soit avant, soit après, soit à l'intérieur de la séquence).

À la différence des résultats obtenus précédemment, où les répétitions directes étaient les plus fréquentes, celles-ci sont selon cette répartition minoritaires, puisqu'elles comptabilisent 5536 occurrences, soit 45,41 % du total. Nos résultats vont dans le sens de ceux de Candéa (2000a), qui constatait que beaucoup des répétitions de son corpus entraient dans une séquence mixte. Nous n'irons néanmoins pas jusqu'à dire, comme l'auteure le fait que

il est très rare que la répétition soit utilisée toute seule pour marquer le travail de formulation. (Candéa 2000a : 321)

²⁸ Notons que c'est le seul exemple que nous avons trouvé où le *non* dans la séquence répétition peut être attribué à une forme de correction de la part du locuteur. Dans les autres cas, il s'agit d'un *non* qui signifie une négation en réponse à une question. Ce cas se produit pour beaucoup des *oui* du corpus, où il est parfois difficile de faire le départ entre le *oui* qui confirme ce que le locuteur dit et le *oui* qui consiste à répondre affirmativement à une question de l'interlocuteur.

²⁹ La liste de ces marques reste inchangée.

5.3.4 Le schéma des séquences mixtes

Les séquences mixtes ont des formats très variés. On peut en effet trouver une marque seulement avant ou seulement après la répétition, avant et après la répétition et, dans ces trois cas de figure, la séquence répétition peut elle-même être contigüe ou associée. Voici un exemple de chacun de ces six formats :

iljDV1 (...) il y a non c'est vraiment curieux [il y a il y a]
/ euh / surprenant / (...) [iljDV1r]

ilpMJ1 (...) le système démocratique dans lequel ils |
 vivent <ilpGV0> mm -| / les institutions **euh** [qui les qui
 les] servent et cetera [ilpMJ1r]

iljGE1 (...) et en pluS on a une on un accent **euh** [les
 les] **euh** assez typé et qui est assez / assez reconnu (...)
 [iljGE1r]

norKJ1 oui -| / oui je pense [j'irai chercher le // l/ j' irai
 chercher le] / programme des cours après si vous |-
 voulez ou [norKJ1r]

famRM1 (...) et puis eux ils sont partis habiter euh /
 euh au Luxembourg pendant très longtemps // [et euh
 / et] maintenant ils sont ils sont revenus [famRM1r]

accPH1 **bè** [c'est / c'est moi // |- c'est moi] **hein**
 [accPH1r]

De plus, on peut avoir plus d'une marque dans le contexte de la répétition (et là aussi les marques sont soit antérieures, soit subséquentes, soit dans les deux contextes).

Nous appelons *marque contigüe* une marque présente dans le contexte direct de la répétition ; une *marque non contigüe* est séparée de la répétition par une autre marque, ou par plusieurs autres marques.

Tous schémas de répétitions confondus, la répartition des marques antéposées et postposées aux répétitions est la suivante. Nous n'avons pas fait de sous-classement dans les marques non contigües en fonction de leur éloignement par rapport à la répétition :

	marques contexte gauche		marques contexte droit		total	
	n=	pourcentage	n=	pourcentage	n=	pourcentage
marque contigüe	2609	52,64 %	1026	20,70 %	3635	73,35 %
marque non contigüe	1018	20,54 %	303	6,11 %	1321	26,65 %
total	3627	73,18 %	1329	26,82 %	4956	100 %

Tableau 17. Répartition du contexte des marques

Au vu du tableau ci-dessus, on constate que les marques sont beaucoup plus fréquentes dans le contexte gauche que dans le contexte droit de la répétition³⁰. Les séquences mixtes sont donc préférentiellement des séquences où la disfluence commence avant la présence de la répétition.

Proportionnellement, les marques non contigües sont également plus fréquentes à gauche qu'à droite.

Les données du tableau 18 prennent en compte le nombre de marques antérieures ou subséquentes, indépendamment du nombre de séquences répétitions concernées par cette cooccurrence. C'est ce à quoi nous allons nous intéresser maintenant, en distinguant les répétitions selon qu'elles sont directes ou associées. Pour rappel, les répétitions directes concernent 7695 occurrences et les répétitions associées 4497 occurrences. Le tableau suivant, où les résultats sont synthétisés en pourcentage pour chacune de ces deux catégories, ne doit donc pas occulter le fait que les effectifs des répétitions directes sont plus nombreux.

	répétitions directes			répétitions associées		
	0 après	x après	total	0 après	x après	total
0 avant	71,94 %	5,81 %	77,75 %	73,07 %	7,24 %	80,31 %
x avant	20,14 %	2,11 %	22,25 %	17,61 %	2,08 %	19,69 %
total	92,08 %	7,92 %	100 %	90,68 %	9,32 %	100 %

Tableau 18. Répartition des répétitions directes et associées en fonction des marques

³⁰ Nous n'avons, dans notre corpus, aucune suite de répétitions différentes qui feraient qu'une marque comptabilisée ici dans le contexte droit d'une répétition le serait également dans le contexte gauche d'une autre répétition.

Le tableau est à lire de la manière suivante : *0 avant* signifie qu'aucune marque ne précède la répétition, *0 après* qu'aucune marque ne la suit ; *x avant* veut dire qu'au moins une marque précède la répétition et *x après* qu'au moins une marque lui est subséquente.

En croisant ces quatre paramètres, on constate que la configuration la plus fréquente est celle où aucune marque ne précède ni ne suit la répétition. En effet, dans 71,94 % des cas de répétitions strictement contigües, la répétition apparaît seule, non marquée par un autre élément dans son contexte direct³¹. Viennent ensuite les répétitions ayant une marque à gauche et rien à droite (20,14 %), puis celles ayant une marque à droite et aucune marque à gauche (5,81 %) et enfin les répétitions entourées de marques, avec seulement 2,11 % des occurrences.

On constate un même ordre de fréquence décroissant parmi les répétitions associées. Ceci peut étonner. En effet, on aurait pu s'attendre à ce que les répétitions associées, d'une certaine manière plus disfluentes puisqu'elles cumulent déjà les marques de différentes sortes, soient accompagnées de marques supplémentaires dans leur contexte gauche et/ou droit. Or on constate qu'il n'en est rien, les répétitions associées sans autre marque représentant près de trois-quarts des occurrences (73 %).

Toutefois, le test du Khi² montre une différence significative entre répétitions contigües et répétitions associées : une répétition contigüe est plus fréquemment précédée d'une autre marque disfluente qu'une associée (20,14 % contre 17,61 %), et inversement, qu'une répétition associée est plus fréquemment suivie d'une autre marque (7,24 % contre 5,81 %).

On ne voit pas clairement, à l'analyse de nos données, quelles configurations seraient absentes des séquences mixtes. Comme nous l'avons vu pour les répétitions associées, si certaines tendances apparaissent, on ne peut en exclure d'autres. C'est pourtant ce que fait Candéa, quand elle affirme que certaines combinaisons sont impossibles, voire interdites :

(...) la position située après la répétition ne peut pratiquement pas être occupée par une pause silencieuse, si ce n'est lorsque le MO2 ou le MP2 est suivi par une autocorrection immédiate. (...) nous pouvons formuler très sérieusement l'hypothèse que la combinaison double *répétition* + *p.sil.* (de type **le le /p.sil./ cordonnier*) est totalement interdite. (Candéa 2000 : 333-334)

³¹ Nous parlons ici de contexte direct et des marques prises en considération dans ce travail. Il faudrait bien sûr étendre l'étude à d'autres marques et ne pas prendre le facteur de stricte contigüité.

Candéa (2000 : 338) décrit encore une autre configuration qu'elle ne possède pas et dont elle suppose la rareté voire l'inexistence ; il s'agit de la suite *eah + MO1 + MO2 + mot plein cible*. Or, nous en avons des occurrences dans le corpus.

ileAO1 oui di/ disons que là on est dans le sujet qu'on
on traite bon il y a des sujets **eah le le style** d'écriture
qui varie selon qu'on retrace un peu le billet d'humeur
[ileAO1r]

Cela dit, deux facteurs nécessitent que nous apportions un bémol, facteurs qui rendent parfois fragiles les comparaisons que l'on peut faire :

- dans nos textes, la notation de la pause est subjective et la tendance est de ne pas la placer à l'intérieur d'un groupe nominal (cf. chap. 4.4 ; dans ce dernier extrait, pas de pause dans tout le tour de parole) ;
- nous ne notons pas l'allongement. Quand Candéa parle de l'interdiction ci-dessus, c'est en considérant que ni le répétable ni le répété ne subissent un allongement final, ce dont nous sommes incapable de rendre compte, cette indication ne figurant pas dans nos textes.

Toutefois, rappelons que Candéa travaille sur un nombre restreint d'occurrences (seulement 100 cas), et que l'on ne peut exclure que ses résultats soient un effet du corpus plutôt qu'une tendance générale du parler.

5.3.5 Les marques les plus fréquentes dans les séquences mixtes

Voyons maintenant quelles sont les marques les plus fréquentes qui précèdent et suivent une séquence répétition. Le tableau 19 reprend les fréquences d'occurrences des marques. Il distingue les marques selon leur position à gauche ou à droite de la répétition et selon qu'elles sont contigües ou non contigües (dans ce dernier cas, il n'y a pas de sous-catégorisation en fonction de l'éloignement). Une colonne donne le pourcentage des occurrences sur l'ensemble des marques dans la même position (antérieure ou postérieure à la répétition).

	avant contigu	avant non contigu	avant total	pourcentage avant total	après contigu	après non contigu	après total	pourcentage après total
/	818	205	10230	28,21%	336	57	393	29,57%
euh	411	222	633	17,45%	105	48	153	11,51%
//	284	92	376	10,37%	132	18	150	11,29%
oui	138	105	243	6,70%	40	36	76	5,72%
ah	156	49	205	5,65%	12	11	23	1,73%
[[amorce]]	120	47	167	4,60%	117	21	138	10,38%
hein	91	45	136	3,75%	17	10	27	2,03%
enfin	108	18	126	3,47%	25	14	39	2,93%
non	74	49	123	3,39%	23	14	37	2,78%
bon	82	31	113	3,12%	31	11	42	3,16%
ben	70	25	95	2,62%	8	10	18	1,35%
(silence)	53	16	69	1,90%	48	9	57	4,29%
(rire)	25	18	43	1,19%	23	4	27	2,03%
oh	28	13	41	1,13%	1	2	3	0,23%
bè	32	9	41	1,13%	4	2	6	0,45%
mm	19	15	34	0,94%	4	5	9	0,68%
ouais	21	9	30	0,83%	2	4	6	0,45%
m	16	12	28	0,77%	12	9	21	1,58%
pf	12	9	21	0,58%	7	2	9	0,68%
disons	11	2	13	0,36%	12	3	15	1,13%
(xxx)	8	2	10	0,28%	15	0	15	1,13%
(x)	5	4	9	0,25%	7	1	8	0,60%
pff	5	5	10	0,28%	3	1	4	0,30%
(xx)	5	1	6	0,17%	10	1	11	0,83%
(toux)	5	3	8	0,22%	2	1	3	0,23%
comment	5	2	7	0,19%	25	9	34	2,56%
hum	1	5	6	0,17%	1	0	1	0,08%
(soupir)	2	4	6	0,17%	1	0	2	0,15%
(rires)	2	1	3	0,08%	2	0	2	0,15%
boh	2	0	2	0,06%	1	0	1	0,08%
	2609	1018	3627		1026	303	1329	

Tableau 19. Fréquence des marques dans les séquences mixtes

Le classement ci-dessus est effectué suivant le nombre décroissant d'occurrences des marques antéposées à la répétition.

Plusieurs remarques peuvent être faites à la lecture du tableau :

- le classement ne suit pas le même ordre de fréquence pour les marques antérieures et les marques subséquentes à la répétition ;
- la pause brève est toujours la marque la plus fréquente, comme c'était le cas pour les marques insérées à l'intérieur d'une séquence répétition ;
- néanmoins, la pause brève obtient un score plus faible qu'à l'intérieur de la séquence répétition (29 % ici contre 50 % en interne) ;
- la pause longue et le *euh* restent les autres marques les plus fréquentes, ainsi que l'amorce de morphèmes, particulièrement en position subséquente ;
- plusieurs autres marques obtiennent des scores plus élevés qu'à l'intérieur de la répétition, notamment les particules *bon, hein, non* ;
- certaines particules obtiennent des scores insignifiants, comme c'était déjà le cas à l'intérieur des répétitions (*boh, hum, pff*, etc.).

Nous avons également appliqué le test de Pearson sur ces données. Les conclusions que l'on peut tirer de ce test sont les suivantes :

- on constate une très forte corrélation entre la fréquence d'occurrences des marques placées dans le contexte gauche d'une répétition et la fréquence de ces marques dans le corpus global : la valeur observée est de 0,914 ;
- la corrélation est également très élevée pour les marques subséquentes à la répétition, puisque la valeur est de 0,887 ;
- si la comparaison se fait sur les rangs des fréquences des marques avant et après la répétition, la corrélation est également très forte, avec une valeur de 0,837.

Dans les trois cas, le risque que ces résultats soient dus au hasard est inférieur à un millième.

5.4 Analyse morphosyntaxique des répétitions

Tout mot, appartenant à n'importe quelle catégorie grammaticale, peut théoriquement être l'objet d'une répétition. De plus, comme nous avons eu déjà l'occasion de le voir, le répétable peut aussi bien être un mot unique qu'une suite de plusieurs mots. Cette séquence peut constituer un groupe

syntactique complet ou celui-ci peut être inachevé. Les deux exemples suivants illustrent ces deux derniers cas de figure :

chaBR1 ah oui hein / elle était venue au monde dans cette maison-là elle était morte à quatre-vingt-quatre ans dans la même maison / dans la même maison / et et / et ses parents faisaient déjà café et elle est venue au monde dans le café (...) [chaBR1r]

iljDV1 redondant / eh ben c'est parfois à force de vouloir donner / une information / précise par billet on a tendance à se répéter euh / et à // à répéter un peu trop souvent la même chose dans dans le même dans le même billet / je crois que c'est pas c'est c'est un souci de se faire comprendre (...) [iljDV1r]

Or, si tout mot et si tout type de séquence peut être répété, nous avons vu dans ce qui précède que certaines régularités émergeaient, qui concernent aussi bien le répétable que le(s) répété(s). Nous avons notamment mis en évidence à plusieurs reprises la forte fréquence des mots grammaticaux concernés par les répétitions.

Nous avons donc voulu analyser plus en détail les répétitions, en ce qui concerne leur catégorie grammaticale. Pour cela, nous avons analysé de manière approfondie un sous-corpus d'un peu plus de 50 000 mots, composé de sept entrevues (appelé ci-après *Corpus Rimouski*)³². Nous avons obtenu 1151 séquences répétitions. La distribution de celles-ci correspond à celle que nous avons observée pour notre corpus :

- les répétables d'un seul mot sont de loin les plus fréquents, avec 73,3 % des occurrences ;
- les occurrences les plus nombreuses sont celles où le nombre *n* de répété égale 1. On obtient effectivement 88,8 % des cas sur ce schéma.

On constate donc une indéniable régularité dans le format général de la répétition, puisque ces pourcentages sont pratiquement identiques à ceux que nous avons obtenus sur notre corpus, où l'on avait 74,19 % de répétables

³² Cette analyse a été présentée lors du colloque international « Description linguistique pour le traitement automatique du français » qui s'est déroulé dans le cadre du Congrès de l'ACFAS, en mai 2003 à Rimouski. La communication était intitulée : « j'ai l'impression parfois qu'E: les le voca/ le vocabulaire enfin les les / (toux) / le langage est confus ».

Les sept entrevues semi-dirigées n'ont pas toutes été reprises dans le corpus que nous avons constitué ici, principalement pour des raisons de qualité sonore (leur médiocre qualité n'aurait pas permis une exploitation dans des études prosodiques). Néanmoins, cela n'invalide pas les résultats que nous avons obtenus pour cette étude ; ils permettent même de mettre en évidence une certaine régularité du phénomène.

d'un seul mot et 87,44 % de séquences répétitions où $n=1$. Cette régularité correspond également à celle observée sur le français par Blanche-Benveniste (2003b) Candéa (2000a), Grosjean et Deschamps (1972), Henry (2001 et 2005) ou encore Jeanjean (1984).

Nous voudrions maintenant analyser de façon plus approfondie cette classe de répétables composés d'une seul mot.

5.4.1 Catégorie grammaticale des répétables

Le *Corpus Rimouski* comprend 727 répétables où $l=1$. La ventilation de ces formes en fonction de leur catégorie grammaticale peut être synthétisée dans les deux tableaux suivants, le premier reprenant le nombre d'occurrences et le second donnant les pourcentages :

	n=1	n=2	n=3	n=4	TOTAL
Préposition	163	18	1	2	182
Prepdet ³³	14	2	4	1	21
Déterminant	143	21	4	1	168
Pronom personnel	86	8			94
Pronom relatif	34	4	2		40
Pronoms divers	31	1	1		33
Conjonction de subordination	42	3			45
Conjonction de coordination	45	7	1		53
Onomatopée	4	1	1		6
Adverbe	130	29	2	1	162
Nom	6				6
Adjectif	1				1
Verbe conjugué	21	1	1		23
Verbe à l'infinitif	2				2
amorces	5				5
	727	95	17	5	841

Tableau 20. Catégorie grammaticale des répétables de longueur 1 (valeur absolue)

³³ Il s'agit des contractions (préposition + déterminant).

	n=1	n=2	n=3	n=4	Total
Préposition	19,38 %	2,14 %	0,12 %	0,24 %	21,64 %
Prepdet	1,66 %	0,24 %	0,48 %	0,12 %	2,50 %
Déterminant	17,00 %	2,50 %	0,48 %	0,12 %	19,98 %
Pronom personnel	10,23 %	0,95 %	0,00 %	0,00 %	11,18 %
Pronom relatif	4,04 %	0,48 %	0,24 %	0,00 %	4,76 %
Pronoms divers	3,69 %	0,12 %	0,12 %	0,00 %	3,92 %
Conjonction de subordination	4,99 %	0,36 %	0,00 %	0,00 %	5,35 %
Conjonction de coordination	5,35 %	0,83 %	0,12 %	0,00 %	6,30 %
Onomatopée	0,48 %	0,12 %	0,12 %	0,00 %	0,71 %
Adverbe	15,46 %	3,45 %	0,24 %	0,12 %	19,26 %
Nom	0,71 %	0,00 %	0,00 %	0,00 %	0,71 %
Adjectif	0,12 %	0,00 %	0,00 %	0,00 %	0,12 %
Verbe conjugué	2,50 %	0,12 %	0,12 %	0,00 %	2,73 %
Verbe à l'infinitif	0,24 %	0,00 %	0,00 %	0,00 %	0,24 %
amorces	0,59 %	0,00 %	0,00 %	0,00 %	0,59 %
	86,44 %	11,30 %	2,04 %	0,60 %	100,00 %

Tableau 21. Catégorie grammaticale des répétables de longueur 1 (pourcentage)

Les données ci-dessus confirment la tendance (règle ?) que nous avons déjà mise en évidence à plusieurs reprises : lorsque le répétable est constitué d'un seul mot, il appartient dans la majorité des cas à la métacatégorie des mots grammaticaux. En effet, si on laisse les amorces de côté (dont il est parfois incertain de « prédire » la catégorie grammaticale du terme entier, cf. chap. 1 sur les conventions de transcription), seules 23,06 % des occurrences appartiennent à la catégorie des mots lexicaux. Il s'agit majoritairement, dans ce groupe, de répétables qui appartiennent à la classe des adverbes (19,26 %).

En ce qui concerne les mots grammaticaux, on voit que les positions de tête, en termes de fréquence, obtiennent pratiquement des scores

équivalents : les prépositions sont légèrement les plus fréquentes (21,64 %), suivies par les déterminants (19,98 %).

Les autres catégories obtiennent des scores plus faibles. Néanmoins, les pronoms, que nous avons répartis dans le tableau en trois sous-catégories, totalisent un score équivalent aux deux catégories précédentes, avec 19,86 % des occurrences.

Nos résultats vont tendanciellement dans le même sens que ce que relevaient déjà certains chercheurs. Des divergences apparaissent néanmoins liées notamment à des facteurs déjà évoqués plusieurs fois qui tiennent à une manière différente d'effectuer les décomptes ou à une définition différente du phénomène. Ainsi, Blanche-Benveniste (1987) cite une étude de Jeanjean dans laquelle 97 % des répétitions concernent des pronoms clitiques sujets (50%), des prépositions (23 %), des articles (17 %) et des conjonctions (7 %).

Grosjean et Deschamps (1975) comptabilisent quant à eux 70 % de « mots grammaticaux ou groupes de mots grammaticaux ». On regrette que les auteurs ne donnent pas le détail chiffré dans la répartition des mots grammaticaux, ni la distinction de catégorie grammaticale, entre déterminant et pronom par exemple pour *le*, *la* et *les*³⁴.

Candéa (2000a) montre que le phénomène concerne quatre fois plus souvent un mot outil (MO) qu'un mot plein (MP). Cette constatation l'incite à penser que

les répétitions de MO en français oral ont un caractère systématique, alors que les répétitions de MP ont un caractère plutôt accidentel. (Candéa 2000a : 316)

Henry (2005), dans un travail centré exclusivement sur les répétitions, a analysé un corpus d'un million de mots qui fait intervenir 1200 locuteurs. Son étude quantitative sur 15 786 répétitions montre que les répétitions impliquant des mots outils sont 5 fois plus nombreuses que celles comportant un mot plein³⁵.

Henry (2002) obtient des scores différents des nôtres, avec à l'intérieur de la catégorie des mots grammaticaux, 41,5 % de déterminants, 26 % de pronoms et 13 % de prépositions. En fait, cette différence s'explique notamment par le fait que l'auteure ne tient pas compte uniquement des

³⁴ Grosjean et Deschamps insistent également sur le nombre important de mots grammaticaux concernés par l'allongement et les faux départs (cf. leur terminologie).

³⁵ L'auteure ne s'attache qu'aux répétables qui totalisent aux moins 100 occurrences. Nous avons vu que les hapax étaient plus nombreux parmi les mots lexicaux, ce qui explique peut-être ses résultats.

séquences où la longueur du répétable est de 1, mais aussi des séquences où le répétable est composé de plusieurs mots.

Grosjean et Deschamps, dans leur étude de 1972, relèvent :

(...) le rôle majeur joué par les mots grammaticaux. Il faut aussi mentionner leur fréquence d'occurrence (où l'on relève par ordre d'importance les mots suivants : *le/la/les, de, je, des, un, et, à, il[s]*)³⁶ est fortement corrélée ($r = 0,80$) avec la fréquence d'usage des mots français (...). (Grosjean et Deschamps 1972 : 154)

Cette remarque va dans le sens de celle de Morel et Danon-Boileau (1998 : 84) qui constatent :

On ne sera pas surpris de constater que ceux-ci [les mots-outils les plus fréquemment répétés] se classent parmi les soixante-quinze premiers mots de la liste de fréquence du français fondamental (établie en 1950 par G. Gougenheim *et alii*).

Nous avons déjà remarqué ce fait lors de l'analyse des répétables les plus fréquents du corpus.

En ce qui concerne ces formes très fréquentes, nous avons voulu voir dans notre corpus comment se fait la répartition entre les formes possiblement ambiguës dans les catégories grammaticales déterminant et pronom. Ainsi, nous obtenons des résultats comparables à ceux de Henry (2001) où les répétables *les, le* et *la* sont beaucoup plus fréquemment des déterminants que des pronoms, ce qui correspond également à la tendance observée pour ces formes ambiguës lorsqu'elles ne sont pas répétées. En effet, Vergne et Giguët (1998) dénombrent seulement 2,24 % de pronoms sur 1054 occurrences de *le, l', la* et *les*.

³⁶ Dans leur étude de 1973, les mots grammaticaux les plus répétés sont, par ordre de fréquence décroissante : *qui, le, la, les, de, un(e), sur, sa, au, à*.

5.4.2 Le répétable est une séquence de mots grammaticaux³⁷

Nous venons d'analyser les répétables composés d'un seul mot, majoritairement constitués de mots grammaticaux, principalement des prépositions et des déterminants.

Quand on y regarde de près, on constate que la plupart des répétables de deux mots contiennent également au moins un mot grammatical.

Si l'on prend l'exemple de la préposition *dans*, on a 204 occurrences où *dans* est répétable seul. À cela s'ajoutent les occurrences où il est suivi d'un déterminant : *dans ce*, *dans certains*, *dans cette*, *dans des* (4), *dans l'* (2), *dans la* (4), *dans le* (6), *dans les* (13), *dans leurs*, *dans ma* (2), *dans mes*, *dans mon* (2), *dans nos*, *dans notre*, *dans son* (3), *dans un* (5) et *dans une* (6). On a aussi des répétables plus longs comme *dans d'autres (pays) ou dans le même (billet)*.

Pour l'anglais, Blankenship et Kay (1964), ainsi que Cook (1971), ont montré qu'en général, la répétition se fait à l'initiale du syntagme. C'est cette régularité de reprise à l'initiale du syntagme qu'illustrent les exemples de *dans*.

On peut citer le même type de fonctionnement avec les autres prépositions *à*, *avec*, *entre*, *pour*, etc.

On le répète encore, le phénomène de la répétition est soumis à certaines contraintes : la répétition disfluente n'apparaît pas n'importe où dans l'énoncé, et elle concerne de manière privilégiée les mots grammaticaux monosyllabiques (cf. Blanche-Benveniste 2003b).

Cette manière de reprise à l'initiale, qui semble relever de la systématisme, explique le petit nombre d'occurrences qui ne sont pas des mots grammaticaux que l'on observait dans les répétables de longueur 1. En fait, en général, un substantif n'est pas répété seul. S'il fait l'objet d'un problème de sélection lexicale, ce problème se manifeste par la répétition des mots-outils qui le précèdent. C'est donc tout le groupe syntaxique qui est repris à l'initiale, comme l'illustrent les exemples suivants :

³⁷ L'hypothèse développée récemment par Constantin de Chanay (2005), dans une approche pluridimensionnelle – l'oral étant « un tout sémiotiquement complexe », une « totalité "verbo-voco-posturo-mimo-gestuelle" » (p. 2) – est que l'analyse traditionnellement faite de certaines répétitions en répétitions paradigmatiques ne peut être validée si l'on tient compte de facteurs non verbaux et/ou de la dimension interlocutive. Ainsi, certaines répétitions de mots-outils, interprétées comme disfluentes, font en fait progresser l'énoncé sur l'axe syntagmatique. L'auteur tire ses exemples de deux interventions télévisées, l'une de Jean-Claude Van Damme, l'autre de Nicolas Sarkozy, dans lesquelles il analyse la posture, les gestes et mimiques des locuteurs.

ilcCF1 l'accent c'est la la tonalité que l'on met sur les / sur les syllabes [ilcCF1r]

ileGG1 sur une / sur une classe de de de quinze ou vingt que nous étions il y en avait peut-être un ou deux ou trois et encore c'était en pleine Ardenne (...)

ileGF0 (rire) mais revenons à nos à nos histoires [ileDC1]

fameS1 alors on avait donné rendez-vous à un / à un autre endroit /aux aux parents / pour pas qu' ils voient les |- toilettes

ilrCP1 (...) il y a différentes euh / distances différentes à parcourir et il y a une couleur pour chaque pour chaque parcours par exemple bleu cinq kilomètres euh jaune dix kilomètres euh / rouge vingt kilomètres (...) [ilrCP1r]

Le phénomène est le même pour les verbes, où l'on remarque la reprise de la séquence à gauche de celui-ci :

ilpMJ1 (...) moi je je / je connais peu de jeunes qui qui parlent wallon de A à à Z / par contre qui utilisent des termes wallons ou des expressions et cetera je crois que ca se ca se transmet / exactement de la même manière (...) [ilpMJ1, ilpMJ1r]

ileDC1 les profs étaient obligés de redonner des cours de français / à leurs étudiants en physique / ce qui euh ce qui est assez / assez révélateur (silence) (bruits de feuilles) [ileDC1, ileDC1r]

Cela explique le nombre réduit de mots lexicaux seuls que l'on a dans nos données. Ce n'est pas sur le lexème seul que se manifeste le travail de formulation, mais sur toute l'unité dont il est la tête (le syntagme, pourrait-on dire). Comme le fait remarquer Blanche-Benveniste (1985 : 113) :

(...) on observe que les positions syntaxiques sont en place dès le début de l'énoncé, même si elles sont simplement « annoncées » (...). La répétition-hésitation concerne non pas la syntaxe elle-même, mais le remplissage lexical des positions syntaxiques.

Néanmoins, le fait que le mot grammatical soit répété ne donne pas toujours lieu ensuite à un syntagme immédiatement achevé. Ainsi l'exemple suivant :

ilrCP1 oh il y avait des des euh hum / des tests de programmes de nouveaux programmes qui étaient sortis

euh des jeux-vidéo il y avait euh / sinon qu'est-ce qu'il y
 avait / il y avait des euh pf des commentaires sur euh
un sur un comment / oui sur euh Bureau-nonante /
 parce que bon j'avais été à Bureau-nonante quoi (...)
 [IlrCP1r]

On a ici une séquence relativement complexe où la répétition repérée par le programme est *sur euh un sur un*. La suite de cette séquence, même si elle contient des formes que nous avons permises comme insertions (*comment, oui, euh* et la pause brève), n'appartient plus à la répétition au sens où nous l'avons définie. En effet, le répétable précédemment identifié est la suite préposition + déterminant *sur un*. Il manque donc le déterminant pour que la suite *sur euh Bureau-nonante* soit analysée comme le dernier répété de la séquence. Ainsi, selon notre méthode, cette séquence répétition initie un syntagme qui n'est pas achevé *immédiatement*. Il se peut aussi que le syntagme initié ne soit pas achevé du tout, même « à retardement ». On a donc une rupture syntaxique de l'énoncé. C'est ce point de la rupture syntaxique que nous allons maintenant aborder.

5.5 Continuité ou rupture syntaxique de l'énoncé

La répétition des mots-outils constitue-t-elle une rupture dans la construction syntaxique de l'énoncé ? En d'autres mots, la place syntaxique qu'ils ouvrent est-elle actualisée dans le discours ou la répétition du mot-outil marque-t-elle une auto-interruption de l'énoncé ? Selon (Morel et Danon-Boileau 1998 : 85),

La répétition multiple [$n > 1$] et sans pause du mot-outil semble relever d'une gestion à très court terme de la formulation. Elle aboutit toujours à l'énoncé d'un mot.

Dans ce cas, il n'y aurait donc pas de rupture syntaxique, la construction ouverte par le mot-outil étant au final achevée. Pour ces cas, Henry (2005 : 85) parle d'un « vide lexical » momentané, et d'un « remplissage lexical qui s'effectue dans un second temps pour aboutir à un énoncé achevé ».

À l'analyse de notre sous-corpus (*corpus Rimouski*), nous avons pu observer que 5,7 % des répétitions (soit 66 cas) étaient le lieu d'une rupture dans la construction syntaxique de l'énoncé.

Ainsi donc, dans la plupart des cas, la place ouverte par la répétition est achevée. Mais on constate des cas, comme celui mentionné ci-dessus, où la place est achevée, mais pas immédiatement. Dans le cadre de l'étiquetage morphosyntaxique, cela pose évidemment problème.

Le plus fréquemment, la rupture se produit après la répétition, comme dans l'exemple suivant :

ilrMS1 (...) je vois pas pourquoi pour euh pour aller à à
/ chez les Flamands ben il faudrait plus avoir son
 accent liégeois (...) [ilrMS1r]

Néanmoins, nous avons beaucoup d'occurrences où la rupture précède la répétition. En fait, la répétition marque le début d'une nouvelle construction syntaxique, laissant la séquence antérieure inachevée, comme dans :

norkJ1 (...) euh pf je je sais bien qu'il vaudrait mieux
 que mes enfants aient l'orthographe de // qu'on /
qu'on souhaite voir // euh dans les journaux et sur les
 murs // hein [norkJ1r]

Dans cet exemple, la rupture est marquée par la pause longue. Mais ce n'est pas toujours le cas, comme le montre l'énoncé suivant :

norkJ1 alors là évidemment -| oui / hein oui bien sûr //
 euh les la société est pour le moment ainsi faite / que /
 que les les manquements // tant sur le plan grammatical
 // euh / à ces normes qui ne sont donc pas encore très
 très bien étudiées mais dont l'évidence s'impose
 néanmoins // on les les manquements sont
 sanctionnés // (...) [norkJ1r]

Ici, on a bien une pause longue dans le contexte gauche de la répétition, mais celle-ci ne la précède pas directement comme dans l'exemple précédent. La rupture de construction se produit sans aucune marque dans le texte : la construction entamée avec le pronom sujet *on* est laissée inachevée (la séquence verbale ne sera pas actualisée, en tout cas pas suite à un pronom sujet), pour repartir sur une nouvelle construction qui débute par la répétition du déterminant. Le sujet n'est plus ici pronominal mais nominal, et il débute par une répétition à l'initiale du syntagme.

On voit donc bien, à l'analyse d'un cas comme celui-ci, que les indices manquent pour déterminer sur des bases formelles les cas d'inachèvements de construction.

6 CONCLUSIONS

Si l'on s'est intéressée un peu longuement aux répétitions, c'est parce que ces disfluences sont particulièrement fréquentes dans le corpus ; les laisser telles quelles dans le texte diminuerait considérablement les performances de l'étiqueteur morphosyntaxique. De plus, il est plus facile de les traiter dans une phase de prétraitement (cf. chap. 4) que d'intégrer leur description dans les grammaires de levée d'ambiguïtés.

Nous avons constaté que la répétition, telle que nous l'avons définie ici en tout cas, obéit à d'indéniables régularités, qu'il s'agisse du nombre de répétés, de la longueur du répétable ou encore des séquences mixtes, plus

nombreuses que les répétitions contigües. Mais malgré les régularités observées et les tendances qui se dégagent, on est face à des structures complexes et à de nombreuses variations dans notre corpus

Nous avons vu également que la répétition affecte principalement les mots grammaticaux (des déterminants, des prépositions et des pronoms) ou des suites de ceux-ci. En fait, il est capital de les traiter correctement car nous verrons que les mots grammaticaux sont de bonnes bases pour la construction de grammaires de levée d'ambiguïtés (chap. 6).

Néanmoins, malgré les régularités observées, nous échouons à trouver des indices qui nous permettraient de détecter les cas où la répétition est le lieu d'une rupture syntaxique dans la construction de l'énoncé. Ce problème est crucial, et nous y reviendrons pour les autres disfluences et également dans le chapitre concernant l'étiquetage.

CHAPITRE 3. SECTION 3

L'AUTOCORRECTION IMMÉDIATE

accFJ1 bè c'est-à-dire que euh / oui le le
les Liégeois // euh // ont un ac/ ont ont
une euh comment dirais-je ont un contact
sont généralement des gens chaleureux
(...) [accFJ1r]

1 DÉFINITION

Nous appelons *autocorrection (immédiate)* le phénomène langagier qui consiste pour un locuteur à énoncer un morphème suite à un autre morphème différent qui appartient à la même catégorie grammaticale. Ce deuxième morphème vise à corriger le premier morphème énoncé.

Une fois de plus, la terminologie est loin d'être unifiée, et l'on rencontre dans la littérature le terme d'*autocorrection* pour n'importe quelle séquence dans laquelle le locuteur se corrige, indépendamment de la forme que peut prendre cette correction. Fornel et Marandin (1996) utilisent quant à eux le terme *auto-réparation* ; Bénard (2005) emploie le terme *révision* ; Grosjean et Deschamps (1972) préfèrent *faux-départ*, équivalent de l'anglais *false-start*. Candéa (2000a) distingue quant à elle *autocorrections immédiates* et *faux départs*, ceux-ci étant plus complexes et concernant aussi des mots pleins ou des structures syntaxiques inachevées.

Le fonctionnement de l'autocorrection immédiate telle que nous l'avons définie est très proche de celui de la répétition, à la différence qu'il n'y a pas d'identité formelle entre répétable et répété(s). En effet, comme dans le cas des répétitions, la place syntaxique qu'occupe l'élément remplaçant (*repair*, selon la terminologie de Shriberg (1994)) est la même que celle de l'élément remplacé (*reparandum*).

Voici un exemple type d'autocorrection :

ilePA2 (...) je me dis après tout l'accent c'est un peu **le**
la carte d'identité c'est un peu l'appartenance à une
région (...) [ilePA2r]

Dans cet énoncé, le locuteur remplace le déterminant défini masculin par sa variante au féminin, mieux adaptée au lexème qu'il prononce ensuite.

Parfois, il y a non pas remplacement total du *reparandum* par un autre morphème, mais « aménagement » phonétique de celui-ci. C'est le cas dans l'énoncé suivant où le déterminant reste le même mais est présent une fois sous sa forme pleine avant d'être repris sous sa forme élidée :

iljVM1 parce que c'est vrai que (silence) chez nous c'est plutôt les vieux / qui parlent wallon mais beaucoup de jeunes parlent wallon aussi mais ils ont euh moins **le** l'occasion de le parler [iljVM1r]

On rencontre ce cas pour l'élision du déterminant, mais aussi pour l'élision du pronom *que* ou encore de la conjonction :

iljVM1 (...) dans la majorité des cas euh il a une fluidité verbale **que qu'**on n'a pas ailleurs (...) [iljVM1r]

Ces exemples n'avaient pas été répertoriés dans les répétitions du fait de l'absence d'identité formelle entre répétable et répété. C'est la raison pour laquelle ils seront analysés ici.

Dans cette section, nous nous centrerons uniquement sur les autocorrections de certains mots grammaticaux.

Nous réduisons donc encore le champ d'application du terme *autocorrection*. Ce choix s'est imposé pour plusieurs raisons :

- nous faisons l'hypothèse d'un fonctionnement comparable à celui des répétitions, et nous avons vu la forte fréquence des mots grammaticaux dans ce type de séquences ;
- nous voulions pouvoir repérer automatiquement ces séquences, et étendre la recherche à tout type de réparation (en termes de catégorie grammaticale et/ou de structure syntaxique) aurait rendu non seulement l'automatisation du repérage impossible, mais aurait engendré un bruit ingérable qui, sur un corpus de la taille du nôtre, n'aurait permis aucune analyse sérieuse ;
- se centrer sur une catégorie particulière d'autocorrections, qu'il est possible de repérer automatiquement, permet également d'envisager leur formalisation dans le cadre des grammaires de levées d'ambiguïtés et de les intégrer dans une description unifiée avec les répétitions.

C'est donc principalement pour des raisons pratiques que nous nous limitons ici aux mots grammaticaux, gardant en tête l'objectif final : la

possibilité d'analyse de phénomène en vue d'une annotation semi-automatique des corpus.

Ainsi, nous ne nous intéresserons pas ici aux formes d'autocorrections suivantes :

a) *celles qui concernent la correction du lexème*

iljDV1 mais je dirais dans / m dans la plupart des des régions très industrielles et / et très pauvres so/ socio-économiquement // où euh il y a un mélange de cultures où il y a euh énormément d'immigrés **qui ont** // **qui ne sont pas** parvenus à / maîtriser correctement la langue (...) [iljDV1r]

Nous avons déjà abordé dans la section consacrée aux répétitions ce type d'exemples, qui relèvent d'une recherche sur la dénomination (cf. Blanche-Benveniste 1985).

Ces réparations, qui concernent le niveau lexical, sont appelées *lemma substitutions* par Levelt (1983). En fait, comme le signale l'auteur, elles doivent être analysées davantage en termes sémantiques qu'en termes proprement syntaxiques. Elles sont par ailleurs impossibles à formaliser, de manière à être repérées automatiquement.

b) *celles qui concernent une correction mettant en jeu des mots grammaticaux, mais qui ne sont pas en contigüité.* Les autocorrections que nous avons choisi d'étudier sont donc très locales. Dans l'exemple suivant, le locuteur corrige en modifiant le pronom sujet (*on* à la place de *je*), mais en ayant déjà énoncé le verbe :

ilrLD3 je réserve bon je vais chez / quand j'ai fini mes devoirs je vais chez mon amie et tout bon / **je vais on** **va** au cinéma ou boire un verre ou des trucs comme ça [ilrLD3r]

Dans l'énoncé suivant, c'est à une correction du déterminant que se livre le locuteur, *quelques* remplaçant le possessif *mes* :

ileDC1 (...) bon je je plaisante **ce sont mes ce sont quelques** collègues féminines qui disent ça (...) [ileDC1r]

En termes d'analyse, les deux cas que nous regroupons ici sont très différents. En effet, dans le premier, le groupe constitué par le pronom sujet et le verbe est syntaxiquement complet, aussi bien dans la séquence initiale que dans la séquence réparée. Ce n'est pas le cas dans le second énoncé où le syntagme initié par le déterminant possessif est laissé inachevé, entraînant une rupture syntaxique entre *reparandum* et *repair*.

- c) **les changements de temps pour les verbes.** Dans l'énoncé suivant, à deux reprises, la locutrice corrige le temps de son verbe : la première fois préférant un imparfait au présent (*allait* à *va*), et la deuxième fois dans un mouvement contraire (*est* à *était*). On remarque ici un fonctionnement régulier que nous retrouverons dans les amorces de morphèmes : dans la séquence corrigée, qui concerne le verbe, le pronom sujet est également repris :

magMM1 mais on allant on allait regarder les magasins
/ on ira voir les magasins ou on ira voir les vitrines //
voilà / on va on allait voir les vitrines // et il y avait
du il y avait déjà des¹ des des magasins il y avait des
des vêtements il y avait des chaussures et cetera /
c'était c'est c'est déjà comme m/ comme maintenant
(...) [magMM1r]

On pourrait encore citer d'autres types d'autocorrections, qui fonctionnent sur des schémas différents les uns des autres. Ce n'est pas par manque d'intérêt que nous limitons notre propos ici, mais pour des raisons de faisabilité : le phénomène de l'autocorrection envisagé sous toutes ses réalisations effectives nécessiterait en effet un travail à part entière².

Nous avons donc décidé de nous focaliser particulièrement sur certains mots-outils qui initialisent un syntagme, verbal ou nominal.

Dans le cas des autocorrections que nous avons choisi d'aborder ici, à la différence de certaines répétitions pour lesquelles il est parfois très difficile, même en recourant à la prosodie, de trancher pour savoir si on a affaire à une disfluence ou une poursuite sur l'axe syntagmatique, il apparaît clairement que nous sommes dans le premier cas de figure.

2 LES DIFFÉRENTS TYPES D'AUTOCORRECTIONS RETENUS

Les autocorrections que nous allons aborder ici concernent les déterminants (définis, indéfinis, possessifs et démonstratifs), les pronoms sujets et les pronoms relatifs. Nous aborderons également la question des prépositions et des suites prépositions + déterminants.

¹ Cette réparation est équivalente à celle de l'énoncé précédent : le syntagme ouvert par la forme *du* est laissé inachevé.

² Cf. l'article issu de la thèse de Martinie (2001) sur les énoncés réparés.

2.1 Les déterminants

2.1.1 Les déterminants définis : *le, la, les, l'*

Nous avons recherché toutes les séquences où cooccurrent des déterminants définis. Comme pour la recherche des répétitions, nous avons prévu l'insertion d'éléments, à savoir : *ah, bè, ben, boh, bon, comment, disons, enfin, euh, ha, hein, hum, m, mh, mm, non, oh, oui, ouais, pf, pff*, la pause brève (/), la pause longue (/ /), le silence (*silence*) et l'amorce de morphème.

Les 210 séquences d'autocorrection immédiates du déterminant défini se ventilent comme suit dans le tableau 22. La colonne en abscisse indique la forme initialement énoncée par le locuteur (= la forme au début de la séquence disfluente), la colonne en ordonnée représente la forme qui constitue la correction (= la fin de la séquence disfluente avec la reprise du déroulement sur l'axe syntagmatique).

Nous présentons les résultats en pour cent, et également, entre parenthèses, en nombre d'occurrences.

	<i>l'</i>	<i>le</i>	<i>la</i>	<i>les</i>	TOTAL
<i>l'</i>	0,48 % (1)	1,44 % (3)	0 % (0)	1,44 % (3)	3,37 % (7)
<i>le</i>	19,71 % (41)	0,48 % (1)	20,19 % (42)	13,46 % (28)	53,85 % (112)
<i>la</i>	8,17 % (17)	13,94 % (29)	0,48 % (1)	4,81 % (10)	27,40 % (57)
<i>les</i>	2,88 % (6)	6,25 % (13)	5,77 % (12)	0,48 % (1)	15,38 % (32)
TOTAL	31,25 % (65)	22,12 % (46)	26,44 % (55)	20,19 % (42)	100 % (208)

Tableau 22. Autocorrections du déterminant défini

Des données présentées dans ce tableau, nous pouvons dégager certaines tendances générales. En effet, mise à part la correction du *l'* en *la*, le corpus atteste toutes les possibilités de corrections du déterminant défini par un autre déterminant défini. Nous verrons ci-dessous que ce n'est pas le cas pour les autres sous-catégories de déterminants dans notre corpus.

Si toutes les possibilités sont représentées, c'est néanmoins en fréquence variable. La forme *le* est celle sur laquelle porte le plus fréquemment

l'autocorrection, avec près de 54 % des occurrences. Vient ensuite le *la* (27,40 %), le *les* (15,38 %) et enfin le *l'*, concerné par seulement 7 occurrences. Le petit nombre de cas de *l'* s'explique selon nous par les choix de transcription. En effet, nous formulons l'hypothèse que dans ces cas, le transcritteur a noté l'élision (ou ce qui pourrait en tout cas être interprété raisonnablement comme une élision) non pas en tant que telle mais comme une amorce de morphème – ce qui explique notamment la forte représentation de ce type d'amorce, corrigée immédiatement par la forme considérée comme pleine (cf. chap. 3.4). Ceci explique aussi sans doute qu'on n'a pas de répétitions du *l'*.

On a ainsi :

accBF1 non / mais / de par les Français qui / qui
imitent // l'accent bruxellois // ils disent toujours euh
une fois mais si/ sinon [accBF1r]

En fait, toutes les occurrences de *l'* prises en compte ici sont suivies d'une amorce de morphème, comme le montre l'exemple suivant :

norGA1 donc l'en/ le maître est important mais il n'est
pas tout seulement un mauvais maître ben on perd son
temps [norGA1r]

Le contexte droit du son [l] orienterait ainsi les choix de transcription³ : forme élidée quand une amorce de morphème suit, amorce de morphème dans tous les autres cas.

En ce qui concerne le *repair*, la répartition des données en termes de fréquence est moins inégale que pour la forme qui subit la correction (le *reparandum*). On constate néanmoins une préférence pour la forme élidée, avec 31,25 % des séquences qui s'achèvent finalement par cette forme. Viennent ensuite *la* (26,44 %), *le* (22,12 %) et enfin *les* (20,19 %).

³ Notons qu'on n'a aucune indication de ce type dans les conventions de transcriptions de VALIBEL, mais il semble bien que tous les transcritteurs aient procédé de la même manière au fil des années.

On a aussi de nombreux exemples où, à la seule lecture de la transcription, on peut se demander si un autre choix n'aurait pas été possible (autre choix peut-être invalidé par des indices prosodiques que nous ne possédons pas). Ici, la séquence *la la l'accent* aurait peut-être pu être transcrite *l'a/ l'a/ l'accent*. Tout au long de ce travail, nous avons adopté le parti pris de faire confiance à la transcription. La séquence est donc comptabilisée ici dans les autocorrections.

accPH1 parce que la la l'accent de la mer belge // je
comprends // si vous allez / tout doucement je comprends mais
ici elle pouvait parler tout doucement comme tu |- veux
[accPH1r]

Si l'on se centre non plus maintenant sur les formes en tant que telles mais bien sur le nombre du déterminant, on constate que le singulier est largement dominant, aussi bien en ce qui concerne la forme de départ (84,62 % des occurrences) que la forme choisie au final (77,88 %). Néanmoins, lorsqu'il y a changement de nombre dans la réparation, on se rend compte que celui-ci concerne plus souvent un passage du singulier vers un pluriel (19,34 %) que l'inverse (14,9 %).

L'autocorrection peut également consister, pour le locuteur, à ajuster le genre de son déterminant au substantif qu'il va énoncer.

Plusieurs hypothèses ont été avancées en ce qui concerne les hésitations sur le genre. Pour Sauvageot (1962 et 1972), il s'agirait d'une hésitation dans le chef du locuteur, qui hésite sur le genre du mot qui va suivre. Cappeau (1998) remet en cause cette hypothèse :

Le genre propre au nom impose la forme du déterminant qui le précède. D'où la nécessité de se livrer à une « analyse anticipative » et les risques de conflit qui peuvent conduire à ces ratés. (...) [les exemples cités] concernent des termes courants qui ont peu de chance d'être mal maîtrisés par les locuteurs. L'explication par une méconnaissance du genre serait donc peu plausible. (Cappeau 1998 : 305)

Au vu des substantifs concernés dans nos exemples, on peut difficilement suivre cette hypothèse.

Outre les formes *le* corrigées en *la* (20 % des occurrences) et inversement (près de 14 %), on a les corrections qui concernent la forme élidée. Sous cette forme, le genre du déterminant est indistinct. Dans certaines corrections, il s'agit non pas d'un changement de genre mais simplement d'une adaptation phonétique : le locuteur a énoncé un déterminant, dont le genre correspond bien à celui du substantif qu'il anticipe, mais qui doit en fait être élidé puisque celui-ci commence soit par une voyelle, soit par un *h* aspiré :

ilpCM1 (...) je vais pas faire ici **le l'**hypocrite (...)
[ilpCM1r]

Mais dans d'autres cas, le déterminant élidé n'est pas du même genre que celui du substantif qui le précède :

ileGF0 mm mais vous êtes quand même si si je vous suis bien donc **le l'**écriture est un un moyen de traduction euh de la réalité [ileGG1r]

On a un processus comparable pour la forme corrigée *les*. Le locuteur corrige le nombre, passant d'un singulier à un pluriel, sans qu'il y ait

nécessairement correction du genre. Dans l'énoncé suivant, on passe d'un masculin singulier à un masculin pluriel :

ilrDT1 je ne sais pas moi (silence) mais tous tous l/ les
coins de Belgique où **le / les** revenus sont un peu pluS
élevés [ilrDT1r]

Mais la correction peut porter non seulement sur le nombre, mais également sur le genre.

Pour ces autocorrections liées au genre, Cappeau parle d'une séquence

qui fait se succéder la faute et la réussite. (Cappeau
1998 : 305)

Dans l'énoncé suivant, le locuteur entame le syntagme avec un déterminant féminin singulier pour le corriger en un masculin pluriel :

ileAO1 vous verrez que **la les** trois quarts de la
rédaction ont déjà quitté les lieux [ileAO1r]

Dans notre corpus, les résultats pour les corrections en *l'* et *les* peuvent être synthétisés comme suit :

la → *l'* : 10 adaptations phonétiques, 7 changements de genre ;

le → *l'* : 23 adaptations phonétiques, 18 changements de genre ;

la → *les* : 3 féminins, 5 masculins (changement de genre) ;

le → *les* : 16 masculins, 9 féminins (changement de genre).

On le voit, hormis pour le passage du *la* en *les*, les corrections qui aboutissent au déterminant éliminé ou à *les* consistent plus souvent à conserver le même genre qu'à le modifier. Néanmoins, les résultats entre ces deux « stratégies » ne sont pas véritablement tranchés puisqu'on a au total 57,14 % d'occurrences pour lesquelles on garde le même genre, contre 42,86 % des cas dans lesquels le genre est modifié.

C'est pour le masculin que le genre est le plus souvent conservé, avec 59 % des occurrences, tandis que pour le féminin, le score tombe à une conservation du genre dans 52 % des cas.

Avec ces cas de changement de genre, ajoutés aux changements déjà repérés dans le tableau ci-dessus, on comptabilise 110⁴ occurrences dans lesquelles la correction consiste au moins en un changement de genre (avec parfois aussi un changement de nombre), soit près de 53 % du total des occurrences d'autocorrections. Ces changements de genre montrent une nette préférence pour le changement qui va du masculin vers le féminin

⁴ 42+29+7+18+5+9=110.

(62,72 %), les corrections aboutissant au déterminant *la* étant moins nombreuses (37,27 %).

On voit donc ici une nette préférence pour le masculin (au départ comme à l'arrivée). On a également plutôt un passage du pluriel au singulier que l'inverse. Néanmoins, cette forte tendance n'est pas une généralité puisque de nombreux cas attestent le mouvement inverse. Nous ne pouvons donc aller dans le sens de la conclusion à laquelle aboutit Candéa (2000a : 357), qui affirme que

toutes ces autocorrections [à part une exception] vont du
plus neutre au *plus spécifié* (masculin → féminin,
présent → non présent, singulier → pluriel).

L'hypothèse sous-jacente est que la première apparition de la séquence serait en quelque sorte automatique, rectifiée ensuite pour aller du plus général (*grosso modo* le masculin singulier), au plus spécifique (le féminin et/ou le pluriel). Cela va dans le sens de Cappeau (1998 : 307) :

La prédominance du masculin pourrait être due au
caractère non marqué de ce genre : le locuteur qui
cherche un terme l'utiliserait donc en priorité.

Cela dit, le total de 53 % de changement de genre que l'on obtient est finalement faible en regard de ce à quoi on pouvait s'attendre. Le résultat qui nous semble finalement intéressant ici est le nombre important d'occurrences qui participent d'une adaptation morphologique/phonétique.

Dans le tableau 22 donné ci-dessus, nous avons considéré la forme de départ et la forme d'arrivée de la séquence. On a donc pu s'étonner de trouver dans les autocorrections un *le reparandum* et un *le repair* au final de la séquence, comme s'il s'agissait d'une répétition. C'est en fait parce que nous avons recherché les séquences disfluentes maximales, c'est-à-dire qui s'achèvent avec la réparation. Or, il arrive que cette réparation ait la même forme que le terme de départ, avec entre les deux un changement de déterminant. C'est le cas de l'exemple suivant :

ileGG1 euh / finalement le / la le contenu est quand
même servi par la forme [ileGG1r]

Néanmoins, ces cas de retour au déterminant initial sont rares, puisque l'on n'en rencontre que 4 occurrences dans le corpus.

Nous avons ici examiné les formes *le*, *la*, *les* et *l'* appartenant à la catégorie des déterminants. Nous avons donc fait un tri manuel puisque ces formes peuvent aussi être des pronoms.

En fait, on constate qu'on n'a que deux occurrences d'autocorrections qui concernent ces formes lorsqu'elles sont pronoms. On a donc dans notre

corpus la même tendance que pour les répétitions, avec une écrasante majorité en faveur de la catégorie des déterminants par rapport à celle des pronoms pour les formes potentiellement ambiguës.

Les deux occurrences recensées dans le corpus suivent la même structure : le pronom *le* est corrigé immédiatement en un *l'* suivi d'un verbe à l'infinitif qui débute par une voyelle (*le l'entretenir* et *le l'enseigner*).

2.1.2 Les déterminants indéfinis

Pour les déterminants indéfinis, nous sommes directement confrontée au problème de l'ambiguïté virtuelle : une suite de deux formes qui sont possiblement deux déterminants indéfinis en langue ne sont pas nécessairement des déterminants indéfinis en discours. Si l'on cherche ainsi une suite de deux déterminants indéfinis, dans un texte non désambiguïsé (qui est le format de texte sur lequel nous travaillons actuellement), on recense des exemples comme les suivants :

chaGG0 et vous / vous souvenez **d'un** problème d'héritage qu'il y aurait eu [chaSR1r]

chaBR1 et voilà voilà **un des** enterrements où il y avait beaucoup des gens à cet enterrement-là (...) [chaBR1]

puisque *un*, *d'* et *des* sont possiblement des déterminants.

En fait, les exemples ci-dessus ne participent pas du tout de la disfluence. Nous rencontrerons le même problème lié à l'ambiguïté pour les prépositions et afin de travailler sur des autocorrections certaines, nous avons décidé de nous centrer uniquement sur les autocorrections de *un* en *une* et inversement. Les résultats sont les suivants :

	un	une
un	1	41
une	21	2

Tableau 23. Autocorrections des déterminants indéfinis

Ici, la tendance que l'on observait pour le déterminant défini, en ce qui concerne le genre, est encore accentuée. En effet, 63 % des occurrences montrent une correction du masculin en féminin, alors que 32,3 % des items manifestent le passage inverse.

On a trois cas où l'autocorrection aboutit finalement à la forme de départ : une occurrence concerne le masculin et deux le féminin. On ne peut tirer de conclusion d'un échantillon si réduit.

Si avec les suites *un(e) des* on ne peut garantir que l'on a affaire au phénomène de l'autocorrection, ce n'est pas le cas des séquences *des un* et *des une* (le deuxième déterminant étant au singulier). On obtient les scores suivants :

- *des un* : 9 occurrences
- *des une* : 1 occurrence

Étant donné que nous n'avons pas relevé les occurrences d'autocorrections qui vont du singulier vers le pluriel, on ne peut faire de comparaison. Néanmoins, ces cas relevés ici seront pris en compte lors du prétraitement (cf. chap. 4).

2.1.3 les déterminants possessifs

Les cas d'autocorrections qui portent sur la sous-catégorie des déterminants possessifs sont beaucoup plus rares puisqu'on n'a que 16 occurrences. Aucune de ces formes n'apparaît d'ailleurs ni dans les 100 formes les plus fréquentes du corpus, ni dans les 30 mots grammaticaux les plus fréquemment l'objet d'une répétition.

Le détail de ces déterminants corrigés est le suivant : *ma – mon* (2), *mes – ma* (4), *mon – ma* (3), *mon – mes* (1), *sa – ses* (1), *ses – son* (1), *son – sa* (2), *tes – ton* (2).

On ne peut guère tirer de conclusions sur un si petit nombre d'occurrences. On ne peut non plus faire de comparaison avec les autres déterminants du point de vue du genre, étant donné que, à part *ma*, *ta* et *sa*, les autres formes sont épicènes.

Néanmoins, on peut faire trois remarques :

- forme initiale et forme finale appartiennent toutes à la même sous-catégorie : on ne change pas le possesseur (pas d'exemples du type **mon ton*) ;
- on n'a pas de déterminant dont le possesseur est pluriel (*notre*, *votre*, *leur*, *nos*, *vos*, *leurs*). On ne voit pas bien comment nos données auraient pu déterminer cela, et selon nous il ne s'agit pas d'un « effet du corpus ». Néanmoins, ceci nous semble un point qui mériterait d'être vérifié sur d'autres données ;
- on n'a ici aucune rupture syntaxique ; après le déterminant corrigé, le syntagme s'achève normalement.

2.1.4 Les déterminants démonstratifs

Les autocorrections de déterminants démonstratifs sont à peine plus nombreuses que celles des possessifs, avec 18 occurrences, qui se répartissent comme suit : *ce – cet* (3), *ce – cette* (5), *ce – ces* (4), *ces – cette* (1) et *cette – ce* (5).

Pour les démonstratifs, on a aussi des autocorrections liées à un aménagement phonétique, lorsque le *ce* est corrigé en *cet*.

En ce qui concerne le changement de genre, on a 7 occurrences pour le passage du masculin au féminin, et 5 dans l'autre sens. Les occurrences sont trop peu nombreuses pour tirer des conclusions, mais on voit néanmoins que ces résultats montrent une tendance comparable à celle observée pour les déterminants définis.

Dans les occurrences recensées ici, on constate que la réparation n'est pas le lieu d'une coupure syntaxique de l'énoncé, mais que celui-ci poursuit son déroulement syntagmatique après le *repair*.

2.1.5 Les changements de sous-catégorie de déterminants

On a vu pour les déterminants possessifs que la réparation se faisait à l'intérieur d'une même sous-catégorisation en ce qui concerne le possesseur : un syntagme initié par un déterminant dont le possesseur est à la première personne du singulier est réparé par un déterminant dont le possesseur est également à la première personne du singulier, et ainsi de suite.

En fait, nous avons ici cherché les corrections à l'intérieur d'une même sous-catégorie (définis, indéfinis, démonstratifs et possessifs), sans autoriser la possibilité de croisement. En faisant des recherches de ce type sur le corpus, on se rend compte que la récolte est très médiocre, comme le font remarquer Morel et Danon-Boileau :

(...) dans la classe des déterminants du nom, seules sont possibles des corrections de genre ou (plus rarement) de nombre. Dans les cas exceptionnels où l'on finit par changer de classe, ce changement nécessite la présence de marques supplémentaires. (Morel et Danon-Boileau 1998 : 86)

Nous n'avons donc pas exploré davantage cette piste, qui aurait donné une série de séquences dont on n'aurait pu tirer un enseignement sur le fonctionnement global, nous limitant aux mêmes sous-catégories de déterminants.

2.2 Les pronoms sujets

Afin d'éviter d'introduire des erreurs lors du prétraitement, nous avons dû faire le départ entre les pronoms sujets ambigus et ceux qui ne le sont pas, selon la position qu'ils occupent dans l'énoncé. En effet, les formes *elle*, *elles*, *nous* et *vous* peuvent également être des pronoms compléments. On peut donc trouver les suites de pronoms *je vous* ou *il nous* sans qu'il s'agisse d'une autocorrection mais bien d'un pronom sujet suivi d'un pronom complément. Par contre, si l'on a *je elle* ou *il elle*, on est bien dans l'autocorrection, puisque le *elle* complément ne peut se trouver dans cette position entre le sujet et le verbe. En effet, en tant que complément, le pronom *elle* suit une préposition, comme l'illustrent les deux exemples suivants :

blaCD1 ah s/ je ne sais pas ou elle était mais / elle était pas chez **elle je** crois |- (x) [blaJD1l]

ilrLD3 (...) mais elle est au lycée pourquoi parce que / avec **elle je** sais discuter et tout (...) [ilrLD3r]

Ainsi, si la suite *je elle* appartient clairement à la disfluence, ce n'est pas le cas de *elle je*, dans le contexte que nous avons décrit.

Dans le graphe de recherche de ces formes, nous n'avons donc pas tenu compte des pronoms *nous* et *vous*, et nous avons contraint la position des pronoms *elle* et *elles* au *repair* (et non au *reparandum*).

Le tableau suivant donne la répartition des 163 autocorrections que nous avons relevées dans le corpus, en fonction des contraintes que nous venons de décrire. En abscisse on a la forme de départ, et en ordonnée la forme qui achève la séquence disfluente. Les données sont présentées en fréquence relative (pourcentages) et, entre parenthèses, en fréquence absolue.

	je	j'	tu	on	il	ils	elle	elles	TOTAL
je	0,61 % (1)	46,01 % (75)	3,68 % (6)	2,45 % (4)	7,98 % (13)	1,23 % (2)	0,61 % (1)	0 % (0)	62,58 (102)
j'	3,07 % (5)	0 % (0)	0 % (0)	0 % (0)	0 % (0)	0 % (0)	0 % (0)	0 % (0)	3,07 (5)
tu	0 % (0)	0 % (0)	0 % (0)	1,23 % (2)	0,61 % (1)	0 % (0)	0 % (0)	0 % (0)	1,84 (3)
on	6,75 % (11)	1,84 % (3)	0 % (0)	0 % (0)	6,75 % (11)	0 % (0)	0 % (0)	0,61 % (1)	15,95 (26)
il	4,29 % (7)	1,84 % (3)	0 % (0)	3,07 % (5)	1,23 % (2)	0 % (0)	0 % (0)	0 % (0)	10,43 (17)
ils	0,61 % (1)	2,45 % (4)	0 % (0)	2,45 % (4)	0,61 % (1)	0 % (0)	0 % (0)	0 % (0)	6,13 (10)
TOTAL	15,33 % (25)	52,14 % (85)	3,68 % (6)	9,20 % (15)	17,18 % (28)	1,23 % (2)	0,61 % (1)	0,61 % (1)	100 (163)

Tableau 24. Autocorrections du pronom sujet

On le voit d'emblée ici, toutes les possibilités d'autocorrections du pronom sujet ne sont pas actualisées. C'est la première personne du singulier (*je* et *j'*) qui est la plus fréquemment concernée, aussi bien en ce qui concerne la forme de départ (67,49 %) que la forme d'arrivée (65,65 %). Ceci nous semble étroitement lié à notre corpus, composé majoritairement d'entretiens semi-dirigés. En effet, les locuteurs sont invités à donner leur avis, à parler d'eux-mêmes, ils répondent à des questions et s'expriment donc souvent à la première personne. Nous avons d'ailleurs vu au chapitre 2 le classement du *je*, en deuxième position parmi les formes les plus fréquentes du corpus et très bien représenté dans les répétitions. On n'aurait sans doute pas des scores comparables dans d'autres genres représentés à l'oral, comme des conversations ordinaires ou encore des débats⁵. On est sans doute ici devant l'un des effets du corpus que nous avons exclu pour le cas des déterminants définis (cf. § 2.1.1).

La correction la plus fréquente est une adaptation phonétique/morphologique. En effet, 46 % des occurrences concernent une correction du *je* en le pronom élidé *j'*. On a vu que ce type de correction avec l'élision était également très fréquent pour les déterminants définis. La

⁵ Il faut évidemment nuancer le type de débat, même à l'intérieur des débats électoraux. Par exemple, dans les débats préélectoraux des élections communales et provinciales d'octobre 2006, les candidats parlent principalement au nom de leur parti et de l'action de celui-ci. Le *nous* est nettement privilégié par rapport au *je*. Lors d'élections présidentielles, on verrait sans doute le *je* revenir en force.

correction dans le sens inverse est beaucoup plus rare, puisqu'on n'en a que 5 occurrences. Ces 5 occurrences suivent toutes le même format : entre le pronom élide et le pronom sous sa forme pleine, on a une amorce de morphème. Ici aussi, ce sont sans doute les choix de transcription faits par les transpositeurs qui orientent l'analyse que nous faisons des données. En effet, on peut supposer qu'un certain nombre de *j/je* dans le texte auraient pu être transcrits *j' je*, choisissant de faire de l'amorce un pronom élide. Comme pour les séquences *l' le*, l'analyse est clairement orientée par l'établissement des données secondaires.

Après la première personne du singulier, ce sont les pronoms de la troisième personne du singulier *on* et *il* qui entrent le plus fréquemment en jeu. En ce qui concerne la forme de départ, on a 17,18 % de *il* et 9,20 % de *on* ; pour la forme finale de l'autocorrection, on a 15,95 % de *on* et 10,43 % de *il*, soit 26,38 % des occurrences ainsi corrigées.

N'ayant pas pris en compte les formes du pluriel *nous* et *vous*, on ne peut tirer de conclusion en ce qui concerne un éventuel changement de nombre lors de l'autocorrection.

Pour ce qui est du genre, les seules possibilités de détermination concernent le passage du masculin *il/ils* vers le féminin *elle/elles* (puisque le sens inverse n'a pas été cherché pour les raisons développées ci-dessus), et le corpus n'en révèle aucune occurrence.

2.3 Les pronoms relatifs

Dans la catégorie des pronoms relatifs, nous nous sommes penchée sur les formes *qui*, *que*, *qu'* et *dont*.

Parmi les formes simples de pronoms relatifs, nous n'avons pas retenu le *où*, dont les occurrences sont peu nombreuses et qui posait un problème lors de sa cooccurrence avec le *que*, tous deux étant possiblement ambigus. Le seul bon exemple, pour notre propos est le suivant :

norkJ1 je s/ / c'était encore une famille **qui viv/ où**
tout le monde vivait dans la même pièce (...) [norkJ1r]

Nous verrons qu'il sera envisagé lors de la section sur les amorces de morphèmes.

Le *quoi* est également très ambigu, puisqu'il fonctionne fréquemment dans le français parlé comme un phatique. Nous ne l'avons donc pas pris en considération ici.

Le *dont* est concerné par une seule occurrence, où il apparaît en position du *repair*, rectifiant un *qui* préalablement énoncé :

ilePA2 je pense que c'est le hasard // me tout à fait
 bien sûr peut-être certaines personnes **qui dont** le
 métier // dont le métier est justement de parler (...)
 [ilePA2r]

On remarque dans cet énoncé ce que nous appelons un *empilement* de disfluences : un *dont* remplace un *qui*, mais prend part également à une séquence répétition, puisqu'il initialise le répétable *dont le métier*. Nous reviendrons plus loin (§3) sur le traitement de ces séquences complexes.

La seule occurrence mettant en scène un *dont* n'étonnera guère : cette forme est peu fréquente sur l'ensemble du corpus (on insiste régulièrement sur la pauvreté des relatifs autres que *que* à l'oral) mais également dans les répétitions (10 occurrences).

Par contre, le *que* et le *qui* étaient fréquemment l'objet de répétitions : 9^e forme la plus fréquente avec 309 occurrences pour le *qui*, 20^e position pour le *que* avec 157 occurrences. Dans ce dernier cas, ne sont pas séparées les occurrences selon la catégorie grammaticale, pronom ou conjonction. Dans cette section, nous faisons la séparation des occurrences entre ces deux catégories grammaticales.

Pour les formes prises ici en considération, nous avons 61 séquences. Parmi celles-ci, le *que* est la forme la plus fréquente, puisqu'elle est présente dans 88,5 % des séquences. Parmi ces occurrences de *que*, on a plus de conjonctions, puisque celles-ci concernent 32 cas, soit près de 60 % des *que*.

Ici aussi, la correction la plus fréquente, indépendamment de la catégorie grammaticale, concerne un aménagement phonétique/morphologique, avec 73,8 % de l'ensemble des 61 séquences d'autocorrections recensées. L'importance de ces aménagements est donc indéniable et éclaire le processus mis ici en jeu. De nouveau, il s'agit uniquement⁶ du sens qui va de la forme pleine vers la forme élidée, pour les raisons déjà mises en avant antérieurement dans cette section.

Les exemples que nous avons récoltés nous incitent à la prudence au moment de les intégrer dans le prétraitement. En effet, les occurrences montrent des cas où l'on n'est pas dans la disfluence. Ici, la suite *que qu'* ne manifeste pas un piétinement sur l'axe syntagmatique :

ilpCM1 (...) quand on se dit **que qu'**on le veuille ou non
 on est obligé de tenir un discours un peu langue de bois
 euh moi j'aime parler vrai euh c'est un gros défaut en
 politique [ilpCM1r]

On a notamment ce phénomène avec la locution conjonctive *parce que* :

⁶ On n'a pas ici d'exemples de la forme élidée vers la forme pleine, avec une amorce de morphème entre les deux.

ilpCM1 (...) ça peut tenir aussi parfois à l'une ou l'autre formule un peu ratée parce **que qu'on** le veuille ou non on est obligé de ramasser / de ramasser son esprit de ramasser sa synthèse en images / |- euh [ilpCM1r]⁷

Néanmoins ces cas sont minoritaires, et les rares occurrences qui posent un problème ne remettent pas en cause la rentabilité du prétraitement que nous souhaitons faire en repérant ces formes dans le texte (cf. chap. 4).

L'exemple sans doute le plus atypique, dans notre corpus, est le suivant :

iljPF1 (...) c'est le professeur Jean Bernard / euh qui qui disait **que qui** était atterré / de l'évolution / euh des facultés de médecine (...) [iljPF1r]

Nous avons postulé dans la section consacrée aux répétitions que répétables et répétés appartiennent à la même catégorie grammaticale. Ici, en fait, dans la suite *que qui* ce n'est pas le cas, le *que* étant une conjonction. C'est lors de la correction de l'étiquetage que des problèmes de ce type devront être résolus.

2.4 Les prépositions

Le problème qui se pose d'emblée ici est qu'un certain nombre de prépositions sont ambiguës (cf. chap. 5), et une recherche générale sur la catégorie entraîne un bruit considérable.

De plus, la recherche d'une suite de deux prépositions contigües fait émerger des occurrences qui ne relèvent pas de la disfluence mais appartiennent à la langue la plus standard. C'est le cas dans les énoncés suivants :

famED1 j'ai pas assez de sous [suks] / avec X |- pour <famES1> m -| le téléphone j'appellerai demain **de chez** / Maman / ah [famES1r]

ilpGV0 et selon vous qu'est-ce que bien parler on en parle **depuis à** peu près une demi-heure mais / si on devait donner une sorte de définition ou déterminer des critères [ilpMJ1r]

Dans ce dernier exemple, c'est l'adverbe composé *à peu près* qui doit être pris en compte et non la préposition *à* seule. Néanmoins, au stade de prétraitement où nous en sommes, on ne peut travailler sur les catégories grammaticales mais seulement sur les formes en tant que telles. C'est donc le

⁷ Il est amusant de voir cette construction avec l'incise *qu'on le veuille ou non* chez le même locuteur.

à seul qui est reconnu. Or, il ne se substitue en aucun cas à la préposition *depuis* dans cet énoncé, et la séquence ne relève pas d'un phénomène d'autocorrection.

L'ambiguïté de certaines prépositions, liée à la possibilité qu'une séquence soit ou non disfluente, complique à ce stade grandement le travail automatique de repérage.

Ainsi, la forme *en* qui est ambiguë entre les catégories préposition et pronom. Dans un texte standard, le *en* qui suit la préposition *à* est obligatoirement un pronom, comme dans les exemples suivants :

norHJ1 confessionnal a deux N et qu'on n'a qu'à en
mettre un [norHJ1r]

ilrLD3 mais c'est sou/ mais je sais pas bah quand je
parle euh comme ça oui j'arrive à en dire si / si je
commence à réfléchir c'est là j' arriverai pas à en
trouver (...) [ilrLD3r]

Malgré tout, dans notre corpus de français parlé, on rencontre les occurrences suivantes, où le *en* subséquent à la préposition *à* n'est pas un pronom :

famES1 c'est quand même aussi un début de ton
indépendance / |- et / <famHC1> ah ben c'est sûr ça
{ba} -| tu apprends à à vivre à en |- communauté quoi
[famES1r]

ilrLD3 (...) c'est pas possible non plus ou alors il faut
vraiment choisir ce domaine mais à / en Belgique
comme c'est pas suivi le sport (...) [ilrLD3r]

Ici, on est bien dans le cadre de la disfluence : le *en* est une autocorrection de *à*, et tous deux appartiennent à la catégorie grammaticale des prépositions.

Nous venons de le dire, le problème, à ce stade, est que le repérage des disfluences que nous faisons se fait sans analyse grammaticale, sans désambiguïsation préalable. Ainsi, dans notre recherche des prépositions contigües, nous avons recensé des exemples non pertinents pour notre propos, comme le premier présenté ci-dessus. Néanmoins, sur le même schéma formel, on trouve des énoncés qui sont bien ceux qui nous intéressent (le deuxième exemple, même si ces cas sont nettement moins nombreux que ceux du premier type). Le problème est de savoir comment gérer cette diversité d'effets, représentée formellement par le même type de séquences graphiques.

En fait, on aurait pu choisir de traiter uniquement les cas dont on est certaine qu'ils ne posent pas de problèmes, écartant les occurrences

d'autocorrections contenant une préposition que nous n'avons pas retenue pour les raisons évoquées. Nous avons pourtant choisi ne pas traiter les prépositions du tout lors de cette étape de prétraitement, préférant offrir une description unifiée de l'autocorrection (en tout cas à l'intérieur d'une même catégorie) qui ne traite pas le cas par cas.

Ces cas poseront donc inévitablement des problèmes lors de la désambiguïsation, et l'on devra être attentif lors de la correction de l'annotation.

3 EMPILEMENT DE DISFLUENCES : INTERACTION DE L'AUTOCORRECTION ET DE LA RÉPÉTITION

Nous avons choisi d'analyser l'interaction de la répétition et de l'autocorrection pour le déterminant défini, étant donné qu'il s'agit de la catégorie pour laquelle les effectifs sont les plus nombreux.

Nous avons 59 cas où autocorrection et répétition cooccurrent, soit 28 % des cas d'autocorrections recensés. On assiste donc à un piétinement sur l'axe syntagmatique, piétinement renforcé par un empilement des disfluences.

Trois grands cas de figure peuvent se présenter :

1. la répétition apparaît avant l'autocorrection :

norGA1 (...) moi je ne vois pas la la l'intérêt d'imiter la prononciation parisienne (...) [norGA1r]

2. la répétition apparaît après l'autocorrection :

ileDC1 euh parce que d'abord pour le d'abord sur le plan de la forme ils ont une langue assez claire / claire bien bien découpée / un peu chantante sans sans en avoir sans avoir la le le disons le le ton liégeois par exemple (...) [ileDC1r]

3. l'autocorrection est au milieu de la répétition :

accCT0 et n/ / est-ce que pour vous l'accent c'est une euh // c'est ça a à voir avec euh la le euh la catégorie sociale // je vous ai demandé ça mais vous n'avez pas été euh [accPH11r]

Ces cas sont ceux pour lesquels la séquence débute et finit par le même déterminant.

Dans le corpus, les deux premières structures sont le plus fréquemment actualisées, à parts égales, avec 24 occurrences chacune. On n'a que 5 séquences qui suivent le troisième format.

Outre ces trois grands cas de figure, on a une séquence où l'on a une répétition puis une autocorrection répétée (une suite de deux répétitions donc) :

norHJ1 (...) c'était donc le le best-seller comme on dit
|- n'est-ce pas <norFA0> mm -| donc c' était vraiment
la la l' / l'ob/ l'ouvrage est resté en vedette [norHJ1r]

Nous recensons aussi 4 suites plus complexes, où se mêlent plus intimement répétition et autocorrection⁸ :

ileDC1 beaucoup par clichés / la réunion de la dernière
chance euh les euh les carottes sont cuites euh les
comment euh la / le le feu couvre ou bien euh la
comment dirais-je la réunion de la dernière chance ça
(...) [ileDC1r]

Selon Candéa (2000 : 323),

(...) toutes les combinaisons de la répétition avec
l'autocorrection font partie de marques de TdF [travail
de formulation] très complexes comportant des
combinaisons de 4 ou plusieurs phénomènes ; ces
combinaisons occasionnent toujours une rupture
importante de l'énoncé.

Or, au vu de nos données, on ne peut tirer une telle conclusion. En effet, parmi ces séquences d'empilement, nous n'avons que 7 cas où la disfluente marque le lieu d'une rupture syntaxique dans la construction de l'énoncé : le discours, s'il se poursuit sur l'axe syntagmatique, laisse inachevée la place syntaxique ouverte par le déterminant initialisant le syntagme. Par comparaison, nous avons pour les autocorrections de déterminants définis également 7 autres cas de ruptures syntaxiques, dans des séquences où l'on n'a pas d'empilement avec la répétition. Ici encore, nous sommes confrontée au problème de la reconnaissance de la rupture syntaxique. Dans notre corpus, l'empilement, qui pourrait donner à penser à une plus grande complexité (et donc à un « raté » plus important en un point précis), n'est pas un indice certain de rupture : autant de cas de rupture avec l'autocorrection se font sans empilement avec la répétition qu'avec empilement ; plus de cas d'empilements ne donnent pas lieu, comme le suggère Candéa, à une rupture syntaxique. On a d'ailleurs des séquences complexes, où plusieurs marques se succèdent, pour lesquelles le déroulement syntagmatique se poursuit régulièrement passé le point de disfluente, même si celui-ci est long et que s'additionnent divers phénomènes :

⁸ Les 3 autres séquences ont la forme : *le la le le la* ; *les la la l'* ; *les le le la..*

iljNJ1 pff non // il faut dire que // moi le wallon que je connais n'est pas vraiment le wallon pur tel que le parlent le le enfin le les vrais wallingants c'est plutôt vraiment un mélange de patois de de certaines expressions qui sont un peu mélangées (...) [iljNJ1r]

norkJ1 (...) il y a évidemment le le le // euh l'aspect extérieur hein qui // qui qui qui a maquillé cette réalité-là (...) [norkJ1r]

Dans le premier énoncé, on a d'une part une séquence composée d'une répétition, à l'intérieur de laquelle on a un ponctuant (*enfin*), suivie d'une autocorrection ; la disfluence du second exemple voit s'empiler une répétition avec deux répétables, suivie d'une pause longue et d'un *euh*, avant d'arriver à l'autocorrection du déterminant. Ces séquences disfluentes, pourtant longues, n'empêchent pas le déroulement syntaxique de se poursuivre, avec l'achèvement du groupe nominal initié par le déterminant.

4 CONCLUSIONS

On a vu dans cette section que l'autocorrection, telle que nous l'avons définie, est très proche de la répétition. Nous avons mis en évidence certaines régularités :

- les formes les plus souvent concernées par l'autocorrection sont celles qui sont les plus fréquentes dans le corpus général, et aussi celles qui font le plus souvent l'objet d'une répétition ;
- répétition et autocorrection sont souvent cooccurentes dans une même séquence, et il est des cas où les phénomènes sont entremêlés, au point qu'après une autocorrection, le *reparandum* est le même que le *repair* ;
- les changements de genre et/ou de nombre vont préférentiellement du général vers le particulier, c'est-à-dire du singulier vers le pluriel et du masculin vers le féminin, mais l'on ne peut en faire une règle générale ;
- l'autocorrection, pour la catégorie des déterminants, se fait à l'intérieur d'une même sous-catégorie ;
- la fin de l'autocorrection donne lieu en général à la reprise du déroulement syntagmatique de l'énoncé, sans rupture. Lorsque celle-ci a lieu, on a bien du mal à trouver des indices formels qui indiqueraient qu'on laisse l'énoncé inachevé au lieu de l'autocorrection.

Cette section sur l'autocorrection demanderait à être complétée, notamment par un travail approfondi sur les prépositions et sur les suites

préposition + déterminant. Les résultats de l'étiquetage pourraient encore être améliorés en tenant compte de plus de cas de figure d'autocorrections.

CHAPITRE 3. SECTION 4

L'AMORCE DE MORPHÈME

ilpDJ1 oui / ne pas bien parler / c'est
c'est bien sûr commettre des fautes
[fautes] des fautes pardon mais c'est
surtout à mon avis euh euh chercher ses
mots euh a/ a/ avoir des difficultés dans
l'expression quoi [ilpDJ1r]

1 DÉFINITION

Nous appelons *amorce* le phénomène langagier qui consiste en « une interruption de morphèmes en cours d'énonciation » (Pallaud 2002b : 79).

L'exemple suivant est un cas typique d'amorce. Le morphème interrompu – symbolisé par une barre oblique collée directement à la droite de celui-ci, au lieu de l'interruption – est corrigé plus loin dans l'énoncé, où il est repris sous sa forme pleine :

ilrPC1 (...) j'aimerais bien moi ouvrir un ma/ un petit
magasin (...) [ilrPC1r]

Selon Blanche-Benveniste *et al.* (1990), l'amorce participe d'un phénomène d'anticipation, marqué par des allées et venues sur l'axe syntagmatique :

Ces anticipations montrent que le locuteur tient en réserve, lorsqu'il parle, les syntagmes qu'il vient de dire aussi bien que ceux qu'il projette de dire. (Blanche-Benveniste 1990 : 23)

L'énoncé suivant, qui contient deux amorces, illustre bien ces allées et venues sur l'axe syntagmatique, où le locuteur anticipe une première fois l'adjectif *propre* et ensuite le nom *caractère*, avant de reprendre ces deux termes dans un syntagme complet :

ilrLP1 oui pour moi tous tous les personnages qui
viennent à l'encontre de la personnalité euh pro/ enfin

des / du caractère de son **car/** de son **propre**
caractère c'est toujours beaucoup plus difficile à
 interpréter parce qu'on doit on doit / on doit sortir
 beaucoup pluS de soi je veux dire (...) [ilrLP1r]

D'emblée, on distinguera ce phénomène de l'apocope – qu'elle concerne des formes relativement standard comme *prof* ou *ciné*, ou des formes plus rares et plus individuelles comme *mayo(nnaise)* ou encore *pyj(ama)* –, où la troncation du mot est volontaire (Dister 1995)¹.

Jeanjean (1984) et Martinie (1999) désignent sous le terme d'*amorce* non seulement les amorces de morphèmes mais également les amorces de syntagmes. Ainsi, pour ces auteurs, un groupe syntaxique incomplet est qualifié d'amorce. Dans l'exemple suivant, le groupe syntaxique qui démarre avec la répétition du déterminant *un* n'est qu'amorcé puisque la place syntaxique est laissée vide (*vous connaissez* annonce une autre place syntaxique dans l'énoncé). Ici, de plus, on constate que le syntagme n'est pas complété dans la suite de l'énoncé².

norkJ1 (...) je vous signalerai un un vous connaissez
 euh // euh madame Walter là qui a travaillé avec
 Martinet [norkJ1r]

On peut donner ici un statut syntaxique au syntagme amorcé, sans avoir besoin de le reconstituer (ou de l'imaginer) sous sa forme entière. Il en est de même pour un certain nombre d'amorces de morphèmes.

Dans la littérature sur le sujet en français, on trouve également d'autres termes pour désigner l'amorce, telle que nous l'entendons ici, ou des phénomènes proches.

Pour Cappeau (1998), les amorces de syntagmes sont désignées sous le terme de *bribes*, et relèvent plutôt soit de la répétition de déterminants, soit de ce que nous appelons « autocorrection immédiate » (cf. chap. 3.3). Bres et Gardes-Madray (1991 : 96 et sv.) parlent quant à eux de « tâtonnements syntagmatiques ». Martinie (1999) utilise les termes d'« énoncés réparés » et n'isole pas les amorces d'autres disfluences dans son analyse. Dans les études de Grosjean et Deschamps (1972, 1973 et 1975), les amorces sont incluses dans les répétitions de syllabes.

¹ Réapparaît ici le concept d'intention du locuteur que nous avons tenté d'écarter dans ce travail, mais dont nous ne pouvons faire l'économie ici puisqu'il justifie des choix de transcription différents : barre oblique collée directement à la droite du morphème amorcé, aucune marque dans le cas de l'apocope.

² Notons que l'inachèvement ne fait pas nécessairement partie de la définition de l'amorce chez ces auteurs.

Comme nous l'avons déjà fait remarquer, la terminologie n'est donc pas unifiée, ce qui rend la comparaison des résultats de différentes recherches souvent impossible. Il en est de même dans la terminologie anglaise, où les termes *self-repair* et *self-correction* désignent tantôt l'amorce de morphème, tantôt un processus plus large dans lequel l'amorce, telle que nous l'entendons ici, n'est qu'une des réalisations possibles parmi d'autres.

En ce qui concerne le français, le travail le plus abouti sur l'amorce de morphème est à notre connaissance celui de Berthille Pallaud, sur lequel nous reviendrons souvent dans ce chapitre.

2 REPÉRAGE AUTOMATIQUE DES AMORCES

Nous avons vu dans le premier chapitre les problèmes posés au chercheur lors de la transcription des amorces. Néanmoins, une fois la transcription établie, l'amorce de morphème apparaît comme un phénomène non ambigu³, repérable de façon univoque non seulement par le lecteur mais également par programme, de manière totalement automatisée. L'amorce est en effet matérialisée dans le texte par la barre oblique directement accolée à la fin du morphème amorcé⁴.

3 RÉSULTATS GLOBAUX : RÉPARTITION DES AMORCES DANS LE CORPUS

3.1 Fréquence générale

Dans notre corpus, on rencontre 3612 amorces de morphème. Si les amorces sont nettement moins fréquentes que les répétitions (12 192 occurrences), le phénomène n'est pourtant pas rare. En effet, la moyenne dans notre corpus est d'une amorce tous les 123 mots.

Pour Pallaud (1999), les amorces constituent la moitié des achoppements dans les énoncés oraux. Nous sommes ici nettement en-dessous de ces estimations. La moyenne avancée par Henry *et al.* (2004 : 203) est de 4

³ Nous parlons ici du phénomène général de l'amorce et non de l'interprétation de celles-ci. (cf. chap. 1).

⁴ Notons néanmoins que ce choix n'est pas sans poser de problème. En effet, depuis que les transcriptions se font sous le logiciel Praat (Boersman et Weenink 2006), où la lisibilité du texte est moindre, il n'est pas rare de trouver des pauses brèves, symbolisées également par la barre oblique, collées directement au mot qui les précède. Il aurait sans doute mieux valu trouver un signe encore non utilisé dans les conventions VALIBEL pour symboliser l'amorce, afin d'éviter ces confusions.

amorces tous les 1000 mots, soit une amorce toutes les 75 secondes (calcul fait sur un débit moyen de 200 mots par minute).

3.2 Répartition par locuteur⁵

Comme pour les répétitions, on observe ici aussi une forte variation inter-locuteurs. C'est également ce que remarque Pallaud (2002b), dont les résultats mettent en évidence la composante individuelle du phénomène. L'auteure avance l'hypothèse que l'amorce est une « composante stable du style du locuteur » (Pallaud 2002b : 83). En d'autres termes, elle rencontrerait peu de variation intralocuteur.

Si la moyenne, tout locuteur confondu, est d'une amorce tous les 123 mots, on a évidemment des locuteurs chez qui le phénomène est plus fréquent et des locuteurs chez lesquels il l'est beaucoup moins.

Dans notre corpus, la répartition de la fréquence par locuteur⁶ va de 20,6 amorces tous les 1000 mots (ce qui équivaut à une amorce tous les 49 mots) à aucune amorce chez l'intervieweuse ilcMF0 (dans l'entrevue ilcNJ2r), locutrice qui comptabilise tout de même un peu plus de 867 mots pour cette entrevue.

Pallaud (2002b) observait un rapport de 1 à 10. En ce qui nous concerne, en enlevant les données non significatives, nous avons donc un rapport de 1 à 20.

Si le locuteur qui amorce le plus souvent est peu scolarisé (accPH1, la soixantaine, diplôme de primaires), il est suivi de très près (à 0,11 %) par un locuteur très doté culturellement, linguiste, professeur, qui peut faire figure de spécialiste de la parole (norHJ1). Ici aussi, en regardant nos données, on voit mal comment regrouper les locuteurs en fonction de leur niveau d'études ou de leur statut de professionnels de la langue ou de la parole.

De la même manière, la distinction entre entrevues semi-dirigées et conversations spontanées ne donne pas des résultats probants : on voit que les scores, classés par fréquence décroissante, font alterner les locuteurs interviewés et les locuteurs participants à des conversations amicales ou familiales.

Se dessine néanmoins une tendance : en général, les intervieweurs ont des scores plus bas que les autres locuteurs du corpus. On observait déjà cette tendance aux scores bas lors de l'analyse des répétitions : les intervieweurs obtenaient en général des taux de répétitions plus faibles que

⁵ Le tableau en annexe 6 présente le nombre d'amorces par locuteur.

⁶ Nous avons retiré les données non significatives de ces résultats (lorsque le locuteur comptabilise peu de mots).

les informateurs. Ainsi, dans les cinq entrevues dans lesquelles elle intervient, la locutrice chaGG0 a des scores suivants : 2,3 amorces tous les 1000 mots ; 1,9 amorces tous les 1000 mots (deux fois) et aucune amorce dans deux entrevues. L'intervieweuse chez laquelle on observe la plus grande variation est norFA0 : son taux d'amorce va de 13 à 6,6 amorces tous les 1000 mots, soit une fréquence la plus basse supérieure au score le plus élevé de chaGG0. Il est intéressant de noter que là où norFA0 amorce le plus fréquemment, c'est dans l'entrevue norHJ1r, entrevue pour laquelle nous avons déjà relevé chez elle un fort taux de répétitions, que l'on ne trouvait pas dans les autres entrevues de la locutrice. À cet égard, cette entrevue mériterait sans doute que l'on s'y attarde davantage, afin d'observer les raisons qui peuvent expliquer une telle variation par rapport aux autres entrevues dans laquelle la locutrice intervient.

Une donnée qui pourrait être prise en compte, afin d'expliquer ces variations que l'on rencontre entre les locuteurs et, chez les mêmes locuteurs, dans plusieurs textes, serait sans doute le débit de parole. Nous n'avons évidemment aucune donnée de ce type dans les transcriptions dont nous disposons.

3.3 Catégories d'amorces

Se basant sur des cas relativement simples d'amorces, Jeanjean (1984 : 171) affirme que :

Pour assigner une valeur syntaxique à ce genre d'amorces, on n'a pas besoin de reconstituer un quelconque lexème.

Pour des cas plus complexes, l'auteure poursuit :

La référenciation lexicale est donc seulement une question de probabilité, liée à la cohérence sémantique du texte. (Jeanjean 1984 : 172)

Or, il est des occurrences où la probabilité de rétablir la forme pleine correcte est très faible. C'est notamment le cas lorsque la forme amorcée est composée d'un seul phonème, comme on le voit à quatre reprises dans l'énoncé suivant.

norKJ1 (...) euh nous je crois que nous étions trois seulement à ne pas l'avoir // suivi // et nous étions repérés puisque / Warnant on le voyait pour la première fois en première licence // euh il n/ / à ce moment-là j'avais pas eu de cours avec lui avant donc p/ on remettait une fiche / tous ceux du premier cours avaient donc déjà remis leur fiche / au deuxième cours qui n'était pas à la première heure / on re/ remettait une

fiche / euh // euh moi il voit / provenant de Verviers il
me fait lever dites un peu un bon banc blanc / euh j'ai
dû me livrer à toutes sortes de simagrées / hum m/ oui
m/ oui ça va / ça va // mais euh il courait le bruit / que
// bon si on n'avait pas // justement un bon accent / à
l'agrégation on risquait d'avoir des ennuis (...) [norkJ1r]

Dans cet extrait, repris de l'entrevue réalisée auprès d'un professeur de linguistique à l'université, on a cinq amorces de morphèmes dont quatre pour lesquelles on ne peut rétablir ni une forme pleine ni une place syntaxique. En effet, le *n/* peut être l'amorce soit d'un verbe soit de l'adverbe de négation *ne* ; on ne voit pas bien quelle hypothèse on peut raisonnablement formuler quant à l'amorce *p/*, non seulement en termes de forme pleine, mais aussi de catégorie grammaticale ou encore de place syntaxique ; on ne peut identifier la forme entière des deux occurrences de *m/*, même si l'on peut émettre l'hypothèse de l'amorce de la conjonction *mais* que l'on rencontre plus loin dans l'énoncé.

Seule est en fait identifiable l'amorce complétée (cf. § 3.4.1) *re/*, soit un seul cas sur les cinq ici rencontrés.

Dans le cas d'amorce qui suit, le *l/* peut aussi bien être l'amorce d'un pronom clitique (avec comme place syntaxique celle d'un complément direct (*le, la, les*) ou indirect (*lui, leur*)) que d'un verbe :

chaBP1 (...) je vais vous envoyer euh je vais écrire à
l'éditeur qu'il vous envoie un truc ou n'importe quoi ou
bien vous l/ / il vous enverra le bouquin avec un bulletin
de versement vous n'aurez qu'à payer (...) [chaBP1r]

Alors que Jeanjean (1984) semble ne jamais remettre en question le fait qu'il soit possible d'attribuer un statut syntaxique à l'amorce, on voit que l'exercice est pourtant parfois périlleux.

Sur un corpus de 1000 amorces, Pallaud (2002b) n'a pu identifier la place syntaxique de 46 occurrences et la catégorie grammaticale dans 75 cas.

3.4 Typologie des amorces

Nous l'avons dit, l'étude sans doute la plus fouillée sur les amorces en français a été effectuée par Berthille Pallaud, soit sur différents types de corpus (1999), soit sur un corpus recensant 1000 amorces (2002b), soit encore d'un point de vue comparatif (avec Henry, 2003, 2004a et 2004b).

Pallaud (2002b) a établi une typologie des amorces qui les répartit en trois catégories distinctes, selon les effets produits au lieu syntaxique qu'occupe l'amorce : l'amorce complétée, l'amorce corrigée et l'amorce laissée inachevée. Nous reprenons ici cette typologie que nous discuterons ensuite dans la perspective de l'analyse automatisée.

3.4.1 Les amorces complétées

Dans l'amorce complétée, le mot amorcé se trouve complété après l'interruption, sur une même place syntaxique. Dans cette catégorie, nous pouvons distinguer trois cas différents :

3.4.1.1 *Le mot amorcé est complété sans reprise de la partie amorcée*

iljDV1 je blague / enfin // je / à Gand parce que parce qu/ j'ai j'ai env/ j'ai envie de dire Gand parce que les les Flamands qui ont appris le français l'ont l'ont appris euh dans des milieux sociaux très très privilégiés probablement dans des milieux dans lesquels on on a/-ccordait énormément d'importance au au au au français (silence) (...) [iljDV1r]

Ces cas sont très rares. On n'en rencontre que 14 occurrences sur les 3612 amorces que compte notre corpus. Pallaud (2002b), quant elle, obtient des résultats comparables aux nôtres puisqu'elle en recense 3 sur 1000 occurrences d'amorces.

Dans ce type d'amorce, aucun des deux segments entre des séparateurs – l'amorce et le *repair* – ne constitue un mot complet, analysable en tant que tel.

3.4.1.2 *Le mot amorcé est complété avec reprise de la partie déjà énoncée*

iljDV1 apprendre ça c'est transm/ transmettre un savoir donc ça c'est apprendre communiquer euh euh (...) [iljDV1r]

Ici, à la différence du cas précédent, le *repair* constitue un mot complet, répertorié comme tel dans le dictionnaire. La correction suit immédiatement l'amorce.

En formulant des hypothèses quant à la cohérence et la congruence des propos du locuteur, nous considérons qu'un *repair* qui reprend à l'initiale la même forme que l'amorce est bien la forme complète de celle-ci. C'est d'ailleurs en se basant sur une telle hypothèse que le transcripteur a posé des choix de transcription pour certains morphèmes, principalement quand ils ne sont que des phonèmes (cf. chap. 1). Dans cette hypothèse de cohérence, nous basant sur une identité formelle et une stricte contigüité, on recense dans notre corpus 750 amorces qui sont immédiatement complétées par leur forme pleine, soit un peu plus d'un cas d'amorce sur cinq.

Parmi celles-ci, nous avons certains cas particuliers d'amorces dont nous avons déjà parlé dans la section consacrée aux autocorrections immédiates : il s'agit des amorces complétées qui se présentent sous la forme suivante : *j/* ou *l/*. Nous recensons dans le corpus 56 cas d'amorces *j/* complétées sous la forme du pronom personnel sujet *j'* ou *je* ; pour ce qui est de l'amorce notée *l/*, celle-ci est complétée en *l'* (13 occurrences), *les* (9), *la* (15), *le* (17) ou suivie d'une autre amorce *l/* (7). Nous renvoyons à l'hypothèse que nous formulons pour ces occurrences.

Au final, nous verrons que les choix de transcription n'influent pas sur le prétraitement choisi, un codage unique étant utilisé dans les deux cas. (cf. chap. 4, § 5.7.3)

3.4.1.3 *Le mot amorcé est complété avec reprise de la partie amorcée mais aussi d'élément(s) antéposé(s) à l'amorce*

Dans le cas précédent d'amorce complétée, le *repair* suit immédiatement l'amorce. Ce n'est pas le cas dans l'exemple suivant :

ilrMS1 si ça me plaît bien mais // faut dire que / bon
euh avec la boulangerie euh je fai/ je faisais des
pralines chez moi (...) [ilrMS1r]

Ici, la reprise concerne non seulement le verbe amorcé mais également le pronom sujet qui le précède. C'est tout le groupe syntaxique pronom sujet + verbe qui est repris.

Lorsque le mot amorcé est un substantif, le déterminant qui le précède est fréquemment repris dans le *repair* :

iljPF1 oui leur for/ leur formation -| est ils n'y sont
pour rien leur |- formation <iljGF0> oui -| est beaucoup
trop limitée [iljPF1r]

Lorsque le syntagme auquel appartient le lexème amorcé commence par une préposition, celle-ci est souvent reprise également :

norKJ1 (...) j'ai pu observer ce phénomène // d'un d'un
plus grand purisme peut-être sur le plan de la
prononciation // euh chez des Italiens de la d/ de la
deuxième génération [norKJ1r]

On voit déjà les similitudes avec le phénomène de la répétition et de l'autocorrection : outre une immobilisation sur la même place syntaxique, qui suspend l'énoncé, la reprise se fait sur un mode comparable.

3.4.2 Les amorces corrigées.

Dans le cas des amorces corrigées, la correction de l'amorce se fait également sur la même place syntaxique. Comme pour le phénomène de la répétition et de l'amorce complétée, on assiste donc à un piétinement sur la même place syntaxique et non à un déroulement syntagmatique de la suite de l'énoncé. Néanmoins, à la différence de l'amorce complétée, le lexème utilisé dans la correction diffère de celui qui a été amorcé par le locuteur. Comme pour la catégorie précédente, la correction peut soit se faire immédiatement, soit avec reprise d'éléments présents dans le contexte gauche de l'amorce, comme l'illustrent respectivement les deux exemples suivants :

chaBP1 ah oui oui -| ça me surprend oui / c'était dans le Nord de la France hein dans le Nord de la France ils ont dit nous avons un patois / sensiblement sem/pareil n'est-ce pas et il y a des expressions c'est les mêmes ce sont les mêmes (...) [chaBP1r]

accFJ1 (...) j'ai été à plusieurs reprises avec mes parents en Auvergne je trouvais aussi qu'ils avaient aussi un accent qui était pas mal euh // typique / par contre les J/ les Bretons j'ai jamais su / jamais vu qu'ils avaient d'accent moi (...) [accFJ1r]

Pallaud (2002b) parle dans ce type d'amorces d'erreur de sélection lexicale. Ce phénomène est à interpréter pour elle comme un lapsus interrompu et corrigé (2002b : 87). Les lapsus sont évidents dans les énoncés suivants :

norRM1 (...) il y a à peu près une dizaine d'années // il y avait très peu d'adultes // je crois que j'étais la seule // et il se trouve que maintenant qu' on doit déblou/dédoubler des classes d'adultes // ils sont à peu près cent-vingt (...) [norRM1r]

ileGF0 euh on accorde au langage journalistique une fonction de communication

ilePA2 oui

ileGF0 est-ce que vous songez à d' autres communicat/ euh à d'autres fonctions (rire) [ilePA2r]

Dans les amorces corrigées, la catégorie grammaticale de la correction est en général la même que celle de l'amorce. Mais il peut arriver qu'elle soit différente. Néanmoins, ces cas sont les plus rares.

3.4.3 Les amorces inachevées

À la différence des deux types d'amorces que nous venons de voir, la correction de l'amorce inachevée se fait sur une autre place syntaxique. En fait, il serait plus correct de dire qu'il n'y a pas véritablement de correction : le déroulement syntagmatique de l'énoncé est interrompu et la place amorcée est laissée vide. L'énoncé se poursuit en recommençant sur une autre place syntaxique, comme l'illustre l'exemple suivant :

ilrVI2 m quand un néerlandophone parle français / euh
je trouve que ça ne fait pas bien du tout / par rapport à
quelqu'un qui parle bien **fran/** comme un Bruxellois par
exemple (...) [ilrVI2r]

On le voit, le déroulement syntagmatique se poursuit après une rupture syntaxique. Celle-ci peut produire l'effet de liste « coordonnée » (Blanche-Benvensite 1987), comme dans l'exemple suivant :

blaJV1 non -| je me suis dit / va-t'en quoi non tout le
monde {nous , me} dit bè / c' est à lui de partir c'est pas
à toi de partir et tout ça et / rien à **f/** je voulais plus rien
savoi/ et tout le **mon/** on n'a pas arrêté de me sonner
tout le trajet et euh pf // j'ai pas décroché quoi j' ai pff
[blaJV1]

Parmi les amorces inachevées, on a une catégorie un peu particulière où l'amorce est due à la situation de parole. En effet, dans une minorité de cas (92 occurrences, soit 2,55 % des cas d'amorces), l'amorce est provoquée par l'interruption du locuteur par un autre locuteur qui prend la parole et la conserve, comme l'illustre l'exemple suivant :

blaNB1 ben oui / tout à fait mais c'est quoi le sujet de
sa thèse
blaAD1 ah je sais pas |- c'est **p/**
blaNB1 (xxx) -| ah tu ne sais même |- pas ? [blaNB1r]

Le phénomène, s'il apparaît semblable en surface, est évidemment à distinguer d'un point de vue discursif. Là où l'amorce qui participe de la disfluence est la trace, chez un locuteur, de l'énoncé en cours d'élaboration, l'amorce en fin de tour de parole est une forme de coupure de la part de l'interlocuteur ; son « fonctionnement » relève plutôt alors de la gestion des tours de parole. En fait, dans les résultats donnés ici, ces deux cas seront confondus. Cela dit, le second pose moins de problèmes à l'analyse automatisée dans la mesure où la position fin de tour de parole est particulière et où le traitement s'arrête dans cette configuration (cf. chap. 4).

3.5 Répartition des amorces en corpus

Pallaud (2002b : 89), dans une étude sur 1000 amorces issues de 60 locuteurs différents, obtient la répartition suivante, selon les trois catégories d'amorces qu'elle distingue :

type d'amorce	pourcentage
amorces complétées	59,9 %
amorces inachevées	28,8 %
amorces corrigées	11,3 %
total	100 %

Tableau 25. Répartition des types d'amorces chez Pallaud (2002b)

Les amorces les plus nombreuses sont les amorces complétées, avec deux tiers des effectifs. Viennent ensuite les amorces inachevées (28,8 %) et enfin les amorces corrigées avec 11,3 %.

Dans une étude précédente (1999) sur un corpus de 120 amorces, l'auteure constatait que les amorces corrigées étaient les plus nombreuses avec 52,5 % des occurrences, alors que dans une autre étude sur 150 amorces (2004a), la répartition est également en faveur des amorces complétées (59,6 %), avec une ventilation plus égale dans les deux autres catégories (18,5 % pour les amorces modifiées et 21,9 % pour les amorces achevées).

Selon Pallaud, ces différences de résultats tendent à montrer une grande variation inter-locuteur dans ce phénomène langagier.

Dans une étude effectuée sur le *Corpus Rimouski* (cf. chap. 3.2, §5.4), nous obtenons une répartition encore sensiblement différente de celles obtenues dans ces trois études.

En effet, sur 348 amorces que nous avons relevées, le phénomène le plus remarquable est sans doute le nombre d'amorces que nous n'avons pu classer dans aucune des catégories, l'identification non seulement du lexème mais également de sa place syntaxique étant totalement impossible. Ces cas s'élèvent à près de 25 % des occurrences. Pour elles, l'exercice qui consiste à classer l'amorce dans la catégorie des amorces corrigées ou inachevées est impossible.

La catégorie la mieux représentée dans le *Corpus Rimouski* est composée des amorces complétées, avec 112 occurrences, soit près d'un tiers des cas.

En ce qui concerne le schéma du *repair*, nous avons distingué pour ces amorces complétées celles qui sont complétées sans reprise et celles complétées avec reprise d'éléments dans le contexte gauche. Sur les 112

amorces complétées, la répartition est la suivante : elle montre une nette préférence pour l'amorce complétée sans reprise puisque cette configuration représente 62,5 % des occurrences, contre 37,5 % de cas où l'amorce reprend des éléments du contexte gauche. Nous reviendrons ci-dessous sur ces cas d'amorces, qui participent du phénomène de la répétition.

3.6 Les insertions entre *reparandum* et *repair*

Dans les amorces complétées et corrigées, on peut aussi rencontrer l'insertion d'éléments nouveaux entre le *reparandum* et le *repair*. Ceux-ci ne se trouvaient pas présents dans le contexte gauche de l'amorce, à la différence des reprises que nous avons vues

accGP1 non parce que je ne prends pas fort attention à tout ça (en riant) // mais in/ / disons inconsciemment on s'en |- rend [accGP1r]

Ces éléments nouveaux sont parfois des incises énonciatives ou constituent des parenthèses dans l'énoncé. En général, il s'agit plutôt pour le locuteur d'ajuster son propos. Les éléments insérés viennent modifier l'énoncé, offrant la plupart du temps une plus grande précision dans les propos du locuteur. Il peut s'agir d'un adverbe, comme dans l'exemple ci-dessous :

accCP1 l'accent borain est aussi cr/ très critiqué mais on pourrait aussi critiquer l'accent liégeois (...) [accP1r]

Mais toutes les catégories grammaticales sont possibles dans cette fonction d'ajouts modifieurs, puisqu'on rencontre également des noms, des adjectifs, des verbes, des pronoms ou encore des déterminants (Pallaud 2002b : 97)⁷ :

norGA1 seulement il y a c/ c/ il y a un certain nombre de choses qui passent au compte des profits et profits (...) [norGA1r]

blaNB1 dans le hach/ ou/ ou/ oui mais tu l'achètes pré/ be/ tu tu achètes son hachis préparé ? [blaNB1r]

Selon Pallaud, cette sous-catégorie d'amorces, dans laquelle un élément modifieur est inséré,

est particulièrement la marque d'une activité autonymique puisque les amorces concernées ne sont pas le lieu d'une modification mais annoncent une ré-

⁷ Si ces amorces interrompues sont classées par Pallaud dans les amorces complétées, du point de vue de la disfluente, on est bien dans un *reparandum*.

élaboration sur ce qui a été énoncé antérieurement.
(Pallaud 2002b : 95)

On a évidemment dans ces insertions les éléments que nous avons repérés également à l'intérieur des séquences répétition.

Dans le cadre de ce travail, nous sommes limitée par la possibilité d'automatisation de la tâche. Ainsi, pour ce qui est des amorces citées dans les deux énoncés ci-dessus, il est très difficile de déceler de manière automatique si elles sont inachevées, corrigées ou complétées. Ce n'est pas le cas pour les amorces complétées immédiatement, où l'on a une identité formelle entre le *reparandum* et le début du *repair*. Nous écrivons complétées *immédiatement* mais dans ces cas, nous avons aussi permis les insertions que nous avons autorisées pour les répétitions (cf. chap. 3.2, § 5.1.1).

Sur les 750 occurrences d'amorces complétées « immédiatement », on a les 63 éléments suivants insérés⁸ : pause brève (29), *oui* (7), pause longue (5), *enfin* (5), *euh* (12), pause brève + *disons* (1), *disons euh* (1), *non* (1), *hein* (1), *silence* (1).

Comme pour les répétitions, c'est la pause brève qui intervient le plus fréquemment dans ce contexte. Néanmoins, à la différence de ce que l'on observait pour les répétitions, les insertions sont beaucoup moins nombreuses entre *repair* et *reparandum*, puisque dans le cas de l'amorce, seules 8,4 % des occurrences sont concernées, alors que ce chiffre atteignait près de 37 % dans les répétitions.

En observant les occurrences concernées, on ne peut mettre en évidence un quelconque fonctionnement qui serait différent des occurrences sans insertion.

3.7 La catégorie grammaticale des amorces

Même si l'on ne peut identifier la forme pleine du lexème amorcé, le contexte permet, dans certains cas, d'en établir la catégorie grammaticale. Néanmoins, nous l'avons vu ci-dessus, c'est loin d'être faisable dans tous le cas.

Dans son corpus, Pallaud (2002b) a pu identifier la catégorie grammaticale de 925 amorces (sur 1000). Il s'agit en majorité de mots lexicaux (c'est-à-dire chez elle de verbes ou de noms), avec près de 4 amorces sur 5 (82 % des occurrences).

⁸ La recherche porte ici sur les éléments que nous avons sélectionnés. Il est bien évident qu'il se peut que d'autres éléments puissent être insérés et augmenter ainsi le nombre de répétitions complétées de notre corpus.

Nous avons analysé la catégorie grammaticale des 112 amorces complétées du *Corpus Rimouski*. La répartition est la suivante :

catégorie grammaticale	n=	pourcentage
préposition	7	6,2 %
conjonction de coordination	2	1,8 %
conjonction de subordination	3	2,6 %
déterminant	21	18,8 %
nom	19	17 %
pronom	17	15,2 %
adjectif	4	3,6 %
adverbe	14	12,5 %
verbe à l'infinitif	7	6,2 %
verbe conjugué	13	11,6 %
participe passé	5	4,5 %

Tableau 26. Catégorie grammaticale des amorces complétées

Comme dans le corpus de Pallaud, l'avantage est ici nettement en faveur des mots lexicaux (noms, adjectifs, verbes et adverbes) avec 70,6 % des occurrences.

En ce qui concerne cette minorité de mots grammaticaux amorcés, Pallaud (2002b) émet cette hypothèse :

Que cela soit dû à un lien entre la taille du mot et le phénomène d'amorce reste à vérifier. Cela suggère en effet que les mots polysyllabiques puissent être plus fréquemment l'objet d'achoppements que les mots monosyllabiques. (Pallaud 2002b : 93)

On assiste donc pour les amorces à l'inverse du phénomène de la répétition, où ce sont justement les mots grammaticaux, principalement monosyllabiques, qui font majoritairement l'objet de répétitions.

3.8 Amorces et répétitions

Amorces de mots et répétitions sont des phénomènes proches. En effet, s'ils se distinguent toutefois par la catégorie grammaticale des formes concernées, le mécanisme général semble très similaire, du moins en ce qui concerne les amorces complétées et les amorces corrigées. Nous avons vu notamment que les reprises que l'on constate à l'initiale du syntagme sont semblables à celles rencontrées dans les séquences répétitions.

Bien que présentant des fréquences différentes en corpus (les répétitions apparaissent *grosso modo* quatre fois plus souvent), nous avons voulu voir si, chez un locuteur, les deux phénomènes pouvaient être mis en relation. Nous avons déjà fait des remarques en ce sens dans la section concernant l'analyse des amorces par locuteur.

Nous avons voulu aller plus loin dans ce sens et avons appliqué le test de Spearman sur les amorces et les répétitions. L'idée est donc de vérifier s'il y a corrélation, pour un même locuteur, entre le fait de faire des amorces de morphèmes et celui de répéter. Le graphique de dispersion des données est le suivant :

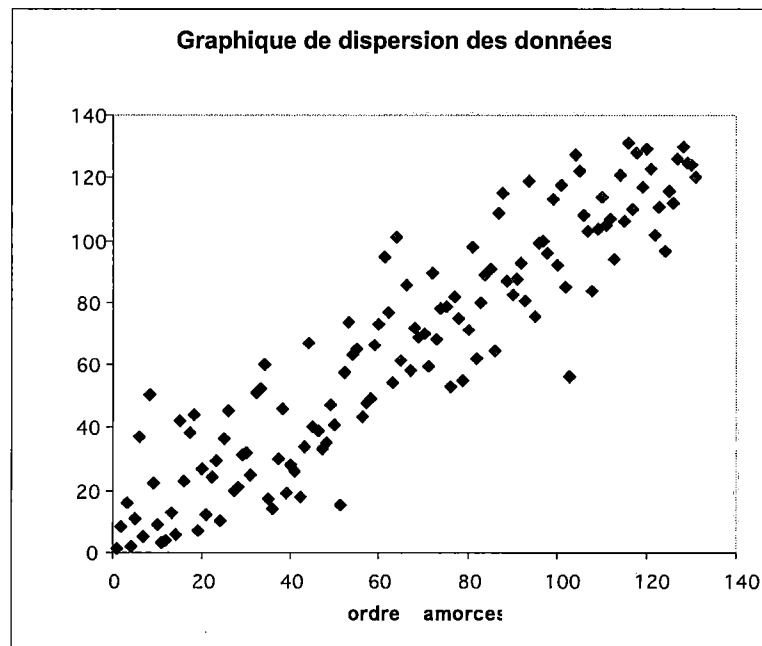


Figure 11. Graphique de dispersion des amorces et des répétitions

La valeur observée est de 0,921, soit une corrélation significative entre les deux paramètres testés. En d'autres termes, en règle générale, plus les répétitions sont fréquentes chez un locuteur, plus les amorces le seront.

également, et inversement. Certains locuteurs échappent à cette tendance. Ils sont représentés par les points éloignés de la courbe centrale sur le graphique, en dessous de celle-ci s'ils amorcent proportionnellement beaucoup plus souvent qu'ils ne répètent ou au-dessus de la courbe dans le cas contraire.

Parmi les « exceptions », le cas d'une locutrice nous semble intéressant. Il s'agit de blaND1, que nous avons déjà mise en évidence pour la longueur des tours de parole (cf. chap. 2). Ici, on voit une similitude très grande pour les deux conversations dans lesquelles elle intervient (blaAT11 et blaJV11). Classée en milieu des effectifs pour ce qui est des amorces, elle répète très peu. Cette manière (ce style ?) se confirme dans les deux textes ici concernés, et on voit que la locutrice est représentée dans le tableau pour ces deux conversations dans des points très proches, qui s'éloignent de la courbe centrale (en haut, à gauche).

En ce qui concerne accPH1, dont nous avons dit qu'il pratiquait proportionnellement le plus d'amorces, on constate qu'il est également dans la partie supérieure de la courbe, ce qui n'étonnera guère. Notons qu'il avait néanmoins tendance à faire beaucoup de répétitions, ce qui, de ce fait, ne l'éloigne pas considérablement de la courbe.

3.8.1 Les séquences d'amorces

Nous appelons séquence d'amorces une suite d'au moins deux amorces. Dans notre corpus, on recense 120 séquences, où les amorces se succèdent de manière contigüe. Parmi celles-ci, on a 13 séquences de trois amorces et une de quatre amorces :

blaJD1 et chose An/ An/ A/ An/ Madame Stassen
ça ? [blaJD1r]

Parmi ces séquences d'amorces, 70 occurrences constituent des répétitions d'amorces, selon la définition que nous avons donnée précédemment de la répétition (identité formelle stricte). Nous n'avons pas repris parmi ces cas des exemples pourtant très proches, dont le fonctionnement est à l'évidence apparenté, mais pour lesquels l'identité formelle n'est pas stricte, comme dans l'énoncé suivant :

magMM1 bè ben maintenant est-ce qu'ils ne font pas
du fas/ f/ fast-food |- comment ce que c'est
[magMM1r]

Les amorces répétées représentent 58 % des séquences d'amorces, mais seulement 1,94 % de l'ensemble des amorces du corpus.

Henry et Pallaud (2004) obtiennent des résultats très proches des nôtres puisqu'elles recensent quant à elles 2,4 % d'amorces répétées parmi les amorces.

Sur l'ensemble des séquences répétées, les répétitions d'amorces représentent 0,98 % de l'ensemble des répétitions de notre corpus, contre 0,93 % chez Henry et Pallaud (2004).

Quand on observe plus attentivement ces amorces répétées, on constate que 38 d'entre elles (soit 54,3 %) sont complétées immédiatement, comme dans l'exemple suivant :

norKJ1 (...) mais enfin moi d' où je suis je le vois / et et
ça peut **t/ t/ te** créer des ennuis / alors euh moi je l'
aime bien il est sympathique cet accent (...) [norKJ1r]

Si l'on s'intéresse aux séquences d'amorces sans identité formelle (c'est-à-dire n'étant pas intégrées dans les répétitions donc), on compte 19 amorces complétées. On a donc proportionnellement plus d'amorces répétées qui sont complétées que de séquences d'amorces non répétées complétées. Henry et Pallaud (2004) aboutissent à des conclusions comparables, en constatant que

c'est dans la catégorie des amorces complétées qu'a lieu, le plus souvent, la répétition. (...) Cela suggère une fonction de la répétition dans cet achoppement que constitue le phénomène d'amorce : elle soutient la recherche lexicale manifestée par l'amorce et contribue à l'aboutissement de cette recherche puisque, finalement, le mot interrompu est plus souvent complété. (Henry et Pallaud 2004 : 208)

Les auteures obtiennent néanmoins des scores beaucoup plus élevés que nous (82 % des amorces répétées sont complétées), ne s'étant pas limitées à une complétion directe de l'amorce, ce que nous avons fait ici.

Les répétitions d'amorces se distinguent également par la catégorie grammaticale des formes concernées. En effet, nous avons vu ci-dessus que l'amorce de morphèmes portait principalement sur les mots lexicaux. Or, dans le cas de répétitions d'amorces, la tendance s'inverse fortement puisque, sur les occurrences dont nous avons pu identifier la catégorie grammaticale avec certitude, nous n'avons pratiquement que des mots grammaticaux. Sept exceptions à cette majorité absolue : un verbe à l'infinitif, un nom propre, quatre adverbes et un participe passé.

On a donc ici un comportement, en termes de catégorie grammaticale, qui rapproche plus les répétitions d'amorces du phénomène de la répétition que de celui de l'amorce en général. La tendance que l'on observe dans notre corpus est beaucoup plus forte que celle de Henry *et al.* (2004), qui obtiennent « seulement » 66,7 % de « mots-outils » dans les répétitions

d'amorces. On ne peut expliquer à quoi cette différence est due, si ce n'est peut-être au fait d'une certaine prudence de notre part. Encore une fois, nous n'avons pas forcé l'interprétation de l'amorce et n'avons pas tranché dans les cas qui nous semblaient indécidables (principalement entre une amorce corrigée et une amorce inachevée). Peut-être ces cas auraient-ils plaidé en faveur d'une interprétation en mots lexicaux. Cela dit, ces cas indécidables n'auraient pas inversé la forte tendance observée ici.

3.8.2 L'amorce à l'intérieur d'une séquence répétition

Nous avons vu dans le chapitre consacré aux répétitions que, après la pause brève, l'amorce était la marque la plus fréquemment insérée entre les termes d'une répétition. Cette configuration représentait 20,29 % des répétitions du corpus.

Sur les occurrences de notre corpus, 1094 cas d'amorces entrent dans le contexte d'une répétition, soit entre le répétable et le répété, soit entre deux répétés. Ces occurrences représentent donc un peu plus de 30 % des amorces recensées. La structure la plus fréquemment représentée est la suivante :

répétable + amorce + répété

L'hypothèse que nous formulons pour ces cas est que l'amorce adopte ici un fonctionnement comparable à celui de la répétition en général. La répétition est le lieu d'un empilement sur la même place syntaxique, mais la réparation permet au déroulement linéaire de l'énoncé de reprendre, sans constituer une rupture syntaxique. Dans cette optique, pour les amorces qui entrent dans cette configuration, on aurait alors affaire soit à des cas d'amorces complétées, soit à des cas d'amorces modifiées. Le schéma de la répétition deviendrait alors le suivant :

Répétable amorcé + répété corrigé (soit complété, soit modifié)

Pour vérifier cette hypothèse, nous avons extrait au hasard de notre un corpus 100 amorces qui entrent dans cette configuration d'une répétition. Parmi celles-ci, 10 amorces étaient suivies d'une autre amorce dans la séquence. Nous les analyserons séparément ci-dessous car leur comportement est plus complexe. Nous avons classé les 90 amorces restantes selon la typologie de Pallaud⁹. La répartition est la suivante :

⁹ Alors que ce classement nous a posé des problèmes à bien des reprises dans le corpus Rimouki, dans cette configuration d'une répétition, il était relativement aisé de faire le départ entre les 3 types d'amorces.

type d'amorces	n=	pourcentage
amorces complétées	58	64,5 %
amorces corrigées	29	32,2 %
amorces inachevées	3	3,3 %
total	90	100 %

Tableau 27. Types d'amorces à l'intérieur d'une séquence répétition

On le voit, notre hypothèse se vérifie effectivement : il est très rare qu'une amorce dans le contexte d'une répétition soit laissée inachevée, et entraîne donc une rupture de la construction de l'énoncé. Dans 96,7 % des cas, les amorces à l'intérieur d'une séquence répétition sont donc soit corrigées, soit achevées, ces dernières étant deux fois plus nombreuses.

Quand on observe les séquences dans lesquelles l'amorce est inachevée, on constate qu'elles comportent toutes les trois une pause longue à l'intérieur de la séquence. Cela dit, cette pause ne peut servir d'indice totalement fiable pour repérer de manière automatique un inachèvement dans la mesure où celle-ci apparaît également dans des cas d'amorces complétées :

accPH1 mais **je ne sau/ -l // je ne saurais** pas te le
dire hein [accPH1r]

Parmi les 100 amorces que nous avons extraites automatiquement, nous venons de dire que 10 d'entre elles constituaient en fait une suite d'amorces. Ces cas sont donc plus complexes à traiter, et nous verrons dans la suite de ce travail comment gérer ces « empilements », ces suites de disfluences dont nous avons déjà dit précédemment qu'elles avaient tendance à se regrouper dans un même point de l'énoncé.

Si les 10 cas que nous recensons ici présentent des structures relativement différentes (en termes de nombre de répétés, de place occupée respectivement par les deux amorces, etc.), on peut néanmoins observer une constante : à part dans un seul cas, le *repair* de ces séquences complexes n'entraîne pas de rupture syntaxique. Les amorces de la séquence sont soit complétées, soit modifiées, soit modifiées et complétées comme l'illustre l'exemple suivant¹⁰ :

¹⁰ Cette séquence a été repérée automatiquement (cf. chap. 3.2, § 5.1.1). Ici, le répétable est le premier *les* en gras dans l'exemple. Le programme isole la séquence jusqu'au dernier *les* (le 3^e), les amorces étant considérées comme des insertions possibles dans une séquence répétition, au même titre que les autres formes que nous avons répertoriées.

accGP1 moi je sais bien quand // si je dois parler anglais
ou enfin // j'ai déjà beaucoup oublié mais // il y a des
choses que // qu' on ne sait pas // faire les accents ne
sont pas mis au au même endroit // et les / donc **les a/**
les into/ les intonations indiquent [accGP1r]

On a dans cet énoncé une première amorce *a/*, qui est ensuite modifiée au profit d'une deuxième amorce *into/* (avec reprise du déterminant), elle-même complétée par la forme pleine *intonation* (une nouvelle fois avec reprise du déterminant).

Nous verrons au chapitre 4 comment nous traitons les amorces lors de l'étiquetage.

4 CONCLUSIONS

Nous avons vu dans cette section un phénomène disfluent relativement fréquent en corpus, puisqu'il apparaît en moyenne une fois tous les 123 mots. Comme les répétitions, l'amorce de morphème est soumise à de fortes variations inter et intralocuteurs, dont on a, au vu de nos données, du mal à déterminer la cause.

Les amorces peuvent être classées en trois grands groupes, selon les effets produits au lieu syntaxique de l'amorce : amorces achevées, complétées ou modifiées. Les amorces se répartissent dans ces trois sous-catégories selon des proportions variables, qui ne changent pas fondamentalement d'un corpus à l'autre. Si ces trois catégories sont particulièrement pertinentes en termes d'analyse linguistique, elles sont peu opérationnelles pour le traitement automatique : il est difficile de distinguer automatiquement les amorces complétées et les amorces inachevées, sauf si les premières sont complétées directement avec identité formelle du *reparandum* et du *repair*.

Si, comme les répétitions, l'amorce de morphème participe des disfluences, nous avons mis en évidence une différence fondamentale quant à la catégorie grammaticale des formes concernées : alors que les répétables appartiennent majoritairement à la classe des mots grammaticaux, les formes faisant l'objet d'amorces sont quant à elles le plus fréquemment des mots lexicaux. C'est sans doute de ce point de vue que les deux phénomènes sont les plus divergents.

Différents du point de vue de la classe grammaticale, ces phénomènes sont souvent étroitement liés puisque 30,29 % des amorces entrent dans une séquence répétition. Nous avons vu que cette configuration permettait de faire l'hypothèse que dans ces cas, formellement identifiables, aucune rupture syntaxique ne se faisait au lieu de l'amorce.

CHAPITRE 3. SECTION 5

LE *EUH* DIT D'HÉSITATION

Trucs et astuces de France Telecom :
quelques conseils

**Le service vocal ne comprend pas le
nom de la ville.**

« *Euh... alors... je crois que c'est vers
Lyon* »

Conseils :

Ne pas décomposer la prononciation des
villes à nom composé : « Saint.../
Etienne.../ les.../ Orgues ».

Il faut prononcer : « Saint Etienne les
Orgues » sans introduire de coupures
entre les mots.

Ne pas répondre : « Euh... à Bordeaux »
mais : « Bordeaux »¹.

1 DÉFINITION

En français, ce que la littérature nomme classiquement *pause remplie*, ou encore *pause sonore* ou *pause pleine*, est matérialisée par la voyelle autonome prononcée [2] et orthographiée traditionnellement *euh* ou, plus rarement, *heu*.

La pause remplie est actualisée différemment selon les langues. En anglais, on a *ah*, *eh*, *er*, *uh* ou encore *um* (*Simple Metadata Annotation Specification* 2004, Schegloff 1981). En général, on constate que la pause remplie se matérialise par une voyelle centrale fortement allongée, dont la durée a une moyenne supérieure à 400 ms (Clerc-Renaud *et al.* 2004).

¹ <http://www.francetelecom.com/fr/groupe/rd/une/118710/trucs-et-astuces.html> (avril 2006).

Selon les chercheurs, le terme de *pause sonore* recouvre également d'autres réalisations dans le parlé. Dans les études de Grosjean et Deschamps (1972, 1973 et 1975), la pause sonore comprend non seulement le *euh* (qu'ils appellent *pause remplie*) mais également les allongements et deux autres marques qu'ils qualifient de marques d'hésitation : les répétitions et les faux départs. Ce classement correspond à celui établi en 1959 pour l'anglais par Maclay et Osgood. C'est également la terminologie adoptée par Duez (1991 : 59) qui classe sous la même étiquette de *pause non silencieuse* ces quatre phénomènes. Campione et Véronis (2004) reprennent quant à eux les allongements et le *euh* sous l'appellation de *pause remplie* et précisent pour les allongements *allongements d'hésitation*.

La terminologie est donc variée, pour un même phénomène classé souvent a priori comme une *pause d'hésitation*.

1.1 Le *euh* et l'allongement

Ce que l'on entend ici par *euh*, et qui est transcrit de la sorte dans les textes de VALIBEL, est à distinguer de l'allongement vocalique. En effet, celui-ci s'attache à un mot du lexique alors que le *euh* constitue une voyelle autonome. Campione et Véronis (2004) parlent d'ailleurs « d'un item quasi-lexical » pour le *euh*.

L'allongement se caractérise par

« une voyelle continue de durée très supérieure à la normale, de qualité vocalique constante et [est] associée à une fréquence fondamentale constante (F0) plate ou très légèrement descendante ». (Campione et Véronis 2004, qui résument Guaïtella 1991)

Phénomène phonétique, qui ne se traduit pas par un morphème dans le flux de la parole, l'allongement ne fait pas l'objet d'une transcription dans les conventions VALIBEL. Les textes sur lesquels nous travaillons ne mentionnent pas cette information, qui relève de la prosodie, que nous ne pourrions donc pas prendre en compte dans notre travail.

La plupart des études associent le *euh* et l'allongement vocalique, regroupés sous l'appellation de pauses pleines qui marquent l'hésitation. Mais, comme le fait remarquer Candéa (2000a : 88), cette association se fait souvent sans justification théorique.

Ainsi, pour Grosjean et Deschamps, qui comparent l'anglais et le français, ces deux marques ont le même rôle (tout comme la répétition et le faux départ). De même, Campione et Véronis (2004) remarquent leur « similitude de propriétés et de fonctionnement » et les englobent sous l'appellation de *pauses remplies*.

Morel et Danon-Boileau (1998), par contre, font l'hypothèse d'une différence de fonctions énonciatives entre l'allongement vocalique et le *euh*. C'est cette hypothèse que reprend Candéa dans sa thèse (2000a), où elle montre que, malgré de nombreux points communs et des propriétés très proches, ces deux marques n'ont pas la même distribution (c'est-à-dire n'apparaissent pas dans les mêmes contextes).

Quoi qu'il en soit, la marque de l'allongement est absente de nos transcriptions, et la seule pause remplie que nous prendrons en considération est donc le *euh*. Si l'on extrapole les résultats des études antérieures sur notre propre corpus, cette absence de l'allongement réduit considérablement le nombre de pauses sonores dans les données sur lesquelles nous travaillons. En effet, Grosjean et Deschamps (1972) relèvent 43 % d'allongements parmi les pauses sonores (donc 57 % de *euh*) ; Candéa, quant à elle, obtient un score plus faible mais néanmoins important : l'allongement représente 34,30 % des pauses remplies de son corpus.

1.2 Le *euh* et le [ə] d'appui

Dans le travail de transcription, il est parfois ardu de faire le départ non seulement entre le *euh* et l'allongement vocalique, mais également entre le *euh* et le son [ə] d'appui à la finale de mot (le *e* postonique parisien, voir Candéa 2002).

À propos de cette distinction parfois difficile à établir entre la présence d'un schwa, d'un allongement du schwa et la présence d'un *euh*, Campione et Véronis (2004) notent :

La question semble totalement indécidable, aussi bien à l'écoute qu'au vu des courbes mélodiques et sonogrammes, et on peut se demander si cette « hésitation » sur l'hésitation n'est pas en fait révélatrice de l'identité de fonction de l'allongement et du *euh*.

Ayant conscience de ces difficultés liées à l'établissement du texte, nous considérons toutefois que le *euh* transcrit dans nos corpus correspond bien à une voyelle autonome, et non à l'allongement ou au son [ə]. Ceux-ci, qui relèvent de la phonétique, ne figurent pas dans les transcriptions. Les *euh* transcrits sont donc bien des formes autonomes, ou en tout cas reconnues comme telles par le transcripteur et l'expert ayant validé la transcription.

1.3 *Euh* dit d'hésitation

Il est classiquement admis que le *euh* 1) constitue une pause dans le flux de la parole et 2) marque une hésitation de la part du locuteur. Nous l'avons dit, ces deux « constats » se retrouvent fréquemment dans la littérature qui

aborde le sujet, parfois sans aucune justification ni théorique, ni même empirique. Ainsi, les conventions de transcription du projet PFC (Phonologie du Français Contemporain) qui généralisent :

Les hésitations seront transcrites *euh*.

Dans les trois études de Grosjean et Deschamps (1972, 1973 et 1975), la pause remplie (c'est-à-dire *euh*) est exclusivement une pause d'hésitation².

C'est également la position adoptée par Clerc-Renaud *et al.* (2004) qui étudient ce qu'elles appellent les *hésitations autonomes* dans plusieurs langues. Les auteures associent d'office le *euh* du français à un phénomène d'hésitation.

Candéa (2000a), quant à elle, remet en cause cette terminologie – qui n'est pas suffisamment neutre – qui consiste à assimiler automatiquement et sans nuance le *euh* à une pause. En effet, selon l'auteure,

rien ne permet *a priori* de classer ces marques [euh et allongements] du côté des pauses, leur rôle n'étant pas uniquement de marquer un temps dans la production d'un énoncé mais également d'indiquer à l'auditeur que ce temps est clairement destiné à poursuivre l'encodage et non pas par exemple à céder la parole à l'auditeur.

Campione et Véronis (2004) vont dans le même sens :

Regrouper ces phénomènes [pauses silencieuses et pauses remplies, c'est-à-dire *euh*] acoustiquement différents sous le même terme de pause revient à poser une hypothèse implicite quant à une identité de fonction. Or, rien n'est moins sûr qu'une telle identité.

En ce qui nous concerne, nous éviterons également dans la suite de ce travail le terme de *pause remplie* pour ne garder que la forme *euh*, quitte à être parfois peut-être un peu répétitive.

² Notons cependant que si cette marque est exclusivement associée à un phénomène d'hésitation, elle n'est pas la seule. Ainsi Grosjean et Deschamps ne se limitent pas aux disfluences en tant que telles pour cerner les marques du travail d'encodage du locuteur :

L'existence de ces variables secondaires [pauses sonores, répétitions et faux départs] est manifestement l'illustration d'un encodage laborieux, qu'il soit momentané ou continu, mais il serait erroné de limiter l'hésitation à ce seul domaine. (...) les sujets, se trouvant confrontés à un problème d'encodage, font également appel aux variables temporelles primaires et simples : ils réduisent leur vitesse d'articulation, et augmentent la durée et la fréquence de leurs pauses non sonores. (Grosjean et Deschamps 1973 : 216)

Candéa reprend les hypothèses de Morel et Danon-Boileau (1998 : 82), pour qui le *eah* du français parlé est « l'indice d'une intention de continuation ». Selon ces auteurs, le *eah* est moins le signe de la recherche du mot chez le locuteur que la marque d'une difficulté à organiser le discours, « à organiser la relation entre le segment déjà énoncé et le segment à venir » (id.).

Selon ces auteurs, le *eah* marquerait donc l'intention de continuer. C'est ce que notaient déjà Clark et Clark (1977, cités par Campione et Véronis 2004), pour qui la pause remplie est un signal conventionnel qui sert au locuteur à « occuper le terrain », manifestant sa volonté de poursuivre et donc de ne pas être interrompu par son interlocuteur.

Mais qu'en est-il dans les faits ? Le locuteur poursuit-il après un *eah*, un peu comme si celui-ci n'était pas présent dans l'énoncé, ou ce *eah* marque-t-il une rupture dans l'énoncé ? Pour le dire autrement, le *eah* constitue-t-il une « pause » dans le flux de parole du locuteur qui s'insère, éventuellement comme la pause silencieuse (chap. 4), entre des syntagmes achevés, ou bien le *eah* intervient-il dans l'énoncé non pas à des charnières de syntagmes mais à l'intérieur même de ceux-ci ? Arrive-t-il que le *eah* marque une rupture syntaxique dans la construction de l'énoncé, dans le mesure où celle-ci serait interrompue après le *eah* ?

2 LES *EAH* DANS NOTRE CORPUS

Avant de présenter l'analyse de nos données, deux remarques s'imposent d'emblée :

1. comme nous l'avons dit, les indications de l'allongement vocalique ne figurent pas dans les transcriptions. Nous ne pouvons donc pas étudier la combinatoire du *eah* et de celui-ci. On peut imaginer, se basant sur les résultats de Candéa (2000a), que les résultats que nous obtenons, en termes de marques combinées (deux marques) ou mixtes (plus de deux marques) notamment, auraient été différents. Cette remarque est également valable pour les combinaisons qui concernent les autres marques de disfluency étudiées dans ce chapitre ;
2. le fait de ne pas prendre en compte l'allongement rendra souvent la comparaison de nos données avec d'autres études relativement difficile. De plus, la terminologie non unifiée, que nous avons déjà mise en évidence, rendra souvent toute éventuelle comparaison impossible. En effet, nous l'avons déjà dit, le terme *hésitation* regroupe chez les auteurs que nous avons consultés des phénomènes différents, tout comme les termes *pause pleine* et *pause remplie*. Nous ferons donc des comparaisons entre nos

résultats et ceux d'autres études lorsque cela sera possible, et que nous serons certaine de nous livrer à des comparaisons pertinentes.

Le *eah* apparaît dans notre corpus très fréquemment, puisque nous comptabilisons 9423 occurrences. C'est l'une des formes graphiques les plus fréquentes, puisqu'elle arrive en cinquième position (cf. chap. 2). Parmi les marques disfluentes que nous étudions ici, elle arrive en deuxième position après les répétitions.

Comme le fait remarquer Candéa (2000a : 260),

dans la plupart des corpus de français spontané, le *eah* est la marque de travail de formulation la plus fréquente (...).

Ce n'est pas le cas chez nous, car nous relevons un nombre plus important de répétitions. Mais il est vrai que nous n'avons pas fait le départ entre les répétitions qui relèvent de la disfluence (ce que Candéa appelle le travail de formulation) et celles qui sont stylistiques ou encore exigées par la grammaire, par exemple. De plus, nous avons recherché toutes les formes identiques, sans limite du nombre de mots du répertoire et avec possibilité d'insertion, ce qui n'est pas le cas de toutes les études qui s'arrêtent parfois à la répétition d'une seule forme graphique (cf. chap. 3.2).

Dans notre corpus, cela signifie que l'on trouve en moyenne un *eah* tous les 47 mots.

3 LE *EAH* COMBINÉ À UNE AUTRE MARQUE

Les études antérieures que nous avons consultées montrent une forte interaction du *eah* avec d'autres marques, comme c'est d'ailleurs souvent le cas pour les disfluences en général.

Ainsi, Grosjean et Deschamps (1972 :151) constatent dans leur corpus que

60 % des pauses remplies [c'est-à-dire *eah*] sont précédées ou suivies par un autre type de pause (pour la moitié environ par des pauses non sonores, puis par des syllabes allongées, des faux départs et des répétitions) (...).

Nous allons donc analyser la distribution des *eah* dans notre corpus, et leur interaction avec d'autres marques.

3.1 Le *euh* et la répétition

3.1.1 Le *euh* dans une séquence répétition

Parmi celles-ci, on a vu que 585 occurrences (soit 6,21 % de l'ensemble des *euh* du corpus) se produisent à l'intérieur d'une séquence répétition (cf. chap. 3.2, §5.3.5).

Nous avons extrait au hasard 100 occurrences de *euh* placées dans une séquence répétition, quelle que soit la place de celui-ci (entre répétable et répété ou entre deux répétés).

Ces 100 cas extraits au hasard nous semblent représentatifs dans la mesure où ils illustrent les configurations différentes qui peuvent se rencontrer (qui ne sont pas exclusives l'une de l'autre, comme le montrent les exemples) :

- *le euh est entre le répétable et le premier répété :*

blaAT1 (...) maintenant faut faut voir comment la vie ira
mais / s'il faut avoir une grosse euh une grosse /
une grosse remise en questions comme ça autant
l'avoir euh (...) [blaAT1]

ilpCM1 (...) mais je comprends que dans / euh / dans
dans dans une vision euh / dans une certaine vision on
puisse avoir cette impression (...) [ilpCM1r]

- *le euh est la seule insertion :*

iljBB1 (...) on est là pour rapporter des faits on n'est
pas là pour euh pour éduquer les gens (...) [iljBB1r]

- *le euh cooccure avec une autre marque :*

iljNJ1 (...) mais d'un autre côté quelqu'un
d'extrêmement intelligent et qui a une // euh une
facilité d'élocution peut dire des choses évidemment très
intéressantes même avec un accent (...) [iljNJ1r]

- *le euh cooccure avec plusieurs autres marques :*

ilrTF1 pouvoir regarder le tableau autant que je veux
pour euh comment / pour bien le regarder pour ne
pas l'oublier (...) [ilrTF1r]

- *plusieurs euh sont insérés :*

ilpCM1 (...) mais je suppose que commencer ses
phrases par euh / euh par par un certain nombre d'
expressions qu' on appelle des gallicismes [ilpCM1r]

iljDV1 (...) on n'utilise le wallon que dans certaines circonstances / p/ pour euh à la limite euh m/ mettre en évidence une situation pour insister sur euh l/ euh euh sur euh sur un mot sur un un / mettre en évidence |- un contexte (...) [iljDV1r]

Nous avons écrit ci-dessus que toutes les configurations étaient rencontrées. Mais en fait, en ce qui concerne la place de la marque, nous ne rencontrons pas la configuration où le *euh* serait placé entre deux répétés. Nous avons vu dans le chapitre sur la répétition que cette place était moins fréquente que la position entre répétable et premier répété. Néanmoins, si on ne trouve pas de *euh* dans cette position parmi les 100 occurrences choisies ici, ce type de séquence n'est pas absent du corpus général.

À l'analyse des 100 cas repris, et quel que soit le schéma concerné, on constate que le *euh* n'a aucune incidence sur le fonctionnement de la répétition, en termes de structuration syntaxique de l'énoncé : la séquence reconnue constitue un piétinement en un point de l'axe linéaire de l'énoncé, piétinement accentué par l'accumulation de marques, mais, on constate une reprise du déroulement syntagmatique de l'énoncé avec le dernier répétable.

3.1.2 Le *euh* objet de la répétition

Le *euh* fait lui-même l'objet de répétitions à 233 reprises, réparties comme suit :

- *euh* est répété une fois dans 229 cas ;
- *euh* est répété 2 fois dans 4 cas.

Cela concerne donc 470 occurrences de *euh*, soit 5 % des *euh* du corpus. Parmi ces occurrences, certaines apparaissent à l'intérieur d'une séquence répétition, comme dans le dernier exemple repris ci-dessus.

En observant attentivement des séquences où le *euh* est répété, on ne constate pas de différence avec les occurrences où il apparaît seul.

3.2 Le *euh* et les autres marques : tableau général

Outre ces interactions du *euh* avec le phénomène de la répétition, on a un certain nombre de *euh* qui sont en distribution avec d'autres marques.

Le tableau suivant reprend les principaux cas de cooccurrences que nous avons relevés dans notre corpus (indépendamment donc des cas de répétitions et d'auto-correction déjà abordés ci-dessus) :

Marque	Position par rapport à <i>euh</i>	Nombre d'occurrences	Pourcentage sur l'ensemble des <i>euh</i>
Indication paraverbale	avant	59	0,63 %
Indication paraverbale	après	100	1,06 %
silence	avant	44	0,47 %
silence	après	151	1,60 %
pause longue	avant	454	4,82 %
pause longue	après	549	5,83 %
pause brève	avant	1143	12,13 %
pause brève	après	1743	18,50 %
amorce	avant	91	0,97 %
amorce	après	108	1,15 %
ponctuant/onomatopée	avant	280	2,97 %
ponctuant/onomatopée	après	353	3,75

Tableau 28. Cooccurrence du *euh* et d'autres marques

3.3 Le *euh* et la pause silencieuse

On constate une importante interaction du *euh* et de la pause silencieuse. En effet, 4084 pauses silencieuses se situent directement avant ou après le *euh*. Parmi les trois types de pause, c'est la pause brève qui est la plus fréquente (2886 occurrences), suivie par la pause longue (1003 occurrences) et enfin par le silence (195 occurrences), cet ordre reflétant l'ordre de fréquence décroissant de chaque type de pause dans le corpus (cf. chap. 4).

Par ailleurs, selon les conventions de VALIBEL, seules les pauses particulièrement prolongées, notées (*silence*), sont transcrites en fin de tour de parole. Les autres plus brèves, même si elles sont présentes, ne doivent pas être transcrites. On peut donc faire l'hypothèse que les cooccurrences du *euh* et de la pause silencieuse auraient été encore plus nombreuses si les deux autres niveaux de pauses (la pause brève et la pause longue) étaient transcrits en fin de tour de parole, les *euh* étant relativement fréquents dans cette position (cf. § 3.5.1).

Campione et Véronis (2004) remarquent également une forte interaction entre les deux types de pauses : dans le corpus qu'ils étudient, 58 % des

séquences³ ayant une pause remplie sont au contact d'une pause silencieuse ; inversement, 33% des pauses silencieuses sont en contact avec une pause remplie. Nous verrons que nous constatons cette tendance dans notre étude de la pause silencieuse sur un sous-corpus constitué de trois enregistrements, le *euh* étant la marque la plus fréquente en cooccurrence directe avec une pause silencieuse (cf. chap. 4).

Dans notre corpus, nous avons 32,85 % des *euh* combinés à une pause. En effet, parmi les 4084 occurrences où une pause silencieuse est contiguë à un *euh*, on a 179 séquences qui constituent des marques mixtes, où une pause silencieuse précède et suit le *euh*. Ce chiffre est inférieur à celui obtenu par Campione et Véronis et s'explique sans doute par deux facteurs qui distinguent leurs données des nôtres (2004) : ils ne mettent pas de seuil minimal à la pause non sonore et comptabilisent parmi les pauses sonores les allongements.

Ce qui apparaît de manière régulière dans notre corpus est la répartition de la pause avant ou après le *euh* : dans les 3 types de pauses, la position après le *euh* est préférentielle, avec 2443 occurrences sur les 4084 de pauses contiguës au *euh*, soit dans près de 60 % des cas. On voit que cette tendance est confirmée par la cooccurrence des deux autres marques prises en considération. Ici aussi, c'est après le *euh* que l'amorce et les ponctuels se placent préférentiellement. Mais peut-être faut-il inverser la perspective et dire que le *euh* se produit plutôt avant ces marques qu'après.

Chez Candéa (2000a : 240), la pause silencieuse est également la marque la plus fréquente en combinaison avec un *euh*, représentant 73 % des combinaisons doubles⁴. Les allongements viennent ensuite avec 15 % des occurrences, suivis des répétitions (7 %) et enfin des autocorrections (5 %).

3.4 Le *euh* et les ponctuels

Les ponctuels/onomatopées que nous avons pris en considération ici sont *p*, *pf*, *pff*, *pfff*, *pt*, *f*, *ff*, *bè*, *ben*, *bah*, *boh*, *ah*, *oh*, *m*, *mm*, *mf*, *hum*, *hein*, *bon* et *quoi*.

Étrangement, on ne trouve pas avec la même fréquence les mêmes ponctuels/onomatopées selon la position avant ou après le *euh*. En effet, les ponctuels/onomatopées les plus représentés sont *bon*, *hein* et *quoi* à gauche du *euh*, et *bon*, *m*, *mm*, *pf* et *pff* à droite de celui-ci, *quoi* n'apparaissant que deux fois en tant que ponctuel dans ce contexte.

³ Les auteurs parlent de séquence car la pause remplie comprend le *euh* et l'allongement qui, concomitants et/ou répétés, forment une séquence.

⁴ Et pas 73 % des occurrences de *euh*.

Il est intéressant de constater que dans toutes les séquences *euh* + *bon* ou *bon* + *euh*, *bon* est effectivement un ponctuant, sans aucune exception (220 occurrences) et non pas un adjectif ou un nom. Cette distribution est un bon indice pour la construction des grammaires de désambiguïsation (cf. chap. 6).

Notons que cette règle ne peut s'appliquer pour la forme *quoi*, qui peut ne pas être un ponctuant dans ce contexte (et n'a donc pas été comptabilisée comme telle ci-dessus).

Par ailleurs, les chiffres du tableau 28 comptent les occurrences précédant le *euh* indépendamment de celles qui lui sont subséquentes. Or, nous avons 17 séquences dans lesquelles un ponctuant entoure le *euh* de part et d'autre, comme dans l'exemple suivant :

ilrCP1 ben euh boh ça dépendait euh pff quinze vingt
kilomètres par jour avec de temps en temps euh de
jours de repos (...) [ilrCP1r]

Nous reviendrons ci-dessous sur ces séquences mixtes.

3.5 Le *euh* « continuatif » : fin de tour de parole et chevauchement

3.5.1 Le *euh* en fin de tour

Dans notre corpus, le *euh* apparaît en fin de tour de parole à 436 reprises (458 si l'on compte aussi les fins de chevauchements internes), concernant ainsi 4,6 % des occurrences de *euh* dans le corpus. À ces occurrences s'ajoutent encore les 21 *euh* suivis d'un silence en fin de tour de parole (parmi les 151 repris dans notre tableau), ce qui nous fait un total de 5 % de l'ensemble des *euh*.

Nous l'avons dit, selon Morel et Danon-Boileau (1998), le *euh* est une marque qui indique que le locuteur a l'intention de continuer son propos. Les auteurs vont même plus loin, en ajoutant que ce signal que donne le locuteur sur son intention de poursuivre son énoncé est identifié comme tel par l'interlocuteur :

Il est intéressant de noter qu'un locuteur n'est jamais interrompu après la production d'un *euh*. (Danon-Boileau et Morel 1998 : 82, note de bas de page)

Candéa (2000a) confirme cette affirmation sur ses données enregistrées en classe de français, avec quelques contre-exemples qu'elle explique aisément (il s'agit toujours d'une interruption de l'institutrice, et les

répliques sont très courtes) et qui, selon elle, ne viennent pas contredire l'hypothèse

selon laquelle la production d'un *euh* est toujours correctement interprétée par l'auditeur comme une marque forte d'inachèvement et comme un indice non-ambigu [sic] de la part du locuteur de son intention de continuer (...) (Candéa 2000a : 242).

L'auteure nuance néanmoins :

(...) un *euh* n'est jamais interprété par l'auditeur comme une marque de clôture et comme une invitation à prendre la parole ; ainsi, un locuteur qui vient de produire un *euh* ne sera pas plus souvent interrompu qu'un locuteur qui marque un léger arrêt à l'intérieur d'un syntagme manifestement inachevé. (Candéa 2000a : 242-243)

Si l'on considère, à l'instar des auteurs cités, que le *euh* est un indice continuatif que donne le locuteur à son interlocuteur, on peut formuler l'hypothèse que lorsque le *euh* apparaît en fin de tour, le locuteur « perd » véritablement la parole. Dans notre corpus, la stratégie du *euh* pour occuper le terrain se révèle non payante dans 5 % des cas. Le *euh*, morphème dans la prise de parole, constitue ainsi pour l'interlocuteur l'occasion de prendre ou de reprendre la parole, comme c'est le cas dans les exemples suivants :

accBF1 mais à part ça ça dépend des gens // quelques-uns me disent que j'ai un accent puis d'autres non

accCT0 et si tu devais dire c'est un accent d'où euh

accBF1 un accent de Charleroi [accBF1r]

blaJV1 ben non non moi m/ non si il change d'avis moi il y a pas de problème mais je sais bien qu'il changera pas d'avis quoi je

blaND0 mais tu vas pas aller pleurer après lui |- pour euh

blaJV1 non sûrement -| pas {ça} alors là / hum [blaJV1l]

blaNB1 un gros manitou // et euh / il / il fait une villa privée / ça on est quand même contents / hein |- parce qu'on nous avait <blaAD1> oui ça sera pas -| quand même parlé de choses euh

blaAD1 d'un lotissement |- ou un truc comme ça

blaNB1 ouais -| ouais ouais (...) [blaNB1l]

3.5.2 Le *euh* dans les chevauchements de parole

Le *euh* apparaît également dans les chevauchements de parole. En début de chevauchement, on a 124 *euh* chez le locuteur qui a la parole. On peut formuler l'hypothèse que le *euh* a été senti par l'interlocuteur comme une opportunité de prendre la parole, et que là aussi la stratégie d'occupation du terrain n'a pas marché. D'ailleurs, dans près de 30 % de ces cas (37 occurrences), ce locuteur non seulement prend la parole mais la garde effectivement, le *euh* étant la seule portion de parole superposée pour le 1^{er} locuteur, comme dans l'exemple suivant :

famVI1 au Laos / euh le principe c'est qu'il y a pas
d'ATM // donc il y a pas d' appareil où tu peux retirer de
l'argent // donc tu dois estimer tes dépenses / |- euh
famVV1 pour un -| mois (rire)

De la même manière, on a des *euh* qui précèdent directement le chevauchement de parole dans 76 cas. Ainsi, le pourcentage des cas où le locuteur perd la parole suite à un *euh*, ou en tout cas où l'interlocuteur tente de la prendre, s'élève à pratiquement 7 %⁵ des occurrences du *euh* en corpus.

3.5.3 Le *euh* en début de tour de parole

En début de tour de parole, on a 547 occurrences de *euh* (dont 12 en début de chevauchement interne).

Cette position, tout comme celle de fin de tour de parole, ne pose pas de problème lors de l'étiquetage (cf. chap. 4).

3.6 Séquences de marques et *euh* seul

En fait, on remarque que souvent, le *euh* se combine non seulement à une autre marque parmi celles envisagées ici, mais à plusieurs d'entre elles⁶. C'est, dans la terminologie de Candéa (2000a), des *marques mixtes*, le terme *marques doubles* étant réservé aux suites de deux marques du travail de formulation.

Ainsi, Grosjean et Deschamps (1973 : 225) constatent

⁵ 436+21+124+76= 657

⁶ Rappelons que les chiffres donnés dans le tableau 22 concernent les marques directes et non des séquences de celles-ci. Ainsi, par exemple, un ponctuant séparé d'un *euh* par une pause n'était pas comptabilisé dans la colonne des ponctuants.

un regroupement de ces pauses entre elles et avec les pauses non sonores [*euh*, allongement, répétition et faux départs] : une seule pause d'hésitation ne suffit plus et le sujet en introduit deux ou trois à la suite afin de retarder d'autant plus le mot ou la phrase qui doit suivre.

Nous avons déjà vu l'interaction du *euh* et du phénomène de la répétition (qu'il s'agisse du *euh* à l'intérieur d'une séquence répétition ou du *euh* lui-même répété), de la pause simple qui entoure le *euh*, ainsi que les séquences *onomatopée/ponctuant* + *euh* + *onomatopée/ponctuant*.

Comme le font remarquer Grosjean et Deschamps (1972 : 151) :

Il est intéressant cependant de constater une certaine corrélation ($r=0,56$ et $0,40$) du nombre de pauses sonores avec le nombre de répétitions et de faux départs ; ceci tend à prouver que le sujet qui met des pauses d'hésitation dans son langage a tendance à se servir de toutes les catégories de pauses possibles qui sont à sa disposition : pauses sonores, répétitions et faux départs.

Là où les pauses non sonores sont distribuées dans le texte à des intervalles plus ou moins réguliers (la fonction démarcative de la pause silencieuse, qui scinde le flux de la parole en segments plus ou moins courts), on constate une répartition moins égale des marques disfluentes, qui peuvent se regrouper à certaines places de l'énoncé. Ce regroupement ajoute encore à l'effet de piétinement, d'entassement paradigmatique, comme dans l'exemple suivant où coexistent répétition, auto-correction, ponctuant et *euh* :

ileDC1 (...) parce que la la langue wallonne **euh bon**
la le le pas la langue bon mais le le dialecte ou le / ou
les langues wallonnes (...) [ileDC1r]

Ici, on voit l'effort auquel se livre le locuteur dans la recherche de la dénomination adéquate, du « mot juste » qui le satisfera, ainsi que son interlocuteur. Le locuteur passe de *la langue wallonne* à *le dialecte* pour enfin arriver à la formule *les langues wallonnes* qui semble mieux lui convenir.⁷ Le travail de formulation est clairement visible ici, où s'accumulent plusieurs marques directement juxtaposées.

⁷ À la lecture de la transcription, on pourrait penser – et, on l'a dit, c'est souvent l'idée reçue en matière de disfluences – que l'on a affaire à un locuteur peu doté en termes de compétences linguistiques. Or, il s'agit ici d'un journaliste de la presse écrite ayant fait des études supérieures, qui s'exprime dans un contexte peu contraint, même s'il est vrai que les rôles sont inversés puisque ce n'est pas lui qui pose les questions.

Or, on l'a dit, les chiffres que nous donnons dans le tableau ci-dessus ne concernent que les marques simples, et non des séquences de celles-ci. Ainsi par exemple, si l'on permet une pause non sonore entre le *euh* et un ponctuant à droite, on obtient 78 séquences supplémentaires qui comprennent un ponctuant⁸ ; de même à gauche du *euh*, où le nombre de séquences ayant un ponctuant augmente alors de 36 occurrences.

En fait, les *euh* « seuls »⁹ sont minoritaires sur l'ensemble de notre corpus, avec 3703 occurrences. C'est donc bien la combinaison des marques qui semble être la règle, plutôt que l'occurrence simple de la pause remplie¹⁰.

Cette constatation peut être étendue aux autres marques de disfluece que nous abordons dans ce chapitre. En effet, si l'on en croit Candéa – qui s'est particulièrement intéressée aux interactions de la pause silencieuse avec les autres marques du travail de formulation, et de celles-ci entre elles – les séquences de marques du travail de formulations sont plus nombreuses que les marques isolées :

Si l'on considère la totalité des marques de travail de formulation attestées dans notre corpus, le nombre d'occurrences simples est **inférieur** au nombre total d'occurrences qui font partie d'une marque mixte ou au moins d'une combinaison double. La situation est différente en ce qui concerne les pauses silencieuses : le nombre d'occurrences simples (encadrées par des séquences sonores sans marque de travail de formulation) est très **largement supérieur** à celui des pauses silencieuses ayant dans leur contexte immédiat au moins une marque de travail de formulation. (Candéa 2000a : 142)

L'auteure poursuit en allant dans le même sens que Grosjean et Deschamps cités plus haut, qui établissent une distinction entre pause non sonore et pause sonore quant à leur caractère obligatoire :

⁸ Ces séquences sont évidemment déjà comptabilisées dans le tableau au niveau de la pause simple.

⁹ Nous entendons également par « seul » que le *euh* ne se trouve pas dans les positions spécifiques de début et de fin de tour de parole, positions qui reçoivent un traitement particulier lors de l'étiquetage.

¹⁰ Notons que chez Candéa (2000a), le *euh* semble faire exception à cette tendance de regroupement des marques, puisque pratiquement 50 % des *euh* apparaissent seuls dans son corpus. Ces résultats se distinguent nettement des nôtres, tendance qui aurait été accentuée sans doute si l'on avait pris en compte l'allongement dans notre étude. Le fait de considérer chez nous les indications paraverbaux comme une marque, démarche purement pratique en vue de la procédure d'étiquetage, ne remet pas en cause cette divergence de résultats.

Cette différence provient du fait que la pause silencieuse est principalement une trace de la structuration maîtrisée de l'énoncé et en cela elle a un caractère obligatoire et dans une large mesure prédictible : une simple pause silencieuse suffit pour remplir son rôle, et la durée dont elle sera affectée participera à la hiérarchisation des constituants qui la précèdent et la suivent. En revanche, les marques de travail de formulation sont principalement la trace d'une structuration temporairement non maîtrisée de l'énoncé, d'où leur caractère facultatif et imprévisible : cela favorise l'accumulation des marques et la réinitialisation des hiérarchies déjà en place. (Candea 2000a : 142-143)

3.7 Place syntaxique du *euh*

La question de la place syntaxique des disfluences est liée à celle du contexte et à la notion de constituant. Elle est cruciale pour l'analyse des différentes marques prises en considération. La définition du constituant relève non seulement d'un choix méthodologique mais également d'un choix théorique.

Ainsi, quand Morel et Danon-Boileau (1998) analysent le contexte des marques de travail de formulation, c'est en fonction du découpage du texte en paragraphes oraux, découpés eux-mêmes en d'autres constituants qui sont les unités pertinentes de leur modèle théorique. Chez ces auteurs, c'est l'aspect intono-syntaxique qui prédomine pour définir ces unités.

Notre critère exclusif sera celui de la syntaxe, mais vue de manière moins rigoureuse que dans certaines études (celles Grosjean et Deschamps par exemple), notamment pas en termes de dépendance. Nous avons déjà eu l'occasion d'aborder les marques précédentes selon la rupture syntaxique ou non de l'énoncé. Dans le cas du *euh*, il nous faut définir ce que nous entendons par groupe syntaxique.

La définition du constituant n'est pas toujours aisée, et on constate qu'elle est parfois relativement implicite. Le constituant pris en compte pour établir la position des disfluences dans l'énoncé varie aussi considérablement d'une étude à l'autre, avec des contours plus ou moins nettement définis. Ainsi, Campione et Véronis (2004) appellent les groupes qu'ils dégagent des énoncés oraux des *segments discursifs*, s'appuyant sur les études des psycholinguistes qui ont montré que

la production de la parole relève d'un cycle planification-exécution, qui résulte en une succession d'unités relativement brèves (que nous appellerons ci-après segments discursifs), séparées par des pauses silencieuses.

Quant à nous, nous appelons constituant syntaxique un groupe relativement comparable au *chunk*, syntagme élémentaire, défini par Abney (1991 et 1996). Il s'agit d'un segment intermédiaire entre le mot et la proposition ou la phrase, dans lequel un élément lexical sert de tête : un nom pour le syntagme nominal et un verbe pour le syntagme verbal. Nous verrons que, comme Vergne (1999), nous considérons le pronom sujet comme faisant partie du syntagme verbal.

Chez Campione et Véronis (2004), la majorité des séquences de pauses remplies¹¹ (81 %) se placent à l'intérieur d'un segment discursif (défini comme « une unité relativement brève ») et « marquent une interruption, suivie ou non d'une reprise et/ou d'une réparation ». On regrette de ne pas avoir de chiffres de sous-catégorisation : on ne sait en effet pas combien de cas constituent une véritable interruption de la construction syntaxique de l'énoncé (c'est-à-dire pour nous, sans que le syntagme soit achevé).

Dans leur étude, seulement 19 % des occurrences se rencontrent en début de segment discursif. Selon Campione et Véronis (2004) :

Dans de tels cas, il ne s'agit pas d'une interruption liée à une difficulté de mise en place lexicale ou syntaxique de la fin du segment. Il s'agit sans doute d'« occuper le terrain » en attendant de trouver une suite au discours, et éviter ainsi que l'interlocuteur prenne la parole.

Bénard (2005) travaille dans le cadre du projet COPTE. Rappelons que ce projet se base sur des transcriptions réalisées à destination de la presse (donc fortement standardisées : sans transcription des marques de disfluences et ponctuées selon les normes de l'écrit). Considérant comme pause remplie les *euh* et les allongements, la chercheuse obtient, dans une étude sur 88 000 mots, 21 % de pauses remplies à la frontière de syntagme (« pour laquelle la transcription de la presse comporte une virgule ») et 14 % à la frontière d'une phrase (on imagine que cela signifie « là où il y a un point » dans la version écrite pour la presse, mais l'auteure ne précise pas). Mais en fait, les autres occurrences ne signifient pas nécessairement qu'il y a rupture syntaxique puisque le *euh* peut très bien apparaître à la frontière de syntagmes sans que ceux-ci soient séparés à l'écrit par une virgule (typiquement entre le sujet et le verbe, par exemple). Ces chiffres ne nous éclairent donc guère, et ne peuvent en tout cas pas servir de point de comparaison pour nous. Néanmoins, ce type de formulation peut faire penser que ces marques d'hésitation se produisent majoritairement en des lieux non adéquats selon la structuration syntaxique.

Dans leurs trois études sur les variables temporelles, Grosjean et Deschamps (1972, 1973 et 1975) analysent également la distribution des

¹¹ Rappelons que les auteurs considèrent également l'allongement parmi la pause remplie.

marques qu'ils étudient, et donnent des résultats chiffrés sur leur position dans l'énoncé. Leur critère est un critère syntaxique : il consiste à observer la place occupée par la marque soit par rapport au constituant *syntagme* (entre deux syntagmes ou à l'intérieur d'un syntagme), soit par rapport au constituant *phrase* (entre deux phrases ou à l'intérieur de la phrase – et donc, alors, en fonction de la position par rapport aux syntagmes).

Les auteurs obtiennent 44,51 % de *pauses remplies* (c'est-à-dire de *euh*) en fin de phrase, c'est-à-dire « au point de rupture grammatical de la chaîne parlée le plus normal » (Grosjean et Deschamps 1972 : 146).

Parmi les 55,49 % des occurrences à l'intérieur de la phrase, Grosjean et Deschamps relèvent 38,5 %¹² des pauses remplies en dehors de ce qu'ils appellent des « points de rupture grammaticaux ». Quand on observe le détail de leurs données (étude de 1972, tableau IV, p. 148), on voit que 24,93 % des pauses apparaissent à l'intérieur du SN ou du SV¹³, et 5,34 % entre une préposition et le SN que celle-ci introduit. Grosjean et Deschamps remarquent également 5,34 % de pauses remplies après un faux départ ou une répétition. Ils ne précisent malheureusement pas si ces marques constituent ou non une rupture dans la construction syntaxique de l'énoncé, dans la mesure où celle-ci peut être abandonnée ou pas après le *euh*.

Quand ils comparent les résultats qu'ils obtiennent sur des interviews radiophoniques (étude de 1972 dont nous venons de présenter les résultats) à ceux issus de l'analyse de descriptions de dessins humoristiques (étude de 1973), les auteurs constatent un plus grand nombre de pauses remplies (c'est-à-dire *euh*) à l'intérieur de phrases dans ce dernier cas (47,4 %). Cette divergence, qu'ils ne constataient pas pour les pauses non sonores, s'explique selon eux par le fait que la tâche de description est plus contraignante pour le locuteur et amène celui-ci à mettre des « marques d'hésitations » à des endroits moins attendus, qui respectent moins la structure grammaticale de la phrase.

Candéa (2000a : 240), quant à elle, remarque que les *euh* suivis d'une pause silencieuse (75 % des cas de combinaisons du *euh* avec une autre marque) n'introduisent pas de rupture syntaxique dans l'énoncé. Elle constate la même régularité pour la combinaison du *euh* et de l'allongement. L'auteure considère d'office la répétition et l'autocorrection comme des occurrences de rupture syntaxique – ce qui n'est pas la position que nous adoptons¹⁴ –, a fortiori dans les cas où ces disfluences sont combinées avec la pause (respectivement 7 % et 5 % des combinaisons).

¹² Contre 25 % pour les pauses sonores.

¹³ Cf. chap. 4 sur la pause vide pour les objections que nous faisons.

¹⁴ Cf. chap. 3.2 et 3.3.

De manière générale, sur les 469 occurrences de *eah* que comporte son corpus, Candéa (2000a : 244) remarque que dans 76 % des cas, l'enchaînement entre ce qui précède et ce qui suit se fait sans aucune rupture syntaxique. Sont concernés, nous venons de le dire, les cas de combinaison double à l'exception de l'auto-correction et de la répétition, mais aussi, et c'est quand même relativement surprenant, tous les cas d'occurrence simple du corpus. En effet, on pourrait s'attendre à ce que cette marque assimilée traditionnellement à des difficultés du locuteur soit un lieu d'achoppement dans la linéarité de l'énoncé, lieu qui indique une rupture syntaxique (plus qu'un « simple » piétinement sur l'axe syntagmatique donc). L'auteure ajoute (p. 244) :

Le *eah* apparaît ainsi être non seulement une marque autonome de l'inachèvement d'une structure en cours mais aussi une marque de **continuité sans rupture syntaxique**¹⁵ : il n'indique pas seulement qu'il y aura une suite mais également que cette suite sera parfaitement intégrée dans la structure syntaxique laissée ouverte avant le *eah*.

En fait, le nombre de ruptures syntaxiques est plus important dans les combinaisons mixtes de Candéa, car l'allongement y est souvent présent. Nous ne pouvons évidemment pas en tenir compte et comparer de ce point de vue ses résultats aux nôtres, des marques doubles chez nous pouvant correspondre à des combinaisons mixtes chez elles (c'est-à-dire marque double + allongement).

Le nombre d'occurrences de *eah* dans la totalité de notre corpus étant trop élevé, nous avons eu recours à l'utilisation d'un sous-corpus afin d'observer le comportement du *eah* en termes de point de rupture syntaxique éventuel de l'énoncé.

Le sous-corpus que nous avons choisi est composé de trois textes : deux conversations informelles (blaAT11 et famVV1r) et une entrevue (accPH1r).

La répartition des *eah* dans les trois textes qui composent ce sous-corpus est la suivante :

¹⁵ Le gras et l'italique sont de l'auteure.

corpus	nombre de mots	nombre de <i>euh</i>	pourcentage
blaAT1l	8177	209	2,56 %
famVV1r	4220	98	2,32 %
accPH1r	8111	107	1,32 %
TOTAL	20 508	414	2,02 %

Tableau 29. Répartition du *euh*

Nous considérons des groupes syntaxiques élémentaires. Les constituants que nous envisageons peuvent donc être très courts et, bien que cela ne soit guère discuté dans leur article, notre constituant correspond sans doute *grosso modo* au « groupe discursif » de Campione et Véronis (2004).

Nous ne redétaillons pas pour ce sous-corpus les cas déjà évoqués ci-dessus où le *euh* est suivi et/ou précédé d'une pause, les cas où il cooccure avec l'amorce de morphème ou entre dans une séquence répétition ou d'autocorrection immédiate. Tous ces cas sont représentés dans le sous-corpus pris ici en considération, de même que les positions début et fin de tour de parole ou encore début de chevauchement. La fréquence relative de ces occurrences pour ce sous-corpus correspond à celle que nous avons observée sur le corpus global.

Il arrive que le *euh* se place entre des syntagmes, dans des positions particulières qui sont celles que nous avons également rencontrées pour la pause silencieuse. Cette position intersyntagme n'est pas « classique » en termes d'analyse syntaxique, puisque le *euh* peut apparaître :

- entre un sujet et un verbe¹⁶ :

blaAT1 (...) tu vois de t/ tous nos copains euh / sont universitaires aussi et je crois que ça doit le trotter un peu tu vois [blaAT1l]

- entre un verbe et son complément d'objet :

accCT0 oui // et est-ce que parfois toujours en français qu'on ait euh plus ou moins d'accent // euh je sais pas moi est-ce qu'un Liégeois va comprendre un Namurois // avec des accents différents [accPH1r]

- entre un nom et un complément du nom :

¹⁶ Notons cependant que c'est le seul cas que nous ayons trouvé dans le sous-corpus.

famVV1 c'est une panne euh d'alternateur
[famVV1r]

accPH1 (...) mais nous étions / avec // tous des // de
Charleroi / |- puisque c'était dans un <accCT0> oui -|
dans un pensionnat // euh de mutuelles
socialistes // alors on savait // (...) [accPH1r]

- entre un nom et la proposition relative qui s'y rattache

blaAT1 enfin {c'est} une formation euh qui a duré
quoi euh six mois je crois (...) [blaAT1]

Comme nous le verrons lors de notre analyse de la distribution des pauses silencieuses (cf. chap. 4), nous ne considérons pas ces emplacements comme gênants. En effet, les grammaires de désambiguïsation que nous allons construire se baseront notamment sur une analyse qui prendra en compte les syntagmes « élémentaires » et pas nécessairement la dépendance entre eux.

Les positions qui nous intéressent donc ici sont celles où le *euh* soit apparaît à l'intérieur d'un syntagme, soit marque une rupture dans la construction de l'énoncé. Voici les deux cas de figure qui peuvent alors se présenter :

- le *euh* à l'intérieur d'un constituant suspend momentanément la linéarité de l'énoncé, qui reprend ensuite normalement, comme dans l'exemple suivant :

ilrCP1 (...) enfin bon ça ne se/ sert pas à grand-chose
concrètement quoi / et euh bon pour les marches euh
fléchées parce qu'il y a aussi des marches fléchées / là
il faut pas euh c'est plus facile il suffit de suivre les
flèches (...) [ilrCP1r]

Le *euh* se place ici à l'intérieur du syntagme, entre le nom et l'adjectif postposé. Il ne fait qu'interrompre momentanément la construction syntaxique du syntagme qui, sans celui-ci, est formé de manière tout à fait classique.

- le *euh* marque l'emplacement d'une rupture à l'intérieur de la construction syntaxique, le locuteur n'achevant pas l'énoncé selon la structure entamée. Deux cas de figure peuvent ici se présenter :
 - le locuteur abandonne la construction entamée. Il s'interrompt complètement dans la structure qu'il a amorcée, éventuellement parce qu'un autre locuteur prend la parole. Précisons que les *euh* en fin de tour de parole n'actualisent pas tous ce cas. Il arrive en effet fréquemment que le syntagme qui le précède soit

achevé ; tous les *euh* dans cette position qui n'est pas rare ne constituent donc pas une rupture syntaxique. Le locuteur peut également laisser sa structure inachevée et entamer une nouvelle construction sans être en position de fin de tour de parole. Voici trois exemples qui illustrent ces configurations :

accCT0 oui / ah il y a encore **une forte euh**

accPH1 oui mais oui

accCT0 entre l'accent de la mer belge [accPH1r]

famVV1 mais vendredi **je / euh Olivier** arrive vers vingt-et-une heures |- à Bruxelles <famDA1> (crie sur Ivan) Ivan [famVV1r]

accPH1 (...) ça a été ma petite copine **pendant euh** enfin tu vois |- façon <blaND1> m -| de parler quoi / enfin je je j'habitais t/ j'habitais quasi là à un |- moment quoi <blaND1> c' est vrai ? -| [accPH1r]

Dans le premier exemple, le syntagme *une forte* est interrompu à l'emplacement du *euh* après l'adjectif *forte*. Celui-ci, en fin d'énoncé, pourrait être analysé erronément comme un nom dans un syntagme complet. Dans le deuxième exemple, on voit la construction laissée en suspens après la préposition. Dans la suite, le locuteur entame une nouvelle construction, qui ne corrige pas celle qu'il avait entamée (en l'occurrence, on n'a pas un syntagme qui commence par une préposition qui pourrait remplacer celui initié par le *pendant*). Le troisième exemple illustre le même phénomène que l'exemple précédent : on voit une nette interruption après le *euh*, puisque la place syntaxique introduite par le pronom sujet *je* reste vide et qu'une autre construction est entamée, avec un nouveau sujet.

- b) le locuteur abandonne la structure syntaxique entamée, mais pour la reprendre en modifiant le lexique. Cette dernière configuration est parfois appelée *réparation* dans la littérature. Elle s'apparente à ce que nous entendons dans ce travail par *autocorrection immédiate*. Il s'agit en effet du même phénomène, qui varie seulement dans le fait que nous avons sélectionné l'autocorrection en fonction des mots grammaticaux, pour des raisons de possibilité de formalisation.

Voici un exemple d'énoncé réparé :

accPH1 non non non non -| non non non non non /
et quand mes grands-parents / habitaient // à

Dampremy // ma grand-mère me parlait flamand et je devais répondre en français / c'est le tort que j'ai eu // |- parce que <accCT0> ah bon -| il est ma grand-mère **c'est une euh c'est elle c'est une Salbet** // c'est un nom flamand ça hein / et je devais répondre en français c' est le tort que j'ai eu [accPH1r]

La répartition dans notre sous-corpus de *euh* intrasyntagmes ou marques de rupture est la suivante :

	<i>euh</i> seul		<i>euh</i> combiné à une autre marque	
	intrasynonyme	rupture de construction	intrasynonyme	rupture de construction
blaAT1l	4	4	2	11
famVV1r	3	0	2	1
accPH1r	1	1	1	9
	8	5	5	21

Tableau 30. Répartition du *euh* intrasyntagme et du *euh* à l'emplacement d'une rupture syntaxique

Sur les 414 occurrences que comporte notre sous-corpus, nous avons 13 occurrences de *euh* à l'intérieur d'un syntagme et 26 à l'emplacement direct d'une rupture syntaxique. Cumulés, ces cas représentent 9,37 % des *euh*. En fait, nous ne tenons pas compte ici des occurrences où le *euh* entre dans une séquence de répétition ou d'autocorrection. En effet, comme nous l'avons vu ci-dessus, ces derniers cas ne constituent pas une rupture syntaxique de l'énoncé (mais ils pourraient être considérés comme des occurrences de *euh* à l'intérieur d'un syntagme achevé).

Malgré ce nombre relativement faible d'occurrences, nous pouvons tirer certaines conclusions :

1. On a plus de *euh* qui marquent l'emplacement d'une rupture syntaxique (26 cas, soit 6,28 %) que de *euh* à l'intérieur d'un syntagme achevé (13 cas, soit 3,14 %). En fait, seuls ces derniers cas sont problématiques pour l'étiquetage, et on aurait pu considérer le *euh* comme une bonne marque de fin de constituant, au même titre que la pause silencieuse.
2. Lorsqu'il est seul, le *euh* est plus souvent intrasyntagme que marque de rupture.
3. Cette tendance s'inverse lorsque le *euh* est combiné à une autre marque : il indique plus fréquemment une rupture de construction qu'une interruption momentanée dans la structure du syntagme.

4. Le *euh* intrasyntagme se place généralement après la préposition.
5. Les marques combinées les plus fréquentes sont la pause et les ponctuants. La fréquence de combinaison des ponctuants est due au locuteur blaAT11, dont nous verrons qu'il affectionne ce type de marque du discours (cf. chap. 4).

On a des cas où une marque mixte relativement longue n'empêche pas le syntagme de poursuivre, sans reprise du mot-outil. L'accumulation des marques ne peut donc être perçue comme un indice de rupture de la construction syntaxique. Ainsi, l'exemple suivant où la préposition est séparée de la suite par la combinaison (l'accumulation) du *euh*, de la pause brève et de deux ponctuants (*enfin* et *tu vois*) :

blaAT1 (...) on avait un trou entre euh / enfin tu vois
midi et cinq heures et alors on aimerait bien aller voir
Le-seigneur-des-anneaux (...) [blaAT1]

Dans l'exemple suivant, le déroulement syntagmatique se poursuit normalement après l'insertion d'une séquence relativement longue (ponctuant + *euh* + ponctuant) entre le sujet nominal et le verbe, sans reprise du sujet par un pronom :

ilrLD3 (...) ici bon les filles oufti euh bon parlent
vraiment euh tu as vu celui-là bè ça ça me répugne quoi
(...) [ilrLD3r]

Dans les chiffres que nous donnons ci-dessus, nous analysons les contextes gauche et droit strictement contigus au *euh* (ou à la séquence disfluente dans laquelle celui-ci entre). Ainsi, nous n'avons pas repris les cas où la rupture ne se marque pas directement à l'emplacement du *euh* mais est décalée (antérieurement ou postérieurement) de plusieurs mots graphiques. Ainsi, dans l'exemple suivant, la rupture se fait après le mot qui suit la séquence disfluente (*euh* + pause brève) :

ilrCP1 euh ben oui euh / je plus souvent avant que
maintenant parce que bon euh disons que mes parents
surtout mon père a il a un un hobby c'est euh la marche
[ilrCP1r]

blaAT1 (...) moi je me reconnais vraiment dans ce
comportement-là tu vois / pour le par rapport euh à
la licence enfin il en parle t/ il en parle toutes les
semaines tu vois (...) [blaAT1]

Nous n'avons pas évalué le nombre de cas concernés par ce phénomène et cela pour plusieurs raisons :

1. il est difficile d'évaluer la distance pertinente à prendre en considération entre l'emplacement du *euh* et l'emplacement de la rupture syntaxique ;
2. il est également difficile – et par ailleurs une analyse de ce type sortirait du cadre de notre travail – d'établir avec certitude si la présence du *euh* est à mettre en rapport ou non avec la rupture syntaxique ;
3. dans l'hypothèse où les cas de ruptures syntaxiques liés au *euh* seraient très fréquents dans ces contextes élargis, cette configuration serait de toute façon difficilement formalisable dans une grammaire locale (tant au niveau du prétraitement qui concerne le découpage du texte qu'au niveau de l'écriture de grammaires de levée d'ambiguïtés, cf. chap. 4 et 6).

4 CONCLUSIONS

Nous avons analysé une marque disfluente qui, parmi celles que nous avons abordées dans cette section, arrive en deuxième position avec 9423 occurrences. Parmi les formes du corpus, c'est l'une des plus fréquentes puisqu'elle se place en cinquième position. On la rencontre en moyenne tous les 47 mots.

On a vu les interactions du *euh* avec les autres marques que nous avons prises en considération, et le fait que dans le cadre d'une séquence répétition par exemple, son comportement syntaxique ne changeait pas par rapport aux cas, minoritaires, où le *euh* est seul (c'est-à-dire sans autres marques). C'est avec la pause que le *euh* cooccure le plus souvent, et nous y reviendrons lorsque nous aborderons la section consacrée aux pauses silencieuses.

Dans la majorité des cas, le *euh* n'a pas d'influence sur le déroulement syntaxique de l'énoncé : il le suspend momentanément, que ce soit à l'intérieur d'un syntage ou à la frontière de deux syntagmes.

Dans les autres cas (6,28 % dans le corpus analysé ici), où il marque l'emplacement d'une rupture syntaxique, il est très difficile de trouver des indices évidents qui nous permettraient de traiter le *euh* de manière assurée lors de la construction de grammaires locales. En effet, certains *euh* seuls constituent une rupture syntaxique de l'énoncé, tandis que des séquences de marques qui intègrent le *euh* sont soit à l'intérieur d'un syntage, soit à la frontière de constituants.

Ainsi, dans l'énoncé suivant, on a six *euh*, que nous numérotons pour plus de facilité :

blaAT1 oh c'était - l à c'était l'horreur hein attends on
va chez les Espagnols il vient avec quoi enfin c'est

l'horreur quoi tu vois tu te souviens |- de ça <blaND1>
 mm -| / enfin **qu'est** qu'est-ce qu'il fout là quoi / et et je
 vois les photos et je dis ah ah un intrus quoi enfin c'est
 vraiment et tout le temps quoi / les sorties c'était hors
 de question je **m/** / combien de fois je me souviens **euh**
[1] il y avait une soirée carnaval-là |- / et <blaND1>
 mm -| c'était c'était je vivais un peu enfin **ce que mes**
parents euh [2] / enfin je sais |- pas si <blaND1> mm
 -| c'était c'est le le gros stéréotype de la fille qui
 reproduit mais **euh [3]** / enfin tu vois **euh [4] les je**
 me souviens d'une soirée carnaval où tu avais même été
 super dégue et / et moi j'étais super mal parce que / on
 on devait y aller puis **Mick euh [5]/ Mick** c'est du
 genre oui parce qu'**il i/ euh [6] i/ je** n'y allais pas sans
 lui c'était sûr je ne sais pas |- pourquoi je n'y allais pas
 <blaND1> mm -| sans lui

Les *euh* [1] et [3] n'ont pas de répercussion sur la linéarité de l'énoncé. Le [5] est à l'intérieur d'une répétition, tout comme le [6], qui est quant à lui dans une séquence beaucoup plus complexe puisqu'on a une suite d'autocorrection, d'amorce et de répétition ; dans ces deux cas, on peut aussi estimer qu'on peut les ignorer. Les exemples [2] et [4] indiquent une rupture, mais d'ordre différent : dans le [2], celle-ci intervient après un groupe syntaxiquement complet tandis qu'en [4], on a bien une rupture mais celle-ci se produit après le morphème qui suit le *euh* (et non après le *euh*).

On voit par ailleurs que dans cet extrait, d'autres lieux marquent un inachèvement.

L'énoncé ci-dessus illustre bien, nous semble-t-il, la complexité des réalisations du *euh*, et il nous faudra traiter au mieux ce morphème lors de l'étiquetage.

CONCLUSIONS SUR LES DISFLUENCES

Dans ce chapitre, nous avons analysé de manière approfondie quatre types de disfluences : les répétitions, les autocorrections immédiates, les amorces de morphèmes et les *euh*. Les répétitions étaient les plus nombreuses, suivies des *euh*, des amorces de morphèmes et enfin des autocorrections. Ces résultats sont étroitement liés à la définition que nous avons donnée de chacun de ces phénomènes. Ainsi, en choisissant de nous centrer sur les répétitions formelles, nous évacuons des cas pourtant très proches de ceux que nous avons analysés ; en englobant dans la même analyse répétitions disfluentes et répétitions non disfluentes, on augmente le nombre des effectifs. De la même manière, le phénomène de l'autocorrection concerne plus d'occurrences que celles que nous avons choisi de prendre en considération dans ce travail. Nous l'avons dit, ce sont des contraintes méthodologiques qui nous ont poussée à procéder ainsi, puisque nous souhaitons faire un repérage automatique de ces disfluences, qui permettra ensuite leur analyse automatisée.

Nous avons analysé ces quatre disfluences séparément, et avons mis en évidence les rapports qu'elles entretiennent non seulement les unes avec les autres, mais également avec d'autres marques telles que les pauses silencieuses ou encore certains ponctuels. On a regretté l'absence de l'allongement vocalique dans les transcriptions, classiquement classé parmi les disfluences, qui a souvent rendu impossible la comparaison de nos résultats avec ceux obtenus par d'autres auteurs sur d'autres types de corpus.

Certaines de marques ont des liens très étroits entre elles : c'est le cas des répétitions et des autocorrections telles que nous les avons définies, mais aussi de l'amorce dont on a vu que près d'un tiers des occurrences entraînent dans une séquence répétition. Souvent en interaction, ces deux phénomènes se distinguaient par la catégorie grammaticale des formes concernées : mots grammaticaux pour les répétitions, mots lexicaux pour les amorces.

Nous avons essayé de relier la fréquence des disfluences à certains paramètres concernant les locuteurs ou la situation de parole, sans obtenir de résultats vraiment concluants de ce point de vue. Les divergences inter locuteur que nous avons mises en évidence pourraient être éclairées sur nos données en les analysant avec la prise en compte d'autres paramètres que nous avons évacués ici, telle que la prosodie.

Tout au long de ce chapitre, nous avons manifesté un intérêt particulier non seulement pour la catégorie grammaticale des formes concernées, mais aussi pour l'impact de ces disfluences en termes de construction syntaxique de l'énoncé. Ainsi, si les répétitions, les autocorrections et les *euh* ne

semblent pas influencer le déroulement syntagmatique de l'énoncé, il en va autrement des amorces qui se répartissent en plusieurs catégories. Les amorces inachevées marquent très clairement une rupture syntaxique dans l'énoncé. Nous avons tenté de voir si l'accumulation de marques pouvait être un indice de rupture syntaxique, mais on a du mal, au vu de nos données, à dégager des régularités.

Les résultats que nous obtenons dans cette section vont maintenant nous permettre de poser des choix quant à la manière de prendre en compte ces disfluences dans le cadre de l'étiquetage morphosyntaxique. En effet, on a vu que même lorsqu'elles ne constituent pas une rupture à proprement parler, les disfluences ont ceci de particulier qu'elles occasionnent un piétinement en un point de l'axe syntagmatique. Pour l'annotation morphosyntaxique avec notre système, cela pose un problème.

Pour prendre un exemple, dans la grammaire du français, un pronom *je* sujet est suivi d'un verbe conjugué à la première personne du singulier ; entre ce pronom et le verbe, on peut trouver d'autres pronoms préverbaux et la particule de négation *ne*. La liste de ces insertions est une liste d'éléments finis. Or, dans nos corpus, on peut trouver après un *je* :

un autre je

ilePA2 (...) don **je je ne pense** pas que c'est une question d'études (...) [ilePA2r]

une pause

iljPF1 **je / vois** pas les qualificatifs pour aller expliquer ça (silence) (...) [iljPF1r]

une amorce

accFJ1 (...) mais / disons que **je n/ n'ai** jamais eu beaucoup l'occasion de de voir des travaux (...) [accFJ1r]

un autre pronom sujet (qui n'appartient pas à la liste des pronoms préverbaux)

ilcDA1 oui / enfin **je j'étais** encore donc vendredi la semaine dernière à Paris (...) [ilcDA1r]

Ces disfluences perturbent donc la description grammaticale classique du pronom sujet *je*.

Dans notre travail, deux grandes options s'offrent à nous pour prendre en compte ces disfluences :

- a. les intégrer directement dans les grammaires de désambiguïsation que nous allons écrire (chap. 6) ;

-
- b. traiter les disfluences avant l'application des grammaires de levée d'ambiguïtés.

C'est cette seconde solution que nous avons choisie, en intégrant les disfluences ici étudiées dans une phase de prétraitement des données. C'est cette phase de prétraitement que nous allons maintenant détailler au chapitre 4.

CHAPITRE 4

LE PRÉTRAITEMENT DU CORPUS

pré-traité, -ée.adj. En adoptant ce riz du Texas (...) vous avez non seulement un riz long grain américain, mais un riz parfaitement « pré-traité ». (Elle, 4 nov. 1968, p. 187 ds QUEM. DDL t.24).

(Trésor de la Langue Française informatisé)

1 INTRODUCTION

Dans le cadre de l'analyse automatique de textes, il est rare de travailler sur des corpus bruts. Ceux-ci sont traditionnellement soumis à une phase de prétraitement ou de normalisation (Silberztein 1998a, Habert *et al.* 1998). Durant cette étape, le texte initial subit des modifications, qui facilitent les traitements ultérieurs. Selon les applications envisagées et l'état originel du texte, cette phase de prétraitement prend des formes parfois très diverses : il peut s'agir de réaccentuer des textes dont les signes diacritiques sont absents, d'harmoniser des graphies, de développer les abréviations ou les sigles, d'isoler des unités telles que les phrases, etc. Dans certains cas, comme par exemple pour la fouille de texte, l'étiquetage morphosyntaxique lui-même relève de la phase de prétraitement appliquée au texte.

Pour des textes oraux, Nelson (1996b) a mis sur pied un système de balisage en vue de leur entrée dans un analyseur syntaxique (parseur) :

Spoken English is characterized by a wide range of nonfluencies which are not found in writing. (...) These phenomena are transcribed as they occur, and the markup for them will be of a particular interest to researchers studying the interaction between speakers. However, they may be seen as disruption of the underlying syntax, and as such are problematic from the point of view of automatic parsing. We use the general term "normalization" to describe the method of using markup to deal with them. (Nelson 1996b : 41)

Cette normalisation peut aller très loin. En effet, les balises de Nelson prévoient, par exemple, la possibilité d'insérer des mots qui auraient été omis (généralement des mots grammaticaux), de compléter des amorces de mots ou encore d'effectuer des corrections appelées « normative insertions », qui consistent à rétablir une forme grammaticalement correcte là où le locuteur avait commis une « erreur ». La normalisation consiste aussi à enlever des mots, voire des segments qui gêneraient le parseur dans son travail. Ainsi, cet exemple donné par Oostdijk (2003 : 65) :

~~She was thinking~~ Lizzie was thinking of coming but

Dans cet exemple, le locuteur corrige en donnant un plus haut degré de précision concernant le sujet, pour passer du pronom personnel *she* au prénom *Lizzie*. Le locuteur abandonne pour ce faire la structure syntaxique en cours. La normalisation proposée par Oostdijk consiste alors à supprimer toute la séquence *she was thinking*, pourtant achevée après le groupe verbal. On voit comment, avec ces modifications, le texte « original » se perd.

De plus, cette normalisation doit être faite manuellement, et au cas par cas selon le corpus, sans possibilité d'automatiser la procédure. Notre démarche va dans un sens totalement opposé : nous cherchons des régularités, qui pourront être formalisées et permettre un traitement des données qui ne sera pas dépendant du corpus considéré, dans la mesure où celui-ci respecte les mêmes conventions de transcription que celles adoptées au centre de recherche VALIBEL.

2 UNITEX ET LE PRÉTRAITEMENT

Le logiciel que nous avons choisi d'utiliser dans le cadre de notre travail est UNITEX, développé à l'Université de Marne-la-Vallée par Sébastien Paumier (2003 et 2006). UNITEX est une plate-forme de traitement de corpus distribuée avec des ressources linguistiques, notamment des dictionnaires électroniques à large couverture (cf. chap. 5). Nous avons choisi UNITEX car il offre de nombreux avantages :

- Il est distribué sous licence LGPL¹, c'est-à-dire que l'utilisateur dispose de son code source qu'il peut modifier, adapter selon ses besoins et ensuite redistribuer gratuitement, avantage incontestable pour la diffusion du corpus étiqueté et sa consultation par d'autres chercheurs ;
- UNITEX offre à l'utilisateur la possibilité d'ajouter ses propres ressources lexicales (cf. chap. 5), ce qui est particulièrement

¹ Voir <http://www.gnu.org/licenses/licenses.html#LGPL>.

intéressant étant donné notre corpus de « français parlé en Belgique » ;

- Est intégré à UNITEX le système de levée d'ambiguïtés morphosyntaxiques ELAG (Elimination of Lexical Ambiguities by Grammars, Laporte et Monceaux 1999) que nous avons décidé de développer (cf. chap. 6).

Traditionnellement, les textes analysés à l'aide du logiciel subissent une étape de prétraitement que nous allons brièvement décrire², en détaillant les raisons qui font que nous ne suivons pas la même procédure pour notre corpus.

Le prétraitement classique avec UNITEX distingue deux grandes étapes, qui sont appliquées d'emblée au texte : la normalisation de certaines formes et le découpage en phrases.

2.1 Normalisation de formes non ambiguës

Cette étape consiste en la normalisation de certaines formes élidées et contractées. Il s'agit de rétablir les formes pleines de certaines formes élidées. Ces formes élidées sont non ambiguës et n'entrent pas dans la construction de mots composés : ces derniers ne seraient pas reconnus lors de l'application des dictionnaires, postérieure à la normalisation. Parmi les formes concernées, on peut citer *lorsqu'* qui est rétablit en *lorsque* ou *puisque'* en *puisque*.

Les formes contractées (morphèmes amalgamés) sont décontractées : une seule forme dans le texte (une suite de mots entre deux blancs) est donc remplacée par deux formes. On a ainsi *auquel*, qui devient *à lequel*, *desquelles* qui devient *de lesquelles*, etc.

2.1.1 Normalisation de formes ambiguës

Les formes élidées ou contractées qui ne sont pas normalisées lors de cette phase de prétraitement le sont après l'application des dictionnaires, lors de la construction de l'automate du texte (cf. chap. 6). En effet, la forme *l'*, par exemple, ne pouvait être rétablie sous sa forme pleine : selon les cas, il s'agit de l'élision de *le* (*Mon fils, je l'aime*) ou de *la* (*Ma fille, je l'aime*). Même chose pour *audit*, qui peut soit être décontracté en *à ledit*, soit garder sa forme originale *audit*, dans le cas où il est substantif (*faire un audit*). Le remplacement de ces formes contractées ou élidées se fait directement sur

² Le lecteur se reportera à Paumier (2006) pour une description détaillée du système.

l'automate du texte, qui permet de représenter les ambiguïtés dans des chemins parallèles (cf. chap. 6).

2.2 Conséquences de la normalisation graphique

Cette normalisation des formes, qu'elle soit effectuée lors de la phase de prétraitement ou plus tard, modifie les données originales. Ainsi, le texte suivant

J'ai raconté qu'on m'avait donné l'armoire du facteur.

devient après normalisation :

Je ai raconté que on me avait donné la armoire du
facteur.

Dans un texte écrit standard, celui qui analyse ces données peut aisément reconstruire le texte de départ. Or, pour ne prendre que le cas des élisions, il arrive en français parlé que des formes standard élidées ne soient pas effectivement réalisées là où elles seraient classiquement attendues, pour des raisons diverses, qui tiennent soit de l'hésitation, soit d'une insistance particulière marquée par une accentuation de la syllabe finale. Les locuteurs utilisent alors la forme pleine, comme dans³ :

ilrTC1 dans la région de Tour enfin à Tour **ie** en
connais assez bien qui parlent le wallon [ilrTC1r]

accCP1 (...) quant à la mère de ma femme / il est
visible **que** elle travaille dans un milieu euh / euh où il y
a plein de néerlandophones / qui parlent éventuellement
quelques mots de français (...) [accCP1r]

L'élision n'étant pas standardisée mais soumise à variation dans les pratiques des locuteurs, cette information nous semble devoir être conservée dans le texte. De la même manière, même si l'on n'en trouve pas d'attestation dans notre corpus, garder la trace que le locuteur a dit *de le* au lieu de *du* est plus que souhaitable. C'est la raison pour laquelle nous n'appliquons pas cette phase de normalisation.

Ce choix a des implications sur les ressources lexicales que nous utilisons. En effet, les formes élidées et contractées doivent figurer en tant que telles comme entrées dans le dictionnaire (cf. chap. 5).

Cet ajout d'entrées doit se faire non seulement pour les formes simples (*qu'*, *l'*, *lorsqu'*, etc.), mais également pour les formes composées (*parce qu'*, *dès qu'*, etc.). De plus, pour les formes composées dans lesquelles on

³ Si les exemples de *je* non élidé sont rares, ils sont beaucoup plus fréquents pour le *que*.

trouve *de*, les entrées doivent être démultipliées en tenant compte des formes élidées et contractées. On a ainsi non plus une seule entrée pour la locution prépositionnelle à *cause de*, mais quatre entrées :

à cause de,.PREP
à cause d',à cause de.PREP
à cause des,à cause de.PREP
à cause du,à cause de.PREP

2.3 Le découpage en phrases

Pour les besoins des traitements ultérieurs appliqués au texte, celui-ci est découpé en unités élémentaires, en « phrases ». La définition de la phrase se base ici sur des critères purement formels et typographiques : « groupe de mots qui commence par une majuscule et qui se termine par un point ». Cette définition, pour aussi « circulaire et à ras de terre » (Catach 1994 : 4) qu'elle soit, n'en pose pas moins d'énormes problèmes de reconnaissance automatique liés à une double ambiguïté :

- l'ambiguïté du point. En effet, celui-ci peut être point final, point abrégatif (c'est le cas dans les sigles ou encore à la fin d'abréviations courantes telles que *etc.*, *cf.*, *ex.*, *etc.*), mais il lui arrive de cumuler la double fonction abrégative et de fin de phrase. « On admet généralement qu'un point abrégatif "absorbe" le point final. » (Drillon 1991 : 138) ;
- l'ambiguïté de la majuscule. Celle-ci est indicative d'un début de phrase mais se retrouve également, indépendamment de la place occupée par le mot, dans les noms de personne ou de lieu, les sigles, les noms d'ouvrages, etc.

Les cas problématiques sont bien gérés par le graphe de reconnaissance des fins de phrase fourni avec UNITEX, qui scinde correctement des textes standard (journalistiques et romanesques) dans 99,99 % des cas, en insérant le symbole {*S*} (pour *sentence*) après chaque phrase reconnue⁴.

Quand nous donnons ce chiffre de 99,99 %, c'est sur base de la définition de la phrase que nous avons adoptée. Mais comme le fait remarquer Benoît Habert à propos de la segmentation en phrases lors de la campagne d'évaluation GRACE des étiqueteurs morphosyntaxiques (cf. chap. 6, § 7) :

⁴ Pour plus de détails, voir Silberztein (1990) et le manuel d'utilisation d'UNITEX (Paumier 2006). Voir également Dister (1998) et Friburger *et al.* (2000) pour des améliorations au découpage de base ainsi que des résultats sur corpus.

(...) un corpus de 450 000 mots selon une segmentation minimale (toutes les ponctuations et les espaces étaient considérées comme séparatrices) a été découpé en 16 897 phrases par un étiqueteur et 33 344 phrases par un autre... (Habert 2006 : 14)

C'est sur cette notion de phrase que nous voudrions maintenant nous pencher plus longuement.

3 LA NOTION DE PHRASE

3.1 La phrase intuitive

La définition de la phrase donnée ci-dessus est strictement typographique, liée à l'écrit. La notion de phrase semble faire partie de l'expérience commune. En effet, chacun a l'intuition de ce qu'est une phrase dans sa langue maternelle. Cette notion, tout comme celle de mot que nous aborderons par la suite, semble « marquée du sceau de l'évidence » (Béguelin 2000 : 49). Elle fait partie du langage commun, et nous la retrouvons dans des expressions telles que « finis ta phrase », « c'est la phrase qui tue », « s'arrêter au milieu d'une phrase », « parler avec des phrases toutes faites », etc. Les instituteurs ne demandent-ils pas régulièrement aux enfants, même très jeunes, de répondre à leur question par une *phrase complète* ?

La phrase est investie culturellement : qu'on pense à la prétendue clarté de la langue française, à l'harmonie et à l'élégance des phrases⁵, qualité indispensable chez les grands écrivains... Il y a la longueur des phrases chez Marcel Proust, chez Claude Simon ou encore chez Céline.

Notre expérience de la phrase est donc acquise dès l'école, où chacun la « reçoit comme une norme officielle de représentation du langage » (Berrendonner 1993 : 21).

Voici la définition de la phrase donnée à des enfants de première année de l'école primaire (6 ans)⁶ :

⁵ On trouve sur internet :

Andreï Makine est l'écrivain de la pudeur par excellence. Sa phrase, élégante sans être ampoulée, retient l'essentiel.

C'est aussi un écrivain, Maraud, à la phrase élégante et aussi joliment ronde qu'on pu l'être certains de ses inspirateurs.

⁶ À l'école communale de Fontin (Esneux).

La phrase est une série de mots qui raconte une petite histoire et qui ne doit pas faire mal aux oreilles. Elle commence toujours par une majuscule et finit toujours pas un point (ou un point d'interrogation ou un point d'exclamation).

Selon cette définition, l'institutrice explique que

Le chat mange la souris.

est une phrase. Par contre,

La souris mange le chat.

n'est pas considérée comme une phrase : une phrase doit avoir du sens, et une souris qui mange un chat, « ça n'a aucun sens, ça fait rire les enfants ». De la même manière, la suppression des déterminants dans

chat mange souris

fait de la séquence une non-phrase, parce que « ça fait mal aux oreilles ».

Il est intéressant de noter ici que, dans l'apprentissage de la notion, les élèves ne sont que secondairement confrontés à la version écrite de la phrase, avec la définition classique de la majuscule et du point. Ce critère semble pourtant essentiel, comme nous allons le voir ci-dessous.

3.2 La phrase intuitive : les représentations

Dans un travail sur les représentations, Marie-Noëlle Gary-Prieur (1985) a proposé à des étudiants 28 séquences, sous forme écrite, afin de cerner la définition intuitive que ceux-ci avaient de la phrase. Elle a analysé les commentaires des étudiants, dont elle a pu faire émerger quatre critères principaux sur lesquels ceux-ci se sont basés pour reconnaître aux différentes séquences le statut de phrase ou de non-phrase.

Tous les étudiants n'accordent pas la même valeur aux mêmes critères.

1^{er} critère : la ponctuation

Celui-ci apparaît comme le critère essentiel auprès des étudiants. En effet, l'absence de point final, et seulement celle-ci, empêche de reconnaître aux séquences suivantes le statut de phrase :

Ma nièce court après sa balle⁷

⁷ Nous reprenons en illustration les exemples proposés aux étudiants par Gary-Prieur (1985 : 33-35).



Il est tombé et il s'est fait mal

Par ailleurs, la présence d'une ponctuation, en l'occurrence un point d'exclamation, permet de distinguer les deux séquences suivantes, composées d'un seul mot, et d'attribuer à la seconde le statut de phrase refusé à la première

Pierre

Pierre !

Il est intéressant de constater que le statut même des exemples proposés induit un jugement différent toujours selon ce critère de la ponctuation. En effet, dans un poème non ponctué de Jacques Prévert présenté explicitement comme tel, les étudiants « s'attachent à trouver des phrases », parce qu'ils « reconnaissent au poète le droit de ne pas ponctuer » (Gary-Prieur, 1985 : 35), alors même qu'ils avaient immédiatement exclu les exemples non ponctués présentés ci-dessus.

2^e critère : l'intonation

On peut s'étonner de voir figurer le critère de l'intonation en deuxième position dans la mesure où l'exercice porte sur des productions écrites. Gary-Prieur ne s'attarde d'ailleurs pas dessus, reconnaissant qu'« on ne peut guère tirer de conclusions directes sur ce plan » (p. 35).

La prise en compte de ce critère paraît être là davantage pour faire bonne mesure et ne pas restreindre l'étude au plan strictement écrit.

3^e critère : le verbe

La présence d'un verbe semble être, chez de nombreux étudiants, un critère définitoire de la phrase.

Ainsi, beaucoup hésitent à qualifier de phrase les séquences suivantes :

(a) Excellent, ce café.

(b) Eh bien, et l'abolition de la peine de mort en matière politique ? (Flaubert)

(c) La liberté ? Un beau mot ?

Gary-Prieur note toutefois une tendance à reconnaître plus facilement *a* que *b* et *c* comme une phrase, dans la mesure où on peut reconstruire une séquence avec le verbe être (*Ce café est excellent*), ce qui est nettement moins évident pour les deux autres exemples.

Néanmoins, il semblerait que la conjonction des deux critères de ponctuation et de présence d'un verbe ne soit pas toujours suffisante. Ainsi,

Pleurer.

n'a pas été reconnu par les étudiants comme une phrase.

De même, l'extrait suivant d'un roman de Philippe Sollers, est ressenti par les étudiants interrogés comme incomplet :

...toujours déportés, décalés, rejetés, troués, ponctués, brisés, enchainés, toujours à nouveau rattrapés et courbés dans l'usure, mais aussi relancés, dénoués, dissous dans le fond à l'entrée du fond... Formant ainsi des plans parallèles tombant ou montant de l'un à l'autre, passant de l'endroit à l'envers, et coulant, nous recomposant, nous croisant, nous reconnaissant, nous désignant, nous oubliant, nous renversant, nous trahissant, nous reconstruisant... (Ph. Sollers, *Nombres*, Le Seuil, 1968, p. 101).

Là où les étudiants semblaient indulgents avec le poète, décelant des phrases dans une séquence non ponctuée de plusieurs lignes, ils le sont visiblement moins pour le romancier.

4^e critère : le sens

Certains étudiants, pour lesquels le critère sémantique est primordial, ont reconnu à toutes les séquences proposées le statut de phrase. En effet, moyennant une recontextualisation adéquate, il leur était toujours possible de trouver un sens, même à une séquence telle que

pierre

que la seule prise en compte des critères 1 et 3 excluait de la catégorie « phrase ».

Certains étudiants, se basant sur le critère du sens, ont proposé d'établir une distinction entre deux types de phrases. D'une part, ils distinguent les phrases dont le sens apparaît de façon autonome, indépendante et, d'autre part, les phrases pour lesquelles le sens ne peut émerger qu'en les insérant dans un dialogue, en inventant un contexte (linguistique ou situationnel). Dans le premier cas, le sens serait « intérieur » à la séquence, alors qu'il serait « extérieur » dans le second cas.

Dans l'expérience menée par Gary-Prieur s'est également posée la question de savoir si l'on avait affaire à une ou à plusieurs phrases. La chercheuse note que les réponses des étudiants dépendaient en général du critère qu'ils considéraient comme le plus important, parmi ceux repris ci-dessus, dans leur définition de la phrase.

Elle note également que parmi les 28 séquences proposées, une seule a été reconnue d'emblée et par tous comme une phrase :

La nuit est tombée depuis longtemps.

L'exercice demandé aux étudiants consiste à tenter d'objectiver cette notion intuitive, qui « appartient au métalangage collectif, validé par le sens commun » (Berrendonner 1993).

3.3 La phrase linguistique

Objet incontournable, la phrase fait partie des unités minimales de la langue au même titre que le phonème ou le morphème. Traditionnellement utilisée en linguistique contemporaine⁸, elle apparaît pourtant à bien des égards comme insaisissable⁹ tant les « tentatives de définitions »¹⁰ foisonnent¹¹.

Dans les critères définitoires de la phrase, avancés par les linguistes, se retrouvent en général deux principes essentiels (Berrendonner 1993, Blanche-Benveniste 1993, Simon 2000) : la complétude et l'autonomie. La complétude concerne en général le sens : une phrase serait un énoncé sémantiquement complet, achevé. Cette idée est déjà présente dans la notion d'*oratio perfecta* (Stefanini 1981). L'autonomie considère quant à elle la syntaxe (la définition du *Trésor de la Langue Française* parle de « cohésion grammaticale ») et, si l'on aborde le point de vue de l'oral, la prosodie. S'ajoute encore le critère de la ponctuation (Arrivé *et al.* 1986), qui permet d'isoler les phrases, et nous avons là réunis les quatre critères dégagés par les étudiants de Gary-Prieur. Les définitions, à base de critères syntaxiques ou sémantiques, considèrent la phrase comme un tout.

À ces critères se superpose encore la catégorisation en types de phrases : la phrase simple ou la phrase complexe (Leeman 2002), selon le modèle de construction ; la phrase verbale ou la phrase averbale (Blanche-Benveniste 1991, Lefeuvre 1999), selon les éléments que la phrase contient ; la phrase énonciative, injonctive, interrogative ou exclamative, distinctions établies en fonction de la nature de la communication et des intentions du locuteur (*Bon Usage*).

⁸ Pour la notion de phrase et son histoire, voir notamment Stefanini (1981).

⁹ « La phrase insaisissable » est le titre d'un chapitre consacré à la phrase dans Béguelin (2000).

¹⁰ En référence au chapitre sur la phrase intitulé « Tentatives de définitions » dans Yaguello (2003).

¹¹ « Dans les années 1950, Fries avait compté 200 définitions de la phrase et il s'en est évidemment rajouté depuis. » (Blanche-Benveniste 1993 : 8)

Dans le domaine de la syntaxe, la phrase constitue une limite supérieure. À l'inverse des essais de définitions de la notion, en grammaire générative, la phrase apparaît comme un postulat et, de ce fait, n'a donc pas à être définie :

La notion de phrase est, en grammaire générative, tenue pour un terme primitif, non défini, de la théorie. (...) nous pouvons (...) nous contenter de nous la représenter comme une suite d'éléments syntaxiques minimaux enchaînés (...) ; ces éléments syntaxiques minimaux sont, approximativement, équivalents aux *morphèmes* des structuralistes. (Ruwet 1967)

La première abstraction qu'il est indispensable d'introduire est la notion de phrase. Ce concept, qui sera noté *P*, n'a pas être défini, il s'identifie au postulat même d'existence de la syntaxe. (Gross 1997b : 5)¹²

3.4 La phrase à l'écrit, la phrase à l'oral

La phrase semble bien être essentiellement une notion liée à l'écrit, raison pour laquelle elle entretient des liens si étroits avec les marques typographiques de la ponctuation.

Selon Berrendonner (1993 : 20),

La seule définition empirique adéquate qu'on ait jamais pu lui trouver repose sur les critères purement orthographiques de la majuscule et du point.

Mais même à l'écrit, la notion n'est évidemment pas sans poser problème, et les exemples de sous-ponctuation et de surponctuation l'illustrent à merveille. Pour preuve, cet extrait de *La Route des Flandres*, de Claude Simon :

Il tenait une lettre à la main, il leva les yeux me regarda puis de nouveau la lettre puis de nouveau moi, derrière lui je pouvais voir aller et venir passer les taches rouges acajou ocre des chevaux qu'on menait à l'abreuvoir, la

¹² On trouve encore, dans le syllabus du cours de Maurice Gross (1997b) :

La description de la notion de phrase s'identifie donc au programme général de recherche de la syntaxe (Gross 1997 : 6).

Le fondement empirique des études syntaxiques réside donc dans la procédure intuitive de reconnaissance des séquences de mots que l'on accepte comme phrase (Gross 1997 : 9).

boue était si profonde qu'on enfonçait dedans jusqu'aux chevilles mais je me rappelle que pendant la nuit il avait brusquement gelé et Wack entra dans la chambre en portant le café disant Les chiens ont mangé la boue, je n'avais jamais entendu l'expression, il me semblait voir les chiens, des sortes de créatures infernales mythiques leurs gueules bordées de rose leurs dents froides et blanches de loups mâchant la boue noire dans les ténèbres de la nuit, peut-être un souvenir, les chiens dévorant nettoyant faisant place nette : maintenant elle était grise et nous nous tordions les pieds en courant, en retard comme toujours pour l'appel du matin, manquant de nous fouler les chevilles dans les profondes empreintes laissées par les sabots et devenues aussi dures que de la pierre, et au bout d'un moment il dit Votre mère m'a écrit. (Claude Simon, *La route des Flandres*, p. 9)

Selon les indices de la majuscule et du point, on a ici une seule phrase¹³. À l'inverse, dans cet extrait d'article de presse, le critère de la ponctuation scinde ce qui pourrait être une seule phrase (*Une façade blanche, discrète et anodine*) en trois phrases distinctes :

Une façade blanche. Discrète. Anodine. (*Le Soir* 2001)

Le style journalistique fourmille de structures de ce type.

Ainsi, non seulement la notion de phrase est liée fortement à l'écrit mais, tout comme celle du mot (Goody 1977), elle serait un produit de l'écrit (Blanche-Benveniste 1993).

L'étude de la langue orale spontanée a mis en avant très tôt la difficulté de segmenter le discours oral en phrases, et cela d'autant plus qu'il n'y a pas de consensus sur une définition de la phrase (autre que graphique). Vendryès le notait dès 1921 :

Dans la langue parlée, la notion de phrase au sens grammatical s'efface. Si je dis : « L'homme que vous voyez là-bas assis sur la grève est celui que j'ai rencontré hier à la gare », je me sers des procédés de la langue écrite et je ne fais qu'une seule phrase. Mais en parlant, j'aurais dit : « Vous voyez bien cet homme, — là-bas, — il est assis sur la grève, — eh bien ! je l'ai rencontré hier, il était à la gare. » Combien y a-t-il de phrases ici ? C'est très difficile à dire : imaginez que je laisse un temps d'arrêt à l'endroit marqué par des tirets, les mots « là-bas » constituent à eux seuls une phrase,

¹³ Selon Védénina (1989 : 2), « l'absence de ponctuation s'adresse à un lecteur hautement cultivé ».

exactement comme si je répondais à une question : « Où est cet homme ? » — Là-bas. » Et la phrase même « il est assis sur la grève » devient facilement un groupe de deux phrases si je m'arrête entre les deux parties qui la composent : « il est assis », « [il est] sur la grève » (ou « [c'est] sur la grève [qu']il est assis »). La limite des phrases grammaticales est ici tellement fuyante, qu'il vaudrait mieux renoncer à en tenter le compte. Mais à certain égard, il n'y a qu'une seule phrase. (Vendryès 1968 : 165, 1921 pour la 1^{re} édition)

La notion de phrase, essentiellement graphique donc, a ainsi rapidement été abandonnée par les linguistes qui s'intéressent à l'oral :

Il nous paraît évident que le français parlé fait sauter les cadres de l'analyse grammaticale traditionnelle (...). Une des notions qui « saute », c'est celle de phrase ; impossible de découper dans le parlé quelque chose qui corresponde à la notion de phrase pour l'écrit. (Blanche-Benveniste et Jeanjean 1987 : 89)

Fixer le discours sous forme de traces visuelles implique une objectivation de ses structures, une analyse explicite et consciente de la chaîne parlée. La phrase, qui est un des instruments de cette objectivation, n'est donc pas une unité de linguiste, mais une unité de lettré (et, plus spécifiquement, de typographe, comme le montre l'histoire de la notion). (Berrendonner 1993 : 21)

C'est la raison pour laquelle les transcriptions réalisées au centre de recherche VALIBEL ne contiennent pas de ponctuation (cf. chap. 1).

Avec l'abandon de la phrase chez les linguistes travaillant sur des données orales, on a vu naître un intérêt pour des contextes plus larges, qui dépassent largement le cadre de la phrase canonique¹⁴.

La notion de macro-syntaxe, développée au GARS, élargit le cadre de l'analyse grammaticale, pour trouver des unités de remplacement à la notion de « phrase ». La macro-syntaxe

(...) traite des unités globales apparentées à ce qu'on entend par « phrase », ou supérieures au niveau de la phrase (Blanche-Benveniste *et al.* 1990 : 40).

¹⁴ Cf. le titre de l'ouvrage édité en 2002 par Charolles, Le Goffic et Morel : *Y a-t-il une syntaxe au-delà de la phrase ?*

Si l'on s'accorde pour dire que la phrase est une unité d'analyse inadéquate, il est néanmoins difficile de s'entendre sur une autre unité (cf. Berrendonner et Reichler-Béguelin 1989).

3.5 Le découpage des productions orales

Le découpage de notre corpus en unités telles que celles de phrases n'est pas envisageable, non seulement pour des raisons théoriques, mais aussi pratiques : les transcriptions ne sont pas ponctuées. Pourtant, dans notre travail, il est nécessaire de découper le texte en unités plus petites. En effet, pour des raisons purement informatiques, un texte long doit être découpé en unités plus petites, afin de permettre une construction aisée de l'automate du texte (cf. chap. 6).

De plus, insérer des marques qui découpent la chaîne de mots s'avère intéressant lors d'applications ultérieures sur le texte. Par exemple, une recherche de motifs syntaxiques ne se fera pas au-delà des frontières ainsi établies lors du prétraitement¹⁵. Dans le cadre de l'analyse morphosyntaxique qui nous occupe, on peut se baser sur ces indices de frontière pour construire des grammaires de levée d'ambiguïtés.

Dans nos transcriptions, le seul découpage effectué a priori est celui qui distribue les énoncés en tours de parole (cf. chap. 1), unité non théorisée mais dont le statut pragmatique est relativement clair. Ces tours de parole peuvent être plus ou moins longs : composés d'un seul mot (les énoncés réponses) à 1110 mots pour le tour de parole le plus long (cf. chap. 2). Nous avons des séquences très courtes, et d'autres beaucoup plus longues, dépassant de très loin la longueur moyenne des phrases dans des corpus écrits standard¹⁶.

Il est donc intéressant, avant l'application de nos grammaires de levée d'ambiguïtés, de procéder à un découpage interne aux tours de parole. Ce découpage constituera une balise au-delà de laquelle les grammaires de levée d'ambiguïtés ne s'appliqueront pas. Nous insérons une marque *{S}* qui correspond pour nous à un *STOP*, plutôt qu'au signifié *Sentence* classique d'UNITEX.

¹⁵ Ou alors, ce sera de manière explicite.

¹⁶ La longueur moyenne des phrases est calculée en divisant le nombre de mots par le nombre de phrases. À l'écrit, cette donnée caractérise la complexité des textes et participe des formules utilisées pour mesurer leur lisibilité.

Cela dit, la longueur moyenne des phrases [ci-après *L*] varie d'un genre de texte à l'autre (Östman 1998), et, à l'intérieur d'un même genre, d'un auteur à l'autre. Brunet (1983 : 124) a ainsi obtenu les résultats suivants : *À la recherche du temps perdu* (*L*=31) ; *Émile* (*L*= 28) ; Chateaubriand (*L*= 22) ; Giraudoux (*L*=21).

Nous allons maintenant nous pencher sur plusieurs marques présentes dans le texte, et qui pourraient être des candidates adéquates pour une segmentation de celui-ci. Il s'agit pour nous de dégager des unités de travail pertinentes pour la tâche de l'étiquetage morphosyntaxique, et non des unités théoriques de l'oral, qui auraient des caractéristiques discursives, telles que s'appliquent à le faire les recherches en analyse du discours par exemple (Degand et Simon 2005 pour l'unité minimale du discours). De la même manière, on sera attentif au fait que nous ne cherchons pas davantage à dégager des unités qui correspondraient aux phrases de l'écrit, c'est-à-dire à des segments complets et autonomes, en termes sémantiques et syntaxiques.

Les marques que nous allons analyser sont les trois types de pauses silencieuses (pause brève, pause longue et silence), et le point d'interrogation.

L'idée est donc de voir à quel moment, dans le déroulement de l'énoncé, le locuteur produit une pause silencieuse. Celle-ci intervient-elle dans la chaîne parlée à un moment quelconque ou plutôt à une frontière de constituants ? Peut-on traiter les pauses des transcriptions comme des formes de ponctuation du texte, malgré ce que l'on sait de la non-correspondance entre ponctuation de l'écrit et pauses orales ?

4 LA PAUSE SILENCIEUSE

La pause silencieuse participe pleinement à l'organisation temporelle des énoncés. Elle a déjà été maintes fois étudiées (cf. Zellner 1998 ou Duez 1997 pour un état de l'art), principalement parce qu'elle constituait déjà un objet de recherche quand les études sur la parole ne se concentraient que sur ce que l'on nomme communément la « parole de laboratoire » (principalement des phrases lues, isolées, dans des études de phonétique acoustique).

Nous appelons ici *pause silencieuse* une interruption perçue du continuum du signal sonore dans une prise de parole. Dans la littérature sur le sujet, la pause silencieuse est également appelée *pause non sonore* ou encore *pause vide*, par opposition à la pause dite pleine, le *euh* (cf. chap. 3.5). Dans nos textes, les reprises de souffle audibles sont également transcrites comme des pauses silencieuses selon les conventions de transcription de VALIBEL.

La pause silencieuse peut se situer à l'intérieur du tour de parole d'un locuteur, ou entre deux tours de parole (« pause de passation de parole », Candéa 2000a). Dans ce dernier cas, elle n'est notée que s'il s'agit d'une pause particulièrement longue, soit un silence noté comme tel : (*silence*).

La pause silencieuse, reprise parfois parmi les phénomènes de disfluency quand elle est perçue comme une marque d'hésitation, occupe des fonctions très différentes, tant dans la planification de l'énoncé par le locuteur que

dans la compréhension de celui-ci par l'auditeur. La pause silencieuse permet au locuteur de reprendre son souffle ; elle lui laisse le temps nécessaire pour planifier la suite de son énoncé ; elle peut aussi avoir pour fonction de donner à l'interlocuteur le temps d'assimiler ce qui vient d'être dit ; elle ménage un suspens (suspens traduit généralement par des points de suspension dans le texte des écrivains) ou vise encore à mettre en évidence la suite de l'énoncé (elle est parfois appelée *pause de focalisation* ou encore *pause stylistique*) ; etc.

Si la pause constitue pour le locuteur l'occasion de reprendre son souffle et de respirer, il a été montré que ce moment de respiration n'est pas imposé au locuteur : celui-ci le « choisit » lors de sa production de parole (Autesserre *et al.* 1992). Ainsi, en général, on ne respire pas au milieu d'un mot, par exemple. Les contraintes physiques et physiologiques, si elles interviennent inévitablement¹⁷, ne sont pas prépondérantes. Comme le dit Duez (1997 : 276) :

La production de la parole est un processus dynamique qui est une fusion d'un ensemble d'actions physiologiques et cognitives. C'est un processus extrêmement complexe qui met en jeu toute une série d'organes et de muscles et qui recouvre différentes fonctions : respiration, recherche lexicale et planification, structuration grammaticale et mise en relief des idées.

Boomer et Dittman (1962) distinguent deux types de pauses silencieuses fonctionnellement différentes : la première est la *pause démarcative*, ou *pause grammaticale*, dont la présence participe de la structuration du discours. Cette pause démarcative ne se place donc pas n'importe où dans un énoncé ; elle participe pleinement à la structuration syntaxique et sémantique de celui-ci, ainsi qu'au regroupement des « mots » en segments cohérents. L'autre type de pause relève quant à lui de la disfluente dans la mesure où il indique une hésitation du locuteur et constitue une marque du « travail de formulation » de celui-ci (Morel et Danon-Boileau 1998). Contrairement aux pauses démarcatives, celles-ci se placent à des endroits non prédictibles, notamment à l'intérieur de syntagmes.

La **pause démarcative** peut être illustrée par l'exemple suivant :

chaBR1 c'était un vieux monsieur // qui élevait des
moutons // qui avait un grand jardin // qui était

¹⁷ Selon Grosjean et Deschamps (1975 : 157) qui comparent l'anglais et le français, « la respiration est un besoin physiologique qui se fait sentir à intervalle régulier indépendamment de la langue envisagée ». Ils constatent également que « plus les suites sonores sont longues, plus le pourcentage de pauses silencieuses contenant une pause de respiration est grand » (idem).

pensionné mineur // qui travaillait encore avec sa
pension // et que j'étais souvent avec lui // et un jour il
était allé voir un film de cinéma (bruit) (...) [chaBR1r]

Comme on le voit dans cet extrait, les pauses isolent des groupes syntaxiques qui correspondent à des propositions : une proposition principale, séparée d'une relative par une pause longue, cette proposition relative étant elle-même séparée par une autre pause longue d'une autre proposition relative juxtaposée, etc. Dans cette transcription, on constate un fonctionnement très régulier de la pause longue, qui structure véritablement l'énoncé en regroupant les mots en groupes cohérents, d'un point de vue syntaxique et sémantique.

La **pause non démarcative** peut être illustrée par l'exemple suivant :

accCTO et est-ce que justement la profession des des
parents ça peut influencer sur la / façon dont les enfants
vont parler avec plus ou moins d'accent [accBF1r]

Ici, la pause intervient à l'intérieur d'un syntagme, entre le déterminant défini *la* et le nom qui suit. Elle ne permet pas d'isoler de quelconques constituants. On remarquera par ailleurs que le tour de parole n'a pas ici d'autre pause que cette pause qui peut être considérée comme une marque de disfluence. Ce type de pause silencieuse non démarcative est fréquemment appelé dans la littérature sur le sujet *pause (silencieuse) d'hésitation* (Goldman-Eisler 1968 :10). En effet, tout comme le *euh*, ces marques considérées comme disfluentes ont été assimilées à des phénomènes d'hésitation de la part du locuteur.

Ces pauses ont fait l'objet de nombreuses études (principalement pour l'anglais, cf. Duez 1997 pour un état de l'art), dont le but est souvent de dégager des critères qui les distingueraient des pauses démarcatives (syntaxiques) : ces critères concernent la distribution des différents types de pauses, souvent en lien avec leur durée et leur fréquence dans l'énoncé, voire encore avec le type de discours.

Les études sur la pause silencieuse et ses fonctions ne manquent donc pas, auxquelles s'ajoutent encore, par exemple, celles qui étudient le lien entre syntaxe et intonation (cf. Rossi 1981). Mais notre objectif, s'il consiste également à étudier les pauses selon qu'elles sont ou non structurantes du point de vue grammatical, se situe dans une perspective relativement différente des études antérieures que nous avons consultées, tant sur le français que sur l'anglais. En effet, s'il s'agit bien de faire émerger une éventuelle différence entre pause démarcative et pause non démarcative, nous le faisons non pas sur base de données objectives – nous entendons par là établies à l'aide d'appareils acoustiques fins qui mesureraient les variables temporelles du français parlé spontané, noteraient des indices prosodiques, etc. – mais bien à partir de données secondaires.

Celles-ci sont constituées par un texte transcrit, dans lequel les pauses sont notées subjectivement par des transcripteurs humains qui n'ont recours qu'à leur oreille¹⁸.

De plus, la plupart des recherches citées se placent dans la perspective de cerner les marques d'hésitation (la pause étant une marque parmi d'autres), et ce point de vue cognitif n'est pas abordé dans notre travail.

4.1 La perception des pauses

La pause ayant une réalité physique (un blanc dans un continuum de signal sonore), on pourrait s'attendre à ce que cette réalité soit objectivable, et donc facilement décelable, et cela de manière identique, par les récepteurs, qu'ils soient auditeurs ou transcripteurs. Or, on constate qu'il n'en est rien et les tests de perception effectués par certains chercheurs concordent sur ce point.

La quasi-totalité des études sur le rôle de la subjectivité de l'auditeur dans la perception de la pause ont été faites sur l'anglais, mais les résultats de deux études sur le français que nous avons consultées vont dans le même sens¹⁹.

Ainsi, Candéa (2000a : 123), dans un test de perception qu'elle effectue dans le cadre de sa thèse, constate qu'aucune des pauses silencieuses de son corpus n'a été notée par la totalité des 30 auditeurs auxquels elle a soumis ses données sonores.

Duez (1991) constate quant à elle que les pauses situées entre des constituants sont mieux perçues que les pauses intraconstituant²⁰, la longueur de la pause ayant une répercussion directe sur son identification : plus une pause silencieuse est longue, mieux elle est perçue par les sujets. En ce qui concerne les pauses subjectives (c'est-à-dire les pauses perçues par les

¹⁸ Notons que les transcriptions se font maintenant avec le logiciel d'aide à la transcription Praat (Boersma et Weenink 2006) dans lequel figure le spectrogramme, c'est-à-dire une représentation graphique de la chaîne sonore qui symbolise la fréquence, l'intensité (degré de noirceur) et la durée. Dans celle-ci, on perçoit visuellement les interruptions dans le continuum sonore. On peut faire l'hypothèse que le transcripteur qui se base sur ces indices indépendants de la forme phonique (et donc des mots et des découpages syntaxiques basés sur la compréhension de ceux-ci) marquerait des pauses là où elles sont physiquement présentes.

¹⁹ Voir Candéa (2000a : 112 et sv.) pour un bref résumé d'autres tests de perception sur les pauses et les phénomènes dits d'hésitation.

²⁰ Notons par ailleurs que Duez, qui s'intéresse à la parole des hommes politiques, recense un certain nombre de pauses intraconstituant dans son corpus, pauses qui correspondent à des pauses stylistiques typiques de ce type de discours.

sujets-auditeurs qui ne correspondent pas à une interruption réelle du signal sonore), l'auteure note que celles-ci sont plus fréquemment identifiées à des frontières de syntagmes qu'à des frontières de propositions ou de phrases. La présence d'un *eah* ou d'un allongement vocalique induit souvent la présence d'une pause subjective :

Dans la plupart des cas, les pauses subjectives correspondent à un allongement final, qui correspond à son tour à la réalisation d'une frontière syntaxique (Duez 1991 : 140).

Les résultats du test de Candéa (2000a) vont dans le même sens : les frontières grammaticales favorisent la perception de pauses subjectives. De même, les pauses subjectives sont davantage perçues par les auditeurs à l'endroit où sont présentes des marques du travail de formulation.

Duez (1991), dans les tests de perception de phénomènes d'hésitation qu'elle met en place, évite de travailler sur des données transcrites, mais place ses sujets uniquement face à la source sonore. En effet, selon elle,

(...) la connaissance du texte écrit, et donc de sa structure syntaxique, peut aussi influencer sur la détection des pauses. Il est même possible qu'elle privilégie la détection des pauses grammaticales. (Duez 1991 : 105)

Duez (1997 : 285) critique à cet égard la procédure de Martin et Strange (1968) qui font marquer les hésitations directement sur le texte transcrit. Les auteurs obtiennent néanmoins des résultats qui corroborent ceux d'autres tests de perception. Ils constatent en effet que les auditeurs déplacent à des frontières syntaxiques des hésitations présentes à l'intérieur de syntagmes.

Candéa (2000a), dont le test de perception montre également que les pauses syntaxiques (réelles) sont mieux perçues que les pauses non syntaxiques, va dans le même sens que Duez, généralisant les propos de celle-ci au-delà du support écrit :

En fait, quelle que soit la procédure utilisée pour tester la perception des pauses, il semble que, de façon générale, tous les résultats issus de tests qui attirent explicitement l'attention des sujets sur le phénomène pausal concordent sur ce point : les frontières syntaxiques (limites de propositions grammaticales, limites de syntagmes, etc.) induisent la perception d'une pause subjective ou favorisent la perception d'une pause réelle située dans une telle configuration. (Candéa 200 : 113)

D'une certaine façon, leurs résultats vont dans le sens des remarques de Boomer et Dittman (1962), selon lesquels les pauses grammaticales (c'est-à-dire démarcatives) sont attendues par les auditeurs.

Comme nous l'avons dit, dans les transcriptions VALIBEL, l'interruption du signal sonore perçue comme une pause est jugée intuitivement par le transcripateur, et non mesurée à l'aide d'appareils acoustiques. Il s'agit donc dans notre corpus d'une marque subjective²¹, laissée à l'appréciation du transcripateur²². Celui-ci prend en compte plusieurs facteurs dont le principal est sans doute la vitesse d'articulation²³. La pause est donc jugée relativement à la vitesse d'articulation et, ainsi, pour une même durée de blanc dans le continuum sonore, selon que le locuteur parle plus ou moins vite, une pause d'une durée pourtant équivalente sera ou non marquée par le transcripateur selon le locuteur auquel elle est attachée.

Comme dans les transcriptions de VALIBEL la notation de la pause est subjective, il nous semble possible de formuler l'hypothèse, à la suite des constatations de Bommer et Dittman (1962), Duez (1991 et 1997) et Candéa (2000a), que les pauses présentes dans notre corpus ont été notées parce que justement perçues comme pertinentes par le transcripateur²⁴, au regard de la syntaxe ou de la morphosyntaxe. Nous faisons l'hypothèse que les pauses ont été transcrites à des endroits préférentiels en termes de structuration de l'énoncé, de regroupement de segments homogènes syntaxiquement et sémantiquement.

4.2 Le degré des pauses

Des études antérieures, notamment sur la lecture oralisée, ont montré la corrélation qu'il existe entre la durée de la pause silencieuse et la position syntaxique de celle-ci. Ainsi par exemple, Vannier et *al.* (1999) distinguent trois positions dans le texte lu – intérieur de phrase, entre deux phrases et

²¹ Dans le sens où elles sont laissées à l'appréciation du transcripateur ; nous ne prétendons pas que toutes les pauses sont subjectives au sens de Duez (1991) et de Candéa (2000a) (c'est-à-dire non présentes dans le signal sonore).

²² Pour rappel, toutes les transcriptions ont fait l'objet d'une validation par un second expert, et il est vrai que sur ce point, les jugements divergent régulièrement.

²³ Lane et Grosjean (1973) ont montré « qu'un sujet qui articule rapidement ne présente pas automatiquement un temps de pause réduit et vice versa » (cité par Grosjean et Deschamps 1975 : 156).

²⁴ On se garde néanmoins d'assimiler ici le transcripateur à un simple auditeur « naïf ». Mais les auditeurs dont il est question dans les recherches des auteurs cités ne sont pas non plus tout à fait naïfs dans la mesure où il leur était clairement demandé de se concentrer sur le phénomène pausal ou sur les marques d'hésitation.

entre deux paragraphes – et constatent des moyennes croissantes de durée de pauses pour ces trois positions.

À partir de données orales, Duez (1991) distingue quant à elle quatre positions – intraconstituant, interconstituant, fin de proposition et fin d'énoncé – et observe également une corrélation entre durée de la pause et frontière syntaxique : les pauses sont plus longues aux frontières syntaxiques majeures (propositions, phrases) qu'aux frontières de syntagmes. L'auteure observe également une variation importante de la durée et de la fréquence des pauses en fonction du type de parole, de la situation de communication. Ainsi, dans la parole des hommes politiques qu'elle étudie, elle constate une divergence entre d'une part les entrevues amicales et politiques et d'autre part les discours, ces derniers ayant une durée totale de pauses 50 % plus élevée que les premiers.

Grosjean et Deschamps (1972) obtiennent également des durées de pauses plus longues en fin de phrase qu'à l'intérieur de celles-ci, autant pour le français que pour l'anglais (1975).

4.3 La répartition des pauses dans le corpus

Les conventions VALIBEL distinguent trois degrés de pauses silencieuses :

- la pause brève (notée par une barre oblique : /)
- la pause longue (notée par deux barres obliques : //)
- la pause prolongée, qui correspond à un silence (notée entre parenthèses : *(silence)*).

Dans notre corpus, la répartition des trois degrés de pauses est la suivante :

marque	nombre d'occurrences n=	pourcentage sur l'ensemble des pauses
silence	2178	7,1 %
pause longue	7377	23,9 %
pause brève	21 316	69 %

Tableau 31. Répartition des trois degrés de pauses

On constate que parmi les trois types de pauses, les pauses brèves sont les plus nombreuses, avec 69 % de l'ensemble des pauses du corpus. Viennent ensuite les pauses longues et enfin les silences. Notons que les

pauses brèves et longues auraient pu obtenir un score encore plus élevé si nos conventions préconisaient de les indiquer aussi en fin de tour de parole et non pas uniquement à l'intérieur de la prise de parole d'un locuteur unique²⁵.

Les trois degrés de pauses confondus, cette indication d'une pause silencieuse se présente dans nos transcriptions en moyenne tous les 14,35 mots graphiques. Il s'agit donc d'une information qui apparaît fréquemment dans les textes, et nous voudrions maintenant voir concrètement de quelle manière elle peut servir d'indice à un éventuel découpage du texte en unités plus petites.

4.4 Le silence

Le silence, noté (*silence*), indique une pause perçue par le transcripteur comme particulièrement longue. Il se peut que certaines de ces pauses très longues appartiennent à ce que Candéa (2000 : 23) appelle le « domaine du comportement humain non-linguistique » :

(...) il ne s'agirait plus d'une *pause silencieuse* mais plutôt d'un *silence*, dans le sens commun de silence en tant que comportement ou attitude qui n'a plus rien en commun avec la pause démarcative, stylistique ou marquant une recherche de formulation dans la parole (Candéa 2000 : 23-24).

Ces pauses en fin de tour de parole, dans notre corpus, sont souvent liées au fait qu'après une question, le locuteur qui va prendre la parole a besoin d'un temps de réflexion (ou de réaction) avant de répondre. Les conventions préconisent de noter la marque du silence en fin de tour, et c'est ici que le terme *convention* prend tout son sens : il ne s'agit donc pas d'attribuer le silence au locuteur dont c'est le tour de parole (cf. chap. 1).

Notre corpus comprend 2178 marques de silence. Parmi celles-ci 760 arrivent effectivement en fin de tour de parole. On a également des *silences* suivis (96 cas) ou précédés (103 cas) d'une autre indication entre parenthèses (*bruit, cris, rire, toux, séquence inaudible*, etc.). Dans ces occurrences, nous verrons qu'une marque de fin de séquence est insérée (cf. § 6).

Restent alors 1219 marques de *silence*, que nous avons analysées au cas par cas. Si comme le prétendent Blanche-Benveniste *et al.* (1987, 1990), la pause se situe à des endroits non pertinents pour la syntaxe, nous pouvons faire l'hypothèse que le silence, qui est une pause particulièrement prolongée, a une distribution relativement régulière dans le corpus, et

²⁵ Pour rappel, seuls les silences sont notés en fin de tour de parole.

constitue une marque fonctionnelle. Nous faisons l'hypothèse que les marques de silence, telles qu'elles ont été notées intuitivement par les transpositeurs, n'apparaissent pas n'importe où dans l'énoncé et qu'il est pertinent dans le cadre de l'analyse morphosyntaxique de les considérer comme des marques d'arrêt dans la structuration de l'énoncé. On pose ici une hypothèse qui se base notamment sur les résultats obtenus par Grosjean et Deschamps (1972 : 149) qui, ayant étudié la distribution syntaxique des pauses silencieuses dans un corpus radiophonique, concluent :

Dans la langue orale spontanée, 75 % des pauses se trouvent à la fin d'un groupe de sens et, parmi celles-ci ; 60 % à la fin d'une phrase.

Les pauses en fin de phrase sont plus longues que les pauses en milieu de phrase.

Ainsi, dans de nombreux cas, le silence se produit à une frontière majeure, en ce sens où la construction syntaxique qui le précède est totalement achevée, comme dans les exemples suivants :

ilrLD3 (...) la euh notre langue c'est quand même le wallon // mais moi c'est ce que je pense (**silence**) je ne sais pas c'est mon avis quoi [ilrLD3r]

famHC1 ah tu es Proximus maintenant (**silence**) moi j'ai jamais eu de problème avec Proximus [famES1r]

Si l'on ponctuait ces énoncés, on aurait sans aucun doute inséré un point à cet endroit, ou tout au moins une ponctuation forte. Ces marques correspondent donc à la fonction démarcative dévolue traditionnellement aux pauses silencieuses, qui consiste à scinder un énoncé en groupes, où l'on pourrait retrouver la phrase classique, telle qu'elle apparaît dans des textes écrits standard.

De manière comparable, pour reprendre la terminologie classique de *proposition*, le silence se place :

entre deux propositions coordonnées :

ilrGS2 ben déjà à leur famille leur entourage (bruit de transcription) / puis il y a leurs connaissances à eux aussi aux gens qu'ils côtoient je ne sais pas à leur travail aussi (**silence**) et puis le fait qu'ils s'intéressent aussi euh / je sais pas à la littérature au cinéma /²⁶ et puis voilà (rire)

iljNJ1 euh (silence) je veux dire que c'est pas un français comme on disait justement à la parisienne

²⁶ On voit ici qu'avant la dernière proposition coordonnée, on a une pause brève.

(silence) mais je crois que je suis suffisamment clair que j'ai une élocution suffisamment correcte / pour que les gens en tout cas ne se posent même pas la question de dire tiens celui-là il a il a tel ou tel défaut donc euh[iljNJ1r]

ou devant une proposition subordonnée :

blaJD1 et / je suis tout près mais je vais retourner / à Saint-Idesbald // chez / chez Raymond et Denise // en taxi // mais alors je ne / me rappelle plus si j'ai dit le nom de la rue parce que je sais bien que |- (rire) <blaCD1> oui -| / un petit peu après / j'entends frapper à ma fenêtre de ma chambre **(silence)** parce que c'était surélevé (silence) ah non je |- dormais <blaJD1> (x) non-| / dans la chambre en bas |- moi

On a aussi des cas où le silence se place devant une proposition relative, séparant le pronom relatif de son antécédent :

iljGE1 (...) je crois que c'est un français de Paris **(silence)** qui gomme toutes les / diversités euh (...) [iljGE1r]

Le silence peut aussi séparer le verbe et le complément d'objet qu'il régit :

norPM1 donc / pour moi on peut améliorer la prononciation disons par / par l'exemple par l'enseignement // par la pratique / la pratique de la bonne ponctuation engendrera (silence) une un emploi // euh correct de la prononciation [norPM1r]

Le silence marque alors une rupture de construction de l'énoncé, en termes d'analyse syntaxique. En effet, classiquement, le verbe n'est pas séparé de son complément d'objet, et aucune marque typographique ne se place à l'écrit. Ce sont des exemples de ce type qui plaident en faveur de la non-correspondance entre pause silencieuse de l'oral et ponctuation de l'écrit. Les exemples suivants vont dans le même sens, puisque la marque du silence apparaît à des endroits non prévus classiquement par la syntaxe. Le silence se produit

entre le sujet et le verbe (6 occurrences)²⁷ :

norFA0 mm / oui mais si ces ces conditions sont réunies / euh l'école doit intervenir // selon vous

²⁷ Notons que les 4 premières de ces 6 occurrences se situent dans la même entrevue. On constate donc une forte régularité. La question est de savoir si elle doit être attribuée au locuteur ou au transcripteur.

norPM1 / le l'école (silence) doit intervenir parce
que c'est rendre service / à l'élève et au futur (...)
[norPM1r]

ou après un présentatif :

ilrTJ1 (...) j'aime bien ça // les slows évidemment ça /
puis (rire) pff ça puis il y a (silence) les animaux ça
[ilrTJ1r]

Dans les trois cas qui précèdent (silence entre la proposition relative et l'antécédent, entre le sujet et le verbe et après un présentatif), si l'on se place au niveau de la construction syntaxique en termes de dépendance, ces marques de la pause silencieuse constituent une rupture. Néanmoins, dans notre optique, on peut considérer cette interruption comme acceptable. En effet, dans ces cas comme dans la majorité des cas que nous rencontrons, le silence intervient entre deux groupes syntaxiques « correctement formés », pour lesquels l'analyse faite par nos grammaires ne sera pas problématique.

Ainsi, si le silence se produit entre un sujet et un verbe ou entre un verbe et un complément régi par celui-ci, le syntagme formé par les différents groupes est pratiquement toujours homogène.

Ces cas sont les plus nombreux (avec plus de 98 % des occurrences). En effet, très rares sont les cas où le silence brise un syntagme. On en rencontre quelques occurrences :

dans le groupe verbal : entre le verbe et son participe passé :

blaJD1 depuis (x) que | - oui (x) Saint-Idesbald
<blaCD1> (xxx) -| (silence) qu'il avait été à Furnes là
hein (silence) là il avait été (silence) bien reçu et
moi aussi sais-tu malgré tout que j'étais de la nuit et que
je ne pouvais pas rester ni rien (...) [blaJD1r]

dans le syntagme nominal, entre le nom et l'adjectif dont il dépend :

ilcBJ2 ben bien parler le français c'est respecter les
règles (silence) grammaticales [ilcBJ2r]

chaBR1 (...) j'ai travaillé vingt-six ans pour lui / c'était
les ateliers Boujat / un homme dictateur (silence)
méchant (silence) avec un caractère de fer (...)
[chaBR1r]²⁸

dans le syntagme nominal, entre le déterminant et le nom :

²⁸ Ici, le silence pourrait guider vers une transcription ponctuée qui mettrait des points à l'emplacement de chaque silence. On aurait alors un phénomène de surponctuation.

norPM1 il y a des (silence) marques de
prononciation à corriger (...) [norPM1r]²⁹

Dans notre corpus, on a aussi des cas où le silence apparaît à un endroit où l'énoncé est inachevé :

blaNB1 |- ouais <blaAD1> il a (x) -| hein et une
thrombose dis / obligés d'aller le / conduire en urgence
euh / enfin / le faire conduire en urgence il me dit / c'est
drôle hein mon bras ne me répond plus / euh moi je
(silence) non fini ça / fini fini fini (...) [blaNB1l]

Les marques de silence apparaissent également au lieu où se fait la recherche de la dénomination (Blanche-Benveniste 1985), avec les tournures caractéristiques d'une telle recherche :

iljVM1 à la limite simplificateur ça aussi ce serait bon
aussi / (xx) réducteur et redondant c'est c'est un peu la
même chose parce que (silence) on va dire / je sais
pas moi on va dire trois fois maison dans le même texte
ou / il est redondant (...) [iljVM1r]

ilrTC1 (...) par exemple je vais encore prendre Liège a
son accent Bruxelles a son accent / les Ardennes ont
leur accent euh / alors en France Paris a son accent euh
Marseille a son accent (silence) c'est comment / je
vais pas dire c'est une méthode de parler mais c'est /
une comment (silence) allez bon oh c'est une voix
entre guillemets [ilrTC1r]

Dans la très grande majorité des cas, on voit donc que la marque du silence n'interrompt pas la structure syntaxique de l'énoncé, dans la mesure où elle intervient après un syntagme achevé et avant un nouveau syntagme, au début de ce syntagme. Pour les rares occurrences où cette indication est concomitante avec une rupture syntaxique de l'énoncé (elle en marque le lieu, sans en être nécessairement la cause), noter cette rupture par une marque STOP est en cohérence avec les traitements que nous souhaitons appliquer au texte.

²⁹ Cette insertion d'une pause silencieuse entre le déterminant et le nom se produit à plusieurs reprises chez ce locuteur. Dans l'extrait suivant, c'est également après le *des* et avant le même syntagme nominal *marques de prononciation*, mais avec une pause brève cette fois. On trouve aussi dans le même tour de parole une autre pause silencieuse entre *des* et *caractéristiques propres* :

norPM1 et il y a des / marques de prononciation que
qu'on ne peut pas appeler des des défauts ni des fautes mais
qui sont des / caractéristiques |- propres // à un <norFA0>
(toux) -| milieu à un milieu ambiant (silence) (...) [norPM1r]

4.5 La pause longue et la pause brève

Les pauses longues, notées par une double barre oblique (//) sont trois fois plus nombreuses que les marques de silence. On en trouve 7377 occurrences dans le corpus. Les pauses brèves sont quant à elles encore plus nombreuses, avec 21 326 occurrences.

Étant donné ce nombre important d'occurrences, il était totalement impossible de tirer de quelconques conclusions qui se seraient basées sur une relecture, qui ne pouvait qu'être plus ou moins attentive, de la concordance. Nous avons donc choisi de nous baser sur l'étude de trois entrevues, afin de voir comment les pauses brèves et longues y étaient distribuées. Nous y avons également étudié les silences, déjà observés ci-dessus, afin de ne pas séparer les trois degrés de pauses et de permettre des comparaisons sur les mêmes données langagières.

4.6 Étude sur un sous-corpus

Le sous-corpus que nous avons choisi est composé de trois textes³⁰ : deux conversations informelles (blaAT1l et famVV1r) et une entrevue (accPH1r). Voici la répartition des pauses pour ces trois textes et pour le sous-corpus qu'ils constituent :

corpus	Nombre de mots	silence	pause longue	pause brève	longueur moyenne entre 2 pauses silencieuses, en nombre de mots
blaAT1l	8177	33	62	450	15
famVV1r	4220	19	77	194	14,55
accPH1r	8111	41	423	287	10,80
TOTAL	20 508	93	562	931	12,93

Tableau 32. Répartition des pauses dans le sous-corpus (valeur absolue)

On remarquera que accPH1r contient plus de pauses longues que de pauses brèves, inversement à la tendance générale du corpus global où les trois degrés de pauses connaissent un nombre décroissant d'occurrences, de la pause brève au silence, en passant par la pause longue.

³⁰ Il s'agit des mêmes textes qui ont déjà servi de base à l'analyse du *eu*h dans le chapitre 3.5.

Cette inversion de la tendance générale rencontrée dans accPH1r concerne en tout six entrevues dans notre corpus global.

La longueur moyenne entre deux pauses du corpus famVV1r (14,55) est très proche de celle de l'ensemble du corpus (14,35). Remarquons que dans ce calcul, nous globalisons à l'ensemble d'un texte, sans faire de distinction selon les locuteurs.

Le sous-corpus ainsi constitué comprend donc au total 1586 pauses dont la répartition est la suivante selon les trois degrés de pauses :

silence	pause longue	pause brève
5,86 %	35,44 %	58,70 %

Tableau 33. Répartition des pauses dans le sous-corpus (pourcentage)

L'objectif spécifique de cette étude du sous-corpus est de voir si la pause, selon son degré, est présente à des endroits pertinents pour le découpage du texte transcrit :

1. La pause est-elle située à l'intérieur d'un constituant (suspend-elle momentanément le déroulement syntagmatique de l'énoncé, qui se poursuit ensuite « normalement ») ou bien se place-t-elle préférentiellement entre deux constituants ? Pour le dire autrement : la pause est-elle démarcative ou non ?
2. La pause constitue-t-elle une rupture dans la construction syntaxique de l'énoncé ? En d'autres termes, intervient-elle à un moment où le déroulement syntagmatique de l'énoncé est interrompu et laisse inachevée la structure syntaxique entamée ?
3. La pause a-t-elle une fonction différente selon qu'elle est ou non combinée à une autre marque ou qu'elle apparait dans certaines positions de l'énoncé ?
4. La pause a-t-elle un comportement différent selon son degré (pause brève, pause longue et silence) ?

4.6.1 Études antérieures

Nous l'avons dit, les études sur les fonctions de la pause, démarcative vs non démarcative, ne manquent pas. Mais force est de constater que, selon les auteurs, les résultats des recherches divergent. Selon certains (Boomer 1965), la pause d'hésitation a une durée plus longue que la pause grammaticale. D'autres (cf. Duez 1997 : 280) ne voient pas de distinction significative entre la durée des pauses, selon qu'elles marquent une hésitation ou structurent le discours en constituants. Ainsi, pour Campione et

Véronis (2004), la pause seule n'est jamais une marque d'hésitation, sauf si elle est très brève (en dessous de 200 ms). Pour remplir cette fonction d'hésitation, elle doit obligatoirement être associée à une autre marque : soit un allongement vocalique, soit un *euh*, soit la combinaison de l'allongement et du *euh*. Les auteurs rejoignent ici Danon-Boileau et Morel (1998), qui estiment qu'en dessous de ce seuil de 200 ms, la pause est en général de nature respiratoire.

Grosjean et Deschamps (1972), dans leur étude sur les variables temporelles du français spontané, s'intéressent à la distribution syntaxique des pauses non sonores. Ils ont étudié un corpus de 30 interviews d'une durée minimale de trois minutes, interviews reprises d'émissions radiophoniques³¹, et ne prennent en considération que les pauses dont la durée est supérieure à 0,25 secondes. Pour leur analyse de la distribution des pauses, les auteurs découpent les énoncés en phrases (« ensemble formé de SN et de SV ») et en syntagmes et étudient la répartition des pauses en fonction de ce découpage. Sans faire de distinction entre les trois types de pauses qu'ils mentionnent (pause de respiration, pause stylistique et pause d'hésitation), ils constatent³² que 60,46 % des pauses non sonores se situent en fin de phrase,

c'est-à-dire au point de rupture grammatical de la chaîne parlée le plus normal. (...) la pause en fin de phrase est tout naturellement considérée comme le point de rupture naturel qui permet la reprise d'air et (...) la pause à l'intérieur de la phrase aurait plutôt tendance à être une pause d'hésitation ou une marque stylistique. (Grosjean et Deschamps 1972 : 146-147)

Les auteurs obtiennent des pauses qui se situent dans 23,36 % des cas à des « coupures grammaticales » (p. 146) (après ou avant un complément de phrase, entre SN et SV ou entre des mots coordonnés ou apposés). Restent enfin 16,18 % de l'ensemble des pauses qui se situent « à l'intérieur d'un groupe de sens (soit à l'intérieur de SN soit à l'intérieur de SV) » (p. 147). Le SV étant composé du verbe et de ses compléments, une pause entre un verbe et un complément d'objet est donc considérée comme marquant une rupture à l'intérieur d'un groupe de sens (nous verrons que nous ne considérons pas qu'il y a rupture dans ce cas). Les auteurs ne mentionnent

³¹ Dans leur étude de 1973 sur des descriptions de dessins humoristiques, ils obtiennent des résultats comparables de distribution syntaxique de la pause non sonore. Alors qu'ils qualifient la description de « langage hésitant » (Grosjean et Deschamps 1973 : 210) – car selon eux plus contraignant pour le locuteur que l'interview –, l'une des explications qu'ils avancent est que, dans le cas des descriptions, « les phrases (...) sont nettement moins complexes et plus courtes ». (Grosjean et Deschamps 1973 : 213)

³² Les auteurs présentent une répartition détaillée des différents cas au tableau IV de la page 148.

aucunement les occurrences où la pause pourrait marquer l'abandon de la construction, sauf à considérer que les faux départs recensent tous ces cas³³.

Grosjean et Deschamps précisent par ailleurs que 74,64 % de leurs pauses sont des pauses de respiration (regroupées ou non avec une pause d'hésitation ou une pause stylistique), sans faire de lien explicite entre le découpage syntaxique et le type de pause (de respiration ou de non-respiration).

Campione et Véronis (2004), qui ont analysé un corpus de parole spontanée de 54 minutes (correspondant à 8500 mots), relèvent seulement 71 % de pauses démarcatives (celles-ci étant repérées grâce à des indices convergents (intonation, syntaxe, etc.)). Ce chiffre est faible si l'on considère que la fonction démarcative est la principale fonction de la pause. Selon ces chiffres, on peut s'attendre à ce que les pauses notées dans notre corpus ne soient pas de bonnes bases à un découpage du texte dans plus d'un cas sur quatre.

Nous l'avons dit ci-dessus, les auteurs constatent que les pauses très brèves (en dessous de 200 ms) ne sont jamais démarcatives. Ils en ont très peu dans le corpus : 24 sur 1163 pauses. En fait, s'ils ont également trois degrés de pauses, on s'aperçoit que ceux-ci ne sont pas comparables aux nôtres, et que nous ne pouvons donc pas extrapoler leurs résultats en ce qui concerne la longueur des pauses.

En fait, à la lecture des études antérieures sur le sujet, nous sommes confrontée à un double problème :

- nous n'avons aucune mesure de durée des pauses, simplement des degrés jugés intuitivement par les transpositeurs ;
- les auteurs consultés ne distinguent pas clairement *hésitation* et *structuration syntaxique*. Ce qui nous intéresse ici n'est pas d'aborder la pause en tant qu'éventuelle marque d'hésitation, appartenant au travail de formulation du locuteur, mais bien de l'analyser en termes de structuration syntaxique de l'énoncé. Or, une pause peut très bien noter une hésitation tout en structurant l'énoncé et en n'introduisant pas de rupture syntaxique de celui-ci. Mais généralement, les auteurs consultés ne font pas le départ entre ces deux points de vue. En tout état de cause, la pause, qu'elle signifie ou non l'hésitation du locuteur, ne nous intéresse que dans le cadre strict de la structuration syntaxique des énoncés.

³³ Dans leur étude comparative français-anglais (1975), ils distinguent les faux départs repris des faux départs non repris.

4.6.2 Combinaison de la pause et d'autres marques

Parmi ce que nous considérons comme une autre marque ou une autre position pertinente dans l'énoncé, nous avons repris : les amorces de morphèmes, les autocorrections immédiates, les répétitions où la pause est interne³⁴, les onomatopées et les ponctuants, les *euh*, les indications paraverbales et les passages incompréhensibles, les positions de fin et de début de chevauchement et de fin de tour de parole, cette position « fin de tour de parole » étant uniquement possible pour les silences.

Parmi les 1586 pauses du sous-corpus, on constate que 888 de ces marques sont contigües à la pause. Nous envisageons ici seulement des cooccurrences immédiates, et non des séquences de ces marques³⁵. Voici les résultats des marques contigües, classées par ordre de fréquence décroissante et indépendamment du contexte gauche ou droit de la marque :

marque	n=
<i>euh</i> dit d'hésitation	170
ponctuants	146
fins de chevauchements	134
débuts de chevauchements	121
onomatopées	110
répétitions	83
fins de tour de parole	45
indications paraverbales	36
amorces	32
passages incompréhensibles	7
autocorrections	4

Tableau 34. Types de marques contigües à la pause

Les marques ou positions considérées ici appellent immédiatement quatre remarques :

1. Nous ne pouvons prendre parmi les marques l'allongement vocalique, dont les études antérieures ont montré la fréquence de cooccurrence avec la pause, et le

³⁴ Et non les répétitions qui précéderaient ou suivraient la pause.

³⁵ Cela signifie que nous prenons en considération les marques strictement contigües.

rôle dans la perception des pauses subjectives (cf. Duez 1997 : 288).

2. L'ordre de fréquence de ces marques est intéressant, puisqu'on constate un classement différent dans la combinaison avec la pause. En effet, le *euh*³⁶ est plus fréquent que les répétitions, ce qui n'est pas le cas de manière globale dans le corpus.
3. Certaines marques contraignent l'analyse que nous faisons de la pause en termes de position intra ou interconstituant, ainsi qu'en termes de rupture syntaxique de l'énoncé. Ainsi,
 - 3.1. la position « fin de tour de parole » est de toute façon considérée comme une fin d'unité (cf. § 6), la notation du silence étant du point de vue du découpage en unités redondante ;
 - 3.2. les amorces de mots, les indications paraverbaux et les passages incompréhensibles sont également considérés comme des endroits où l'analyse des grammaires doit s'arrêter (cf. § 5). La présence de la pause est donc ici également redondante par rapport aux choix de découpages déjà faits pour ces marques.
4. La pause silencieuse est plus fréquente en fin de chevauchement (134 occurrences) qu'en début de chevauchement (121 occurrences). Cela semble étonnant. En fait, une hypothèse qui peut être avancée pour les paroles superposées est que l'interlocuteur se saisit de la pause laissée par le locuteur pour tenter de prendre la parole, comme le suggère Oostdijk (2003 : 64) :

it (...) has been observed that in most communicative situations the time available for pausing is limited as there is the risk of losing the floor to another speaker or losing the audience's attention.

Si cette hypothèse s'avérait exacte, on devrait s'attendre à un plus grand nombre de pauses qui précèdent le

³⁶ Campione et Véronis (2004) notent également une forte interaction de la pause silencieuse et de la pause pleine : 33 % des pauses silencieuses sont en contact avec une pause pleine, et 58 % des pauses pleines cooccurrent avec une pause silencieuse. Ils comptent parmi la pause pleine l'allongement vocalique.

début du chevauchement. Or, on constate pourtant que ce nombre est inférieur au nombre de pauses qui suivent la fin du chevauchement.

4.6.3 Contexte des marques combinées

contexte par rapport à la pause	<i>eah</i>	ponctuant	onomatopée	amorce
avant la pause	146	93	63	21
après la pause	24	53	47	11

Tableau 35. Contexte des marques combinées à la pause

On le voit, pour les quatre marques envisagées, la position avant la pause est beaucoup plus fréquente qu'après la pause. Toutes marques confondues, celles-ci se placent dans 71 % des cas dans le contexte gauche de la pause, contre seulement 29 % des cas dans le contexte droit.

Selon Candéa (2000a), la pause silencieuse seule et la pause silencieuse subséquente à une marque du travail de formulation – chez cette auteure : *eah*, allongement vocalique, répétition ou auto-correction immédiate – ont des fonctionnements très différents et elle y voit deux types de pauses. Pour elle, la configuration *marque de travail de formulation + pause silencieuse* forme une unité (comme si la pause était seule), ce qui n'est pas le cas de la combinaison *pause silencieuse + marque de travail de formulation*. Candéa considère comme seul pertinent le contexte gauche de la pause pour classer celle-ci parmi les pauses non structurantes. Opposée à la pause non structurante, la pause structurante se définit formellement comme

- une pause silencieuse entre deux séquences de parole continue d'un même locuteur ou comme
- une pause silencieuse précédée d'une marque du travail de formulation (à gauche donc, et non à droite comme pour la pause non structurante).

La durée de la pause silencieuse structurante est plus brève que celle de la pause subséquente à une marque de travail de formulation. Dans ces deux configurations, la pause définie formellement comme structurante sépare des constituants ; l'éventuelle présence d'une marque du travail de formulation subséquente à la pause indique le début du constituant suivant auquel elle appartient.

Ainsi, le terme *structurant* chez Candéa n'a pas de correspondance stricte avec le rôle démarcatif tel que l'entend Bommer (1962) par exemple, et tel que nous l'entendons ici. Nous assimilons en fait *démarcatif* à

structurant du point de vue syntaxique (excluant de démarcatif la fonction stylistique des pauses intraconstituant par exemple). Candéa a donc des pauses qu'elle appelle structurantes en position intraconstituant. En fait, nous l'avons déjà dit, Candéa s'intéresse aux phénomènes dits d'hésitation ; quant à nous, c'est l'aspect purement syntaxique des choses que nous souhaitons prendre en compte, ce qui rend les résultats des études antérieures souvent inutilisables pour nous.

De plus, nous avons déjà eu l'occasion de le voir, la définition de ce qu'est l'unité (le constituant) syntaxique varie d'un auteur à l'autre, ce qui rend donc variable l'appréciation en termes de rupture syntaxique, du point de coupure grammatical (cf. principalement les divergences avec les études de Grosjean et Deschamps). Ainsi, Candéa (2000a :170) considère comme une pause intraconstituant la pause placée entre un nom et son expansion, même si cette expansion est de type complément du nom et pas seulement un adjectif épithète.

4.6.4 Pause marque de rupture et pause intraconstituant

Le tableau 36 reprend³⁷, pour les trois degrés de pauses, le nombre d'occurrences (colonne n=) pour lesquelles la construction est soit rompue (le syntagme est laissé inachevé, colonne « rupture »), soit interrompue momentanément à l'intérieur d'un constituant puis achevée (colonne « intraconstituant »). Le pourcentage est calculé sur l'ensemble des pauses de même degré dans le corpus.

³⁷ Nous ne considérons pas comme un point de rupture dans la linéarité syntaxique de l'énoncé les pauses contigües à une marque déjà reconnue comme stoppant la construction, c'est-à-dire certaines amorces de morphèmes, les indications paraverbaux et les passages incompréhensibles (cf. § 5). Ce cas de figure se présente à 75 reprises dans le corpus et les occurrences ne sont donc pas comptabilisées ici.

	rupture		intraconstituant	
	n=	%	n=	%
pause brève	37	3,97 %	100	10,74 %
pause longue	7	1,25 %	47	8,36 %
silence	3	3,23 %	2	2,15 %
total	47	2,96 %	149	9,39 %

Tableau 36. Répartition des pauses intraconstituant et des pauses à l'emplacement d'une rupture

Les pauses brèves sont les plus nombreuses à se placer non seulement à l'endroit d'une rupture syntaxique, mais également en position intraconstituant ; elles sont suivies de loin par les pauses longues et enfin par les silences, classement qui correspond à l'ordre de fréquence de ces pauses dans le corpus. Pour les trois degrés de pauses, les cas de rupture syntaxique sont moins fréquents que les occurrences de pauses placées à l'intérieur d'un constituant (au total, 47 cas de rupture et 149 cas de pauses intraconstituant).

Si l'on considère non pas le nombre d'occurrences mais la fréquence de chaque type de pause sur l'ensemble des pauses de même degré (la colonne pourcentage), on voit que l'analyse diverge sensiblement. En ce qui concerne la rupture syntaxique, les silences sont presque aussi nombreux que les pauses brèves (3,23 % contre 3,97 %), la pause longue ne se positionnant à l'emplacement d'une rupture syntaxique que dans 1,25 % de ses occurrences en corpus. Pour ce qui est de la pause intraconstituant, la fréquence reste la même, la pause brève se plaçant toujours au 1^{er} rang avec 10 % de ses occurrences qui arrivent à l'intérieur d'un même constituant.

4.6.5 La pause à l'emplacement d'une rupture syntaxique

En excluant les pauses cooccurrentes à une autre marque qui constitue un arrêt dans notre procédure de découpage du texte (cf. § 5), on recense 47 cas de rupture syntaxique à l'endroit d'une pause. Nous choisissons la formulation relativement neutre « à l'endroit d'une pause » car nous ne poserons pas d'hypothèse en ce qui concerne la raison de cette rupture syntaxique (la pause en marque-t-elle la place ou en est-elle en partie la cause, par exemple).

En ce qui concerne les indications *silence*, on voit que 3 occurrences marquent une rupture syntaxique. Elles appartiennent toutes au texte

blaAT11, et deux ruptures se produisent en fin de tour de parole, après une conjonction³⁸ :

blaAT1 ouais c'est c'est la folie quoi non mais c'est c'est c'est dingue d'être aussi euh / inconscient et puis / enfin tu vois j'ai l'impression que c'est je vivais que la nuit quoi tu vois // en même temps je crois que c'était pendant l'hiver donc à mon avis c'est cette image-là / qui qui m'est restée mais enfin non pas rien que l'hiver **mais (silence)**

blaAT1 en plus il y a un -| bébé mais non mais / je veux dire euh là c'est / je suis vraiment je suis vraiment bien avec lui quoi enfin c'est clair que **(silence)** oui monsieur oui madame (en parlant fort et en s'adressant à l'enregistreur)

La rupture syntaxique se produit à l'emplacement d'une pause longue à 7 reprises. Ici aussi, toutes les occurrences apparaissent dans un même texte, accPH1r, dont on a vu qu'il comptait plus de pauses longues que de pauses brèves. Devant de tels regroupements des occurrences, on peut formuler deux hypothèses quant aux différences de résultats observés, même devant ce nombre d'occurrences réduit :

1. on est devant une variation interlocuteur ;
2. la variation est moins le fait des locuteurs que de la perception du transcripteur.

En fait, il nous semble plausible que les deux facteurs interviennent de façon concomitante, et nous n'avons pas eu recours au signal sonore pour vérifier, l'objectif étant de travailler uniquement sur base des données transcrites.

En ce qui concerne la pause brève, on constate qu'elle est davantage répartie sur l'ensemble des trois textes (6 pour famVV1r, 9 pour accPH1r et 22 pour blaAT11).

Une typologie des cas de ruptures syntaxiques est très difficile à établir, et il semble bien, au vu de l'analyse de nos données, que celles-ci puissent revêtir des formes très variées, comme le montrent les exemples ci-dessous :

blaAT1 (...) ben si il arrive à à survivre à / à des cours du soir plus un travail à temps plein euh il arrivera **à / parce que** il y a beaucoup moins de cours euh du soir euh (silence) [blaAT1]

³⁸ Notons que 48 % des occurrences de la marque *(silence)* du sous-corpus se produisent en fin de tour de parole.

blaAT1 ouais -| on allait chez elles ouais c'était gai /
non c'était gai / c'était un peu nostalgique tu vois c'est
marrant de / il y avait Nadia Lamantia tu vois
[blaAT1]

accCT0 et vous savez donner un exemple (silence) |- /
ça n'est <accPH1> m a/ -| pas grave non je / donnez
un exemple [accPH1r]

famVI1 merci (rire) / je n'ai pas été à Pattaya / quand
je suis débar/ quand j'ai débarqué de / / de c'était où ?
/ du Laos / quand j'ai |- repris <famVV1> oui -| le bus
pour revenir à / à Bangkok (...) [famVV1r]

Parmi les quatre exemples repris ci-dessus, les deux premiers illustrent une interruption après une préposition. Dans l'exemple (1), une conjonction suit l'interruption, tandis que dans l'exemple (2) le locuteur redémarre avec un présentatif. Dans l'exemple (3), la coupure se produit après un pronom personnel sujet, avec une reprise par un impératif. Dans la définition que nous faisons du constituant et en fonction de nos procédures de levée d'ambiguïtés (cf. chap. 6), ces ruptures qui se produisent après une marque de début de groupe (une préposition, un pronom personnel sujet non ambigu) sont évidemment particulièrement gênantes (cf. chap. 6).

Le dernier exemple montre une pause entre les deux termes d'une répétition de la préposition *de*, celle-ci étant suivie d'une auto-correction (la contraction *du*) avec, entre les deux, l'insertion d'une marque explicite de recherche dans la formulation *c'était où ?* Ces marques sont difficiles à répertorier de manière exhaustive, tant elles peuvent prendre des formes variées. Nous en avons néanmoins repris certaines, principalement des ponctuels, dans la section consacrée aux répétitions (chap. 3.2).

En fait, dans la tâche d'étiquetage, ces inachèvements semblent bien constituer le problème majeur, et cela d'autant plus qu'il est très difficile de les repérer formellement. Nous avons déjà mis ce point en évidence dans le chapitre 3 consacré aux disfluences. En effet, on l'a vu ci-dessus, les cas de rupture ne sont pas préférentiellement marqués par la conjonction de la pause et d'une autre marque. Si c'est le cas pour l'exemple (4) avec la répétition, il ne s'agit pas d'une majorité de nos occurrences, puisque une minorité de cas (17 cas, soit 36,17 %) est concernée par une marque mixte. Les autres cas, presque deux fois plus nombreux, sont uniquement marqués par la présence de la pause seule. De plus, si la combinaison de la pause et de la répétition marque bien en ce cas précis une rupture de la construction de l'énoncé, nous avons vu qu'il s'agit en fait d'une exception à la tendance générale, laquelle montre qu'il n'y a pas de rupture dans cette configuration.

La répartition des marques mixtes, elle aussi très variée, est la suivante :

- avant la pause : 2 *eah* et 3 ponctuels ;

- après la pause : 3 ponctuants ;
- 3 pauses entre les termes d'une répétition ;
- des séquences qui combinent la pause et 2 autres marques : 3 *euh* avant et un ponctuant après la pause ; 2 répétitions et un ponctuant après la pause ; une suite qui combine autocorrection, pause, *euh* et ponctuant.

En fait, nous l'avons déjà dit, on ne peut considérer l'accumulation comme un indice certain de rupture.

Inversément, on recense également de nombreux cas d'inachèvement de la structure entamée qui sont non « marqués », même par une pause. Cela se produit à deux reprises dans l'extrait ci-dessous, chaque fois après le pronom sujet *je*. Il nous semble que l'on pointe là une caractéristique langagière de la locutrice :

iljDV1 euh / d'abord **je en** règle générale **je on** a je
trouve une assez mauvaise diction / enfin prononciation
(...) [iljDV1r]

On est donc bien là, nous semble-t-il, devant une pierre d'achoppement de notre système d'étiquetage (cf. chap. 6).

4.6.6 La pause intraconstituant

Le tableau ci-dessus montre que 9,39 % des pauses du corpus, indépendamment de leur longueur, sont placées à l'intérieur d'un constituant. Ce chiffre est relativement élevé³⁹, et dans l'optique d'étiquetage automatique qui est la nôtre, il constitue un vrai problème. En effet, si l'unité « constituant » telle que nous l'avons définie peut être coupée par une pause pratiquement une fois sur 10, cela rend le système nettement moins performant. Peut-être faut-il envisager de ne pas considérer la pause – du moins la pause brève et la pause longue qui ont respectivement 10,74 % et 8,36 % de leurs occurrences à l'intérieur d'un constituant – comme une marque qui permet d'isoler d'éventuels groupes syntaxiques homogènes. Le silence, lui, resterait alors le seul bon indice de découpage, puisqu'il se place préférentiellement à des frontières majeures (cf. ci-dessus, nos résultats qui corroboraient ceux de Deschamps et Grosjean 1972)

Voyons maintenant le détail de la répartition des 149 pauses intraconstituant. Nous avons relevé 11 catégories différentes concernées par ce phénomène. Dans notre corpus, la pause intervient entre :

³⁹ Il l'est moins que chez Campione et Véronis (2004), mais nous avons vu que l'identification des pauses ne se fait pas de la même façon et que nos données ne sont pas comparables.

-
- un déterminant et un nom : 2,01 % (3 occurrences) ;
 - un nom et un adjectif : 3,35 % (5 occurrences) ;
 - une préposition et un groupe nominal : 4,69% (7 occurrences) ;
 - un pronom sujet et un verbe : 1,34 % (2 occurrences) ;
 - un pronom préverbal (autre que sujet) et un verbe : 1,34 % (2 occurrences) ;
 - l'auxiliaire *être* et le participe passé : 2,68 % (4 occurrences) ;
 - l'auxiliaire *avoir* et le participe passé : 0,67 % (1 occurrence) ;
 - un pronom relatif et la proposition relative : 0,67 % (1 occurrence) ;
 - une forme en *que*⁴⁰ et la proposition : 28,19 % (42 occurrences) ;
 - *et* conjonction de coordination et la suite coordonnée : 55,03 % (81 occurrences) ;
 - une conjonction de subordination et la proposition : 0,67 % (1 occurrence).

La classification établie ci-dessus appelle certaines remarques :

1. nous considérons comme un constituant le groupe formé par un pronom préverbal sujet et par un verbe conjugué, alors que le sujet et le verbe forment deux syntagmes différents. Dans ces derniers cas, la pause qui les sépare n'est d'ailleurs pas comptabilisée dans les données ci-dessus. Nous faisons un traitement différent de ces cas car les grammaires qui traitent la séquence « pronom préverbal + verbe » sont très rentables (cf. chap. 6), et cela d'autant plus dans des corpus oraux, où l'utilisation des pronoms connaît une fréquence particulièrement élevée (cf. chap. 2). C'est la raison pour laquelle nous considérons cette séquence comme un syntagme élémentaire, de même que le syntagme « pronom préverbal non-sujet + verbe », pour lequel nous avons aussi une pause intraconstituant ;
2. le groupe nominal est peu concerné par la rupture : 3 occurrences entre le déterminant et le nom et 8 entre le nom et l'adjectif ; s'ajoutent encore 7 occurrences qui séparent la préposition du GN qu'elle introduit ;
3. de même, le groupe verbal est peu concerné par la pause, avec 9 occurrences en tout ;

⁴⁰ Nous avons ici regroupé sous la proforme *que* les *que* conjonction de subordination, les formes comme *parce que*, etc.

4. les cas de coupure les plus fréquents rencontrés concernent les mots introducteurs (conjonctions ou pronoms) et les propositions que ceux-ci introduisent, avec 84 % des occurrences.

Si l'on exclut ces derniers cas, il reste en fait 24 occurrences de ruptures intraconstituant, ce qui, sur l'ensemble du corpus testé, est finalement un chiffre très faible. Ainsi, hormis ces cas, on pourrait considérer que la pause, dans nos transcriptions, se place donc préférentiellement à des endroits pertinents, et permet d'isoler des groupes syntaxiques homogènes.

La distribution de ces 24 pauses intraconstituant, selon la durée de celles-ci, est la suivante :

pause brève	pause longue	silence
17	7	0

Tableau 37. Distribution des pauses intraconstituant

On le voit ici encore, la pause brève est la plus fréquemment concernée ; vient ensuite la pause longue, avec 7 occurrences. Aucun silence dans le corpus ici testé ne se positionne à l'intérieur d'un constituant. Nous avons vu ci-dessus, sur l'ensemble du corpus, que ces configurations pouvaient néanmoins se présenter, même si elles sont relativement rares.

En ce qui concerne la pause longue, 5 des 7 occurrences apparaissent dans l'entrevue accPH1r, dont on a vu que le nombre de pauses longues était supérieur au nombre de pauses brèves, et cela contrairement à la tendance générale du corpus global. On a néanmoins ce cas de figure dans chacun des deux autres textes, une fois entre l'auxiliaire *avoir* et le participe passé, et une fois entre le nom et l'adjectif.

Pour l'entrevue accPH1r, on notera tout de même que le comportement des pauses selon leur degré ne déroge pas à la tendance générale. En effet, le nombre de pauses brèves intraconstituant est supérieur au nombre de pauses longues intraconstituant (8 occurrences contre 5).

Si nous reprenons notre hypothèse de départ, inspirée de Candéa (2000a), selon laquelle la pause serait préférentiellement intraconstituant si elle est combinée à une marque de disfluente, on constate que peu de cas répondent à ce critère formel. En effet, sur ces 23 occurrences, nous avons seulement quatre cas où la pause est combinée à une autre marque : deux occurrences concernent le *euh* (dont un cas cumulé avec une répétition), une séquence contient une onomatopée et une autre un ponctuant. Ces quatre marques précèdent la pause :

accPH1 oui / non ce n'est pas / parce que / le wallon
mm // grossier c'est pas beau mais le wallon m // un
 beau wallon c'est |- beau ça // ah oui [accPH1r]

blaAT1 ouais c'est c'est la folie quoi non mais c'est c'est
 c'est dingue d'être aussi **euu / inconscient** et (...)
 [blaAT1l]

blaAT1 (...) non vous le ferez **pour euu / pour mardi**
 puis mardi ils ne l'avaient pas fait (...) [blaAT1l]

blaAT1 (...) parce que / j'arrive à la voir super noire **tu**
vois // vraiment noire enfin (...) [blaAT1]

4.7 Conclusions sur la pause

Les trois degrés de pauses notés dans les transcriptions obtiennent chacun des fréquences très différentes : les pauses brèves sont de loin les plus nombreuses, suivies par les pauses longues, puis par les silences. Dans les textes de VALIBEL, ces pauses sont des notations subjectives du transcripateur, ce qui rend la comparaison de nos données avec celles d'autres chercheurs la plupart du temps impossible.

Nous avons constaté que ces pauses sont placées le plus souvent entre des groupes syntaxiques cohérents. Nous avons émis l'hypothèse, renforcée par les tests de perception d'autres chercheurs, que le transcripateur place préférentiellement les pauses à des endroits cohérents pour la structuration syntaxique et sémantique de l'énoncé. Néanmoins, si ces pauses apparaissent à des endroits pertinents pour la structuration en groupes syntaxiques, elles ne correspondent pas aux emplacements des marques typographiques de l'écrit telles que le point, la virgule ou le point virgule, puisqu'on les trouve entre un sujet et un verbe ou encore entre un verbe et un complément d'objet direct.

Lorsque ces pauses se placent à des endroits non pertinents en termes d'analyse syntaxique, nous avons distingué deux cas : le premier constitué des occurrences où la pause est intraconstituant, le deuxième où la pause est placée au lieu d'une rupture syntaxique (le syntagme qui la précède est inachevé). Tous degrés de pauses confondus, le premier groupe contient deux fois plus d'occurrences que le second ; la tendance s'inverse en ce qui concerne le silence, mais le nombre d'occurrences dont nous disposons est très réduit.

L'examen attentif des marques groupées ne montre rien de concluant. Pour ce qui est de la rupture syntaxique, les marques combinées sont même moins fréquentes que les cas où la pause apparaît seule.

Les pauses intraconstituant ont des distributions variables : on a vu que dans le sous-corpus ici analysé, on a 11 configurations différentes, la plus fréquente étant entre une conjonction et la proposition introduite par cette conjonction.

Au vu des résultats que nous obtenons ci-dessus, on peut penser que la marque du silence est un bon candidat au découpage du texte en unités plus petites. En ce qui concerne les pauses brèves et longues, on voit qu'elles le sont dans la majorité des cas. Néanmoins, les occurrences où elles ne le sont pas s'avèrent problématiques en regard des contraintes liées à notre système de levée d'ambiguïtés.

4.8 Le point d'interrogation

Seul emprunt à la ponctuation classique dans les transcriptions, le point d'interrogation apparaît 1920 fois dans le corpus. Sa présence correspond à une intonation montante typique d'une question, lorsque celle-ci prend une forme déclarative, c'est-à-dire sans inversion du sujet ou sans les tournures interrogatives classiques du type *est-ce que* ou *qui est-ce qui*⁴¹. Le point d'interrogation est indiqué dans la transcription afin de désambiguïser ce qui pourrait, sans cette marque, être lu comme une assertion :

accCT0 s'ils parlent en français -| euh il y a pas de problème ? [accBF1r]

Hormis 33 cas, la marque de la question intervient toujours après une structure syntaxique achevée. Les exceptions peuvent sembler étranges à la lecture de la transcription, mais elles apparaissent pourtant comme des questions naturelles à l'oral, comme on le voit dans les exemples suivants :

magPM1 en général je prépare une liste alors quand j'ai une liste de faite / ça va vite
magFM0 une liste de ?
magPM1 une liste / avec toutes mes courses (...)
 [magPM1r]

ilrGS2 ben oui mais c'est pas maintenant c'est c'est pas le carnaval ça c'est le grand feu quoi
ilrMF0 oui ? tu peux me raconter comment ça se passe parce que je ne ?
ilrGS2 le grand feu ben c'est les jeunes du village ils se rassemblent et puis ils vont ramasser / des branches enfin des bois et tout ça pour faire un grand feu et puis

⁴¹ Noter un point d'interrogation pour toutes les questions aurait évidemment donné un nombre beaucoup plus élevé de points d'interrogation dans les transcriptions.

le soir ils font ils allument le feu et puis ils font un
goûter [ilrGS2r]

Du strict point de vue de la structure syntaxique, on considère que le constituant est inachevé, dans le premier exemple après une préposition, dans le second, après le groupe préverbal sujet et la négation. Dans ce dernier cas, se pose la question de ce que nous entendons par rupture syntaxique et constituant. Nous avons déjà abordé ce problème de la définition du constituant *supra*, et nous considérons bien ici qu'il y a rupture.

On notera que cette marque de la question intervient dans les exemples ci-dessus en fin de tour de parole⁴². C'est d'ailleurs le cas pour 75 % des occurrences de cet indice de question. Cela ne nous étonnera pas, dans la mesure où la majorité des entrevues sont des entretiens semi-dirigés, où l'un des locuteurs a le rôle d'intervieweur et pose donc des questions à son interlocuteur. Néanmoins, il est évident que les questions ne sont pas absentes des conversations informelles.

On rencontre également 25 cas où la marque du point d'interrogation se place après une conjonction de coordination :

ilrMF0 oui mm / et ou sinon tu travailles toute la
semaine ou ?
ilrGS2 ah non je vais à l'école toute la semaine sauf le
mercredi [ilrGS2r]

Certains contextes dans lesquels on rencontre le point d'interrogation sont constitués de tours de parole très brefs, parfois composés d'un seul mot. Ils relèvent de la problématique plus générale des énoncés réponses :

famRM1 (...) et alors au fond euh / au début de notre
mariage on faisait toujours une tournante de jeux
famMO1 de jeux ?
famRM1 oui on faisait on jouait au scrabble on jouait
au Trivial-Pursuit (...) [famRM1r]

Dans nos données, le point d'interrogation se place pratiquement toujours après une structure syntaxique achevée. Quand ce n'est pas le cas, il apparaît de toute façon, sauf dans de rares exceptions, à la fin d'un tour de parole. Il constitue donc dans le texte une marque pertinente pour le découpage de celui-ci en unités plus petites.

⁴² Une seule exception parmi les 33 cas de structure syntaxique inachevée :

ilrMF0 donc tu es en dernière année | - **de ?** <ilrGS2> oui -|
aide familiale [ilrGS2r]

5 LE PRÉTRAITEMENT DU CORPUS AVANT UNITEX

Le prétraitement du corpus s'est fait en deux étapes : une première étape *avant* le chargement du corpus dans le logiciel UNITEX, et une seconde étape avec des graphes construits dans UNITEX (§ 6.1).

Nous entendons ici par prétraitement les différentes étapes de marquages ou de transformations du texte (sans perte aucune des éléments qui y figurent), de « préparation » de celui-ci afin de le rendre compatible avec le logiciel de traitement de corpus UNITEX, le système d'étiquetage et les grammaires que nous avons développés.

En effet, si elles sont compatibles avec une informatisation des données, les transcriptions de VALIBEL n'en comportent pas moins des informations qui perturbent la linéarité de la lecture⁴³ : celles-ci sont transcrites entre parenthèses, entre crochets ou encore à l'intérieur de balises. Ces éléments ne doivent évidemment pas recevoir une analyse morphosyntaxique lors de l'étiquetage.

Le prétraitement effectué sur le texte avant son utilisation dans le logiciel UNITEX concerne six points que nous allons détailler ici : la référencement des locuteurs par des codes univoques, la normalisation des conventions de transcription, l'extraction des tours de paroles internes, la notation des multitranscriptions, le balisage des disfluences que nous avons étudiées au chapitre 3 et enfin les éléments notés entre parenthèses.

5.1 La référencement des locuteurs et des textes

Pour constituer notre corpus, nous avons regroupé des textes qui appartenaient à des sous-corpus différents. Ce regroupement a nécessité de préciser la référencement des locuteurs. En effet, dans chaque transcription intégrée dans la banque de données VALIBEL, les locuteurs recevaient un code (L1, L2, L3, etc.) en fonction de leur « arrivée » dans l'enregistrement (L1 pour le premier locuteur, etc.). L0 se voyait réservé au locuteur qui mène l'entretien avec un questionnaire (entretien semi-directif). Des fiches d'identification étant associées à chaque enregistrement et à chaque locuteur (les méta-données), l'identification de celui-ci se faisait aisément par le recours à ces fiches.

Dans la mesure où nous avons regroupé dans un même fichier 60 textes appartenant à 12 sous-corpus différents, il s'avérait essentiel de pouvoir

⁴³ Nous ne parlons pas ici de phénomènes propres à l'oral qu'il est pertinent de conserver pour l'analyse.

identifier chaque locuteur et chaque texte de manière univoque⁴⁴. Un code individuel a donc été associé à chaque locuteur ; il se compose de trois lettres minuscules (pour le nom du sous-corpus auquel il renvoie), de deux lettres majuscules (initiales nom prénom) et d'un chiffre pour différencier les éventuels homonymes.

Le nom de l'entrevue est quant à lui composé du nom du locuteur principal, suivi de la lettre *r* (pour *record*) ou *l* (pour *conversation libre*, dans les entrevues reprises du projet PFC). Une balise en début et en fin de texte permet de noter le passage à un autre texte.

5.2 La normalisation des conventions de transcription⁴⁵

Les conventions de transcription du Centre de recherche VALIBEL ont été revues au cours de cette recherche (Dister *et al.* 2006, Dister et Simon 2007). Bien que touchant des points mineurs, qui ne remettaient pas en cause les choix fondamentaux pris antérieurement, certaines modifications ont dû être apportées. Par exemple, les anciennes conventions prévoyaient de lier par un trait d'union des formes qui devaient être analysées comme une seule unité⁴⁶, bien que ce trait d'union soit absent de l'orthographe standard. On avait ainsi : *au-fur-et-à-mesure*, *à-condition-de*, *au-jour-le-jour*, *au-péril-de*, *avant-que*, *avant-de*, *bien-que*, *de-crainte*, *de-crainte-de*, *dès-lors*, *de-temps-à-autre*, *en-fonction-de*, *grâce-aux*, *grandes-surfaces*, *n'est-ce-pas*, *par-comparaison*, *tout-d'un-coup*, etc. Inversément, afin de permettre la reconnaissance de deux mots dans *l'enfant*, *j'avais*, *prenez-en* ou *voulez-vous*, un blanc était inséré après l'apostrophe et avant le trait d'union : *l' enfant*, *j' avais*, *prenez -en* ou *voulez -vous*.

Ce travail de normalisation a été effectué de manière totalement automatisée, grâce à des programmes en Perl, repris essentiellement du travail de stage effectué par Hubert Naets (Naets 2001).

⁴⁴ Indépendamment de notre travail, cette identification univoque des locuteurs s'est avérée nécessaire pour l'encodage dans le nouveau système de gestion des corpus qu'est [moca] (Kevers *et al.*, 2006). Les choix de codage des locuteurs faits ici sont redevables de nombreuses discussions au sein de l'équipe VALIBEL, et ne sont donc pas spécifiquement liés à notre projet.

⁴⁵ L'étape de normalisation très brièvement décrite dans ce paragraphe consiste également en la vérification minutieuse de points mineurs (en termes de fréquence d'occurrences), comme par exemple la normalisation des sigles par suppression des points, la graphie des abréviations usuelles que nous avons développée (*etc.*, devenu *et cetera*, également afin de supprimer les points) ou encore les hésitations du transcrip-teur, qui prennent maintenant la forme de multitranscriptions. Afin d'homogénéiser autant que faire se peut les données, cette étape a cependant nécessité un important investissement en temps.

⁴⁶ Le logiciel utilisé alors était WordSmith Tools.

Les nouvelles conventions de transcription insistent particulièrement sur la nécessité de privilégier autant que faire se peut l'orthographe standard, les pronoms *il* et *ils*, qui étaient orthographiés *i* quand ils étaient prononcés [i], ont été rétablis en *il* ou *ils*. Là aussi, nous avons procédé de manière automatisée pour le rétablissement de ces graphies, au moyen de graphes construits avec UNITEX (§ 6.1). Le reliquat, non pris en compte par nos programmes, a été corrigé manuellement.

5.3 Chevauchements de parole

Les chevauchements de parole (« overlapping segment ») brisent la linéarité de la lecture du texte transcrit. En effet, quand le chevauchement est interne au tour de parole (c'est-à-dire qu'un locuteur B commence à parler alors qu'un autre locuteur A est déjà en train de parler, mais sans prendre définitivement la parole puisque le locuteur A poursuit son propos), la transcription prend la forme suivante (voir aussi chap. 1) :

accCP1 (...) on a un peu l' **|- impression <accCT0>**
oui -| qu'ils sont complexés d'être Belges // et qu'ils ne
 veulent pas montrer qu'ils ont cet accent-là quoi
 [accCP1r]

ilrMF0 (aspiration) ben je préférerais pas si **|- ça ne te**
<ilrTC1> ça va alors -| dérange pas [ilrTC1r]

Dans la quasi-totalité des cas, comme dans ces deux exemples, ces chevauchements internes ne brisent pas la structure syntaxique de l'énoncé en cours : le locuteur A poursuit, certes avec une intrusion de l'interlocuteur dans son discours, mais il achève néanmoins ses propos selon une structure syntaxique « classique ». Le lieu du chevauchement ne constitue pas une rupture.

Dans notre corpus, deux grands cas de figure se présentent pour ces chevauchements internes, et qui expliquent que la linéarité de la structure syntaxique de l'énoncé n'est pas brisée :

- devant la tentative de prise de parole du locuteur B, le locuteur A continue à parler comme s'il n'était pas interrompu, la stratégie étant justement de poursuivre « comme si de rien n'était » et de garder la parole malgré cette interruption qu'il peut juger intempestive⁴⁷ ;

⁴⁷ Il n'est pas ici question de dire que le chevauchement n'a aucune influence sur le locuteur, et que tout se passe comme s'il n'y avait pas d'intervention du locuteur B. Ainsi, la prise en compte de la prosodie montrerait clairement que dans bien des cas, le locuteur A, pour ne pas être interrompu, développe plusieurs stratégies : il parle plus fort, accélère son débit, etc. Par ailleurs, nous faisons aussi l'hypothèse que le chevauchement provoque certaines répétitions (cf. chap. 3).

- le chevauchement n'est pas une tentative du locuteur B de prendre la parole, mais au contraire une invitation à poursuivre, un encouragement à continuer qu'il lance au locuteur A (cf. Cook 1995 : 36). Ce cas de figure, où l'intervention du locuteur B est non pas une menace mais une marque de coopération à la conversation, constitue la majorité des chevauchements internes rencontrés dans les textes des entrevues semi-dirigées. L'intervieweur invite le locuteur à poursuivre en ponctuant son discours par de multiples interventions, souvent très courtes⁴⁸, signes d'assentiment : *oui, ouais, ouais ouais, c'est ça, ah oui, m, mm, non, non non, etc.* Ces marques sont la preuve d'un échange cordial entre les locuteurs, et ne constituent pas un danger pour la prise de parole du locuteur A. Les locuteurs ne sont pas ici dans un affrontement, dans une lutte pour la prise de parole qui serait envisagée comme un combat.

Dans tous les cas, les chevauchements internes ont donc été extraits des tours de parole dans lesquels ils figurent, pour en faire des tours de parole à part entière. Les deux exemples ci-dessus deviennent donc, après extraction :

accCP1 (...) on a un peu l' |- impression **@67** -| qu'ils sont complexés d'être Belges // et qu'ils ne veulent pas montrer qu'ils ont cet accent-là quoi

@67 accCT0 oui

ilrMFO (aspiration) ben je préférerais pas si |- ça ne te **@1132** -| dérange pas

@1132 ilrTC1 ça va alors

On a inséré le signe @ qui permet de garder la trace de l'existence du chevauchement interne. Celui-ci est noté en dessous du tour de parole dans lequel il figurait. Les tours de parole internes sont numérotés afin d'éviter tout risque de confusion, notamment lorsqu'on a plusieurs chevauchements internes dans un même tour de parole.

Dans le corpus, nous avons 2545 tours de parole internes que nous avons ainsi extraits. Ces chevauchements internes seront réintégrés à leur place originale dans le texte étiqueté, garantissant par là à la banque de données VALIBEL une homogénéisation des transcriptions, que celles-ci soient annotées morphosyntactiquement ou non.

Néanmoins, malgré l'extraction, on voit que des indices restent dans le texte, qui brisent toujours sa linéarité : non seulement les barres de début et de fin de chevauchement (respectivement |- et -|), mais également le signe @ et le nombre associé à l'ordre d'apparition du chevauchement dans le texte. Nous verrons ci-dessous (§ 6.2.2) comment nous résolvons ce problème.

⁴⁸ Voir également le chap. 2.

5.4 Les multitranscriptions

Lorsque le transcripateur n'est pas certain de ce qu'il entend, ou qu'il hésite entre deux possibilités, il utilise un marquage spécifique dans le texte pour noter cette incertitude : s'il hésite mais n'a pas d'autre solution à proposer, il note ce qu'il croit entendre entre accolades ; s'il hésite entre deux possibilités (ou plus), il note celles-ci entre accolades, séparées par une virgule entourée de blancs. Ces deux cas de figure peuvent être représentés comme suit :

{hésitation sans solution alternative}

{hésitation1 , hésitation2}.

Afin de faciliter la suite du traitement, nous réécrivons les chevauchements de la manière suivante, en utilisant un marquage avec des balises :

<m> hésitation sans solution alternative </m>

<m1> hésitation1 </m1> <m2> hésitation2 </m2>⁴⁹

Nous développerons ci-dessous (§ 6.2.2) la nécessité qui nous a poussée à procéder ainsi.

5.5 Les crochets

La phonétique, en alphabet Sampa⁵⁰, est notée entre crochets. Cette information est ajoutée afin de préciser la prononciation, souvent non standard, d'un mot ou d'un groupe de mots, comme dans l'exemple suivant :

ilcDA1 le secteur de Zaventem l'aéroport [aReopoR] /
tous ces gens qui travaillent pour cette région-là / ça ce
sont des Flamands qui ne doivent pas euh répondre du
français [ilcDA1r]

On garde donc la graphie classique du mot, et la prononciation non standard est notée entre crochets. Cette notation figure donc en plus des données langagières et son insertion ne constitue pas une rupture de l'énoncé.

⁴⁹ La pratique de la multitranscription a été adoptée récemment au Centre VALIBEL. Les transcriptions plus anciennes en sont donc exemptes. Dans notre corpus, nous avons recensé 180 cas <m>, 46 cas <m1> <m2> et un seul cas <m1> <m2> <m3>.

⁵⁰ <http://www.phon.ucl.ac.uk/home/sampa/>.

Dans les corpus anciens, qui traitent directement de faits de prononciation (enquête sur la norme, par exemple), on voit que les transpositeurs ont utilisé les notations phonétiques en lieu et place du son lui-même :

norFA0 moins deux caractéristiques / euh la
différenciation entre [e~] et [9~] [norKJ1r]

Ici, la notation phonétique prend intégralement place dans la transcription, et n'est donc aucunement facultative. Dans ces cas relativement nombreux⁵¹, nous avons inséré le son entre balises, indiquant par là qu'il s'agissait bien d'une partie intégrante de l'énoncé du locuteur et non d'un ajout, comme c'est le cas dans l'exemple ci-dessus avec [aReopoR]. La transcription est alors :

norFA0 moins deux caractéristiques / euh la
différenciation entre <son> [e~] </son> et <son>
[9~] </son> [norKJ1r]

De la même manière, des formes transcrites en orthographe standard, accompagnées de la transcription phonétique, ont également reçu des balises de ce type. C'est le cas lorsque le mot est pris par le locuteur comme exemple pour illustrer une prononciation. La forme entre alors dans une structure syntaxique qui n'est pas une structure classique ; la solution la plus adéquate, et permettant une homogénéité dans les transcriptions, est de baliser ces termes choisis uniquement pour leur prononciation, et non suivant des contraintes morphosyntaxiques, de la même manière que les sons. L'exemple suivant illustre cela, où l'on voit que l'une des occurrences de *punch* n'est pas balisée, car elle n'apparaît pas dans la séquence pour sa prononciation mais bien comme une occurrence du mot qui doit être analysée dans l'énoncé :

norKJ1 euh pf bon on fait des petites découvertes à
tout âge quoi hein ce n'est pas <son> **punch**
[p9~nS] </son> mais <son> **punch** [po~S]
</son> / quand c'est une boisson / bon il m'a fallu
attendre la trentaine / |- moi qui suis <norFA0> ah -|
pourtant / amateur de **punch** [po~S] pour savoir que /
fallait prononcer ça <son> **punch** [po~S]
(...)</son> [norKJ1r]

On a donc des crochets qui illustrent deux cas de figure différents : un premier cas où la phonétique est ajoutée aux paroles du locuteur ; un second cas où la phonétique fait partie intégrante du texte et est entourée de balises.

⁵¹ Nous avons également normalisé les cas où [e~] et [9~] étaient transcrits simplement *un* et *in*, sans aucune autre indication phonétique.

5.6 Les balises <s> et </s>

Lorsqu'un locuteur s'exprime dans une autre langue que le français, la portion en langue étrangère est notée entre les balises <s> et </s>, pour noter l'alternance de code. Cette séquence ne doit, et ne peut, évidemment pas être annotée morphosyntaxiquement : les mots qui la composent n'appartiennent pas au français et ne sont pas répertoriés dans le dictionnaire de la langue (cf. chap. 5).

Dans les transcriptions originales, cette séquence était marquée entre les balises (A+) pour le début de l'alternance et (A-) pour la fin de celle-ci. Néanmoins, lorsque l'autre langue était le wallon ou le marollien, la séquence était notée en italique, marquage typographique perdu lors du passage au format texte. Nous avons donc inséré les symboles <s> et </s>⁵² pour baliser cette portion de texte.

Ce balisage pose néanmoins des problèmes théoriques sur le statut que l'on accorde à certaines séquences. En effet, certains cas sont limpides : on a bien affaire à une autre langue que le français, comme dans l'exemple ci-dessous :

chaBP1 après deux trois jours ça revient très vite hein
l'a/ l'allemand hein / ça revient très vite et cetera mais je
ne saurais pas tenir une conversation / euh il faut faut
qu'on soit gentil avec moi il faut des mots simples et
cetera et alors nous étions prisonniers avec des Anglais
j'ai voulu apprendre l'anglais mais / à part **<A> I love**
**you et will you kiss me ** [chaBP1r]

Ici, il est très clair qu'on est en présence d'une portion de texte en anglais, et que faire entrer les différentes formes dans un dictionnaire, leur attribuer une catégorie grammaticale, afin de pouvoir les analyser au moyen d'une grammaire construite pour le français, n'aurait aucun sens.

Outre ces cas tranchés, il arrive que l'on soit en présence de formes dont le statut n'est pas aussi clair. En effet, on se trouve parfois sur un continuum où il faut décider⁵³, pour des questions de méthodologie, si l'on est dans du

⁵² s pour *switch code*, *alternance de code*.

⁵³ Pour éviter tout parti-pris théorique, les conventions actuelles préconisent de ne baliser aucune séquence comme une alternance de code (même dans les cas qui ne prêtent pas à confusion), cette notation relevant d'une analyse préalable. De plus, ce marquage induit un statut différent donné à chaque langue : une langue de base non marquée (en l'occurrence le français) et une autre langue qui apparaît comme marquée. Ce problème se pose de manière cruciale pour des enregistrements faits dans le cadre de l'Action de recherche concertée « Régulation de l'hétérogénéité linguistique en milieu multiculturel. Le français en contact dans les écoles bruxelloises », où plusieurs langues se côtoient, le français étant une langue parmi d'autres.

« hors français » ou dans du français, avec les implications théoriques que cela a sur le statut accordé à ces portions de texte. C'est le cas de formes que l'on pourrait considérer non pas comme du français régional, mais comme appartenant toujours au dialecte. Ainsi les exemples suivants :

blaGD1 ouais ça me pose des problèmes moi parce que
/ enfin non j'ai encore vu un article **kimin** de // Lushel
[lySEl] là [blaGD1l]

blaGD1 mais c'est sûr que je ne sais **nin** il a peut-être
// il a sûrement septante ans hein tout compte |- fait
[blaGD1l]

Le locuteur, un avocat, utilise des mots issus du dialecte (le wallon) : *kimin* veut dire *comment*, et il fonctionne ici exactement comme le mot en français, marquant la recherche lexicale à laquelle se livre le locuteur. C'est le travail de l'énoncé en cours d'élaboration qui est visible ici. Dans le second extrait, il emploie *nin* à la place de l'adverbe français *pas*.

Tout le problème est de savoir si ces formes doivent être intégrées dans un dictionnaire, avec les implications que cela peut avoir sur le statut qui leur est alors conféré.

Pour ces cas, nous avons choisi de les intégrer dans le dictionnaire, afin de pouvoir analyser ces phrases dont la structure syntaxique est celle du français.

Notons que les textes qui composent notre corpus contiennent relativement peu d'alternances de code⁵⁴.

5.7 Le balisage des disfluences étudiées

Les analyses que nous avons faites de quatre disfluences poursuivaient deux objectifs principaux : caractériser le corpus, les entrevues et les locuteurs en fonction de ces disfluences ; étudier celles-ci, principalement d'un point de vue formel et syntaxique, en vue de les intégrer dans le processus de l'annotation morphosyntaxique.

Nous allons maintenant nous pencher sur le traitement que nous faisons des disfluences, au vu des analyses et des résultats que nous avons observés au chapitre 3.

On aura compris que nous ne pouvons nous passer d'un codage particulier et qu'il nous fallait décider ce qui appartient au français (et doit donc être intégré dans le dictionnaire).

⁵⁴ Nous avons d'ailleurs évité de reprendre des entrevues du corpus *Marolles*, dans lesquelles les alternances de codes étaient très nombreuses.

Nous avons choisi de traiter les disfluences dans une phase de préédition des données (cf. Valli et Véronis 1999). Avant même son utilisation dans le logiciel UNITEX, le texte est donc modifié. Cette modification ne consiste pas en un nettoyage brutal des données, qui en supprimerait certaines du texte en le normalisant et en le faisant tendre vers le standard. En fait, il s'agit pour nous de baliser les disfluences, de manière à pouvoir traiter efficacement le texte en vue de l'objectif final qui est le nôtre : l'annotation morphosyntaxique.

5.7.1 Les répétitions

Nous avons choisi de baliser les répétitions, pour ne garder de la séquence répétée que la forme finale corrigée. En d'autres termes, on garde le dernier répété de la séquence. On aurait pu choisir de conserver le répétable, mais garder le dernier répété présente un double avantage, à la fois théorique et pratique : il est en accord avec le processus de réparation de la séquence, où c'est le dernier mot (ou la dernière séquence de mots) qui est gardé par le locuteur – même si formellement, cette séquence n'est pas différente –, et il procède d'une même façon de faire pour les autocorrections et pour les amorces (cf. ci-dessous).

Nous avons donc inséré les balises <r> et </r> respectivement avant le répétable et avant le dernier répété, comme l'illustre l'énoncé suivant :

norPM1 <r> c'est le </r> c'est le risque // il faut savoir ce qu'on fait / quand on fait un dictionnaire de prononciation est-ce qu'on veut faire un dictionnaire / normatif comme le fait monsieur Warnant / avec un choix de témoins / bien sûr limité // **<r> de de </r>** de vingt à quarante ans dans telle euh population // avec les défauts de la norme et l'excès que présente une norme mais / je pense que nous sommes suffisamment adultes pour pouvoir prendre / une distance par rapport à cette norme mais on est / content de l'avoir / tout en sachant qu'il y a des / niveaux / des variances possibles // ou bien alors on fait un dictionnaire descriptif / de la prononciation / dans la francophonie a/ avec / une multitude **<r> de </r> de** notations pour chaque mot // et on perd la notion de norme et ces et le l'ouvrage présente un autre intérêt (silence) **<r> les deux </r>** les deux sont intéressants [norPM1r]

Ces séquences balisées seront ensuite ignorées par les grammaires locales de levée d'ambiguïtés, mais sans perte de la trace de ce travail d'élaboration de l'énoncé dans les transcriptions. Dans une phase de postédition, nous attribuerons à la séquence balisée la même catégorie morphosyntaxique que celle dévolue au dernier répété analysé.

Les marques que nous avons sélectionnées comme pouvant être insérées à l'intérieur d'une séquence répétition sont également mises entre les balises :

blaGD1 il est -| au parquet de Namur lui
blaBD1 ah oui |- carrément
blaGD1 euh -| <r> de // non </r> de Marche (xx)
 [blaGD1r]

blaJD1 (...) mais j'ai déjà resté facilement un / plus d'
<r> un q/ / oh oui </r> un bon quart d'heure /
 malgré tout (silence) [blaJD1l]

Dans les données sur les répétitions que nous avons présentées au chapitre 3.2, chaque séquence était envisagée séparément, même si elle cooccurrait directement avec une autre séquence répétition. Ainsi, dans l'énoncé suivant, nous avons deux répétables différents, et donc deux séquences répétition :

ilrMF0 ah tu n'as pas mis le nom du village de / de
ton de ton village c'est Angreau ? [ilrGS2r]

Nos résultats répertoriaient le répétable *de*, répété une fois, et également le répétable *de ton* répété également une fois. En fait, on assiste là à ce que nous appelons un empilement, et la séquence disfluente totale est en fait la concaténation des deux séquences isolées : *de / de ton de ton*. En définitive, la réparation de cette séquence, à laquelle aboutit le locuteur, est la suite *de ton*. C'est aussi celle qui nous intéresse, celle que nous souhaitons garder dans le texte. Il s'agit donc pour nous de repérer la séquence maximale disfluente, et de ne garder que la réparation finale. L'étiquetage de cette séquence devient alors :

ilrMF0 ah tu n'as pas mis le nom du village <r> de / de
ton </r> de ton village c'est Angreau ? [ilrGS2r]

Des séquences encore plus complexes sont reconnues, comme dans l'énoncé suivant que nous prenions en exemple au début de la section consacrée aux répétitions :

iljDV1 (...) non on a l'habitude maintenant de de parler
 euh devant un micro de parler devant des gens / donc
 ça m'effraie vraiment plus du tout / plus du tout du
tout du tout [iljDV1r]

qui donne après balisage :

iljDV1 (...) non on a l'habitude maintenant <r> de </r>
 de parler euh devant un micro de parler devant des gens
 / donc ça m'effraie vraiment <r> plus du tout / </r>
plus <r> du tout du tout </r> du tout [iljDV1r]

On conserve donc au final la séquence réparée

iljDV1 (...) non on a l'habitude maintenant de parler
euh devant un micro de parler devant des gens / donc
ça m'effraie vraiment **plus du tout** [iljDV1r]

On peut trouver cette solution quelque peu réductrice de la réalité linguistique, car on assiste en fait dans cet énoncé à un empilement de répétitions qui ont une valeur intensive. Néanmoins, rappelons 1) que nous ne prétendons pas baliser uniquement les répétitions disfluentes mais souhaitons intégrer les répétitions formelles dans une description unifiée et 2) qu'un tel traitement ne nuit pas aux données, dans la mesure où il ne s'agit pas de nettoyer définitivement le texte pour le rendre « standard », mais simplement de faciliter son étiquetage morphosyntaxique.

S'il s'agissait d'analyse syntaxique, nous devrions sans doute envisager d'autres manières de procéder.

Ainsi donc, comme nous venons de le redire, notre méthode un peu brute de repérage des répétitions reconnaît des séquences qui relèvent de répétitions de langue et ne sont donc pas disfluentes. Ici, on doit bien distinguer deux cas de figure, où les implications d'un balisage sont de nature totalement différentes :

1. les éléments balisés reçoivent la même analyse morphosyntaxique que le répété conservé et l'énoncé n'est pas radicalement modifié par l'effacement momentané d'une partie de la séquence. En d'autres termes, chaque élément reçoit l'analyse qu'il doit recevoir, indépendamment du balisage. C'est le cas de toutes les répétitions qui relèvent de la disfluence, mais aussi des répétitions intensives. C'est le cas des exemples fournis ci-dessus ;
2. les éléments balisés reçoivent la même analyse morphosyntaxique que le répété conservé mais la structure syntaxique de l'énoncé est modifiée par l'effacement momentané d'une partie de la séquence.

Dans ce deuxième cas de figure, on trouve les répétitions du type *elle elle* où le premier *elle* est le complément d'une première phrase et le second *elle* le sujet d'une autre phrase⁵⁵ :

famRM1 oui / tant pis pour **elle / elle** n'a qu'à pas dire
{que} c'était sa |- famille // hein ? [famRM1r]

On a aussi les cas de répétitions de *ça*, sur le même schéma de répartition des deux *ça* à deux syntagmes différents :

⁵⁵ Dans la section sur les répétitions, nous donnions des exemples de *nous nous*.

accBF1 mais à part ça ça dépend des gens //
quelques-uns me disent que j'ai un accent puis d'autres
non [accBF1r]

Ce type d'occurrences est évidemment gênant. Néanmoins, ces cas sont très rares en regard 1) du nombre d'occurrences effectivement disfluentes de *elle elle* ou de *ça ça* ; 2) du nombre de cas où sans être disfluentes mais emphatiques, répétables et répétés n'appartiennent pas à deux syntagmes différents.

Dans la mesure où on ne fait pas d'analyse syntaxique, et où l'étiquette sera la même pour les deux occurrences, on peut les traiter utilement de la sorte.

5.7.2 Les autocorrections

Les autocorrections que nous avons analysées ont un fonctionnement analogue aux répétitions. Nous balisons le *reparandum* avec les balises <c> et </c>, comme le montre l'exemple suivant :

ileFN1 et le journalisme et puis euh <c> le </c> les
études de journalisme en soi ne me plaisaient pas
[ileFN1r]

On a mentionné l'interaction de l'autocorrection et de la répétition. Dans ces cas, c'est également la séquence maximale qui a été balisée, pour ne conserver que le *repair*. Ainsi, dans l'exemple suivant, on voit qu'entrent en jeu trois phénomènes :

accCP1 il est un peu peut-être chantant sur le les les
les fins / les fins de mots (...) [accCP1r]

On a une autocorrection du *le* en *les* ; le *repair* de l'autocorrection, *les*, est lui-même répétable (avec deux répétés) ; le dernier répété de cette séquence répétition entre dans la composition d'un répétable avec un substantif : *les fins*. Ce qui nous intéresse est de ne garder que le *repair* de toute cette séquence disfluente. Le balisage de celle-ci est le suivant :

accCP1 il est un peu peut-être chantant sur <c> le les
les les fins </c> / les fins de mots (...) [accCP1r]

5.7.3 Les amorces

Notre méthode d'étiquetage se base sur des ressources lexicales constituées par des dictionnaires électroniques. Une forme dans le texte, pour pouvoir être analysée par le système d'annotation, doit préalablement être reconnue par l'analyse lexicale et doit donc figurer dans le dictionnaire. Ce n'est à l'évidence pas le cas pour les formes amorcées. Celles-ci ne

peuvent donc recevoir une analyse, et sont étiquetées par défaut « mot inconnu », ce qui est peu satisfaisant.

Oostdijk (2003), dans un article que nous avons déjà mentionné intitulé *Normalization and disfluencies in spoken language data*, propose de normaliser les amorces en les complétant, c'est-à-dire en les remplaçant par leur forme pleine (normalisation qui appartient à la catégorie des *normative insertion*). Cette procédure nous semble peu en accord avec notre méthode. En effet, il s'agit d'une normalisation exclusivement manuelle, faite obligatoirement par un expert humain, là où le prétraitement que nous effectuons sur le texte est une étape totalement automatisée. De plus, une telle proposition simplifie un peu les problèmes posés par l'interprétation des amorces et fait fi de tous les cas pour lesquels il est impossible de reconstituer la forme entière du mot (cf. chap. 3.4).

En fait, si l'amorce en tant que telle est aisément identifiable, et de manière univoque, dans les transcriptions, toute la complexité réside dans le repérage des types d'amorces. Or, nous avons vu que les amorces avaient un comportement syntaxique différent selon qu'elles étaient complétées/corrigées ou abandonnées : piétinement sur l'axe syntagmatique dans le premier groupe, rupture syntaxique (ou effet de liste coordonnée) dans le cas des amorces abandonnées. De plus, le départ entre les deux dernières catégories (amorces corrigées et amorces abandonnées) est parfois une tâche malaisée lorsque l'expert est humain ; il est pratiquement impossible lorsque celui-ci est une machine (du moins, au stade où nous en sommes, où aucune analyse n'a encore été effectuée). Ainsi, s'il n'est pas possible de distinguer les trois sous-catégories d'amorces établies, on peut néanmoins formuler des heuristiques quant au traitement de certaines d'entre elles.

En effet, en fonction des éléments que nous avons pu faire émerger lors de l'analyse des amorces de notre corpus, qui mettent en évidence certaines régularités, nous pouvons faire certains choix de prétraitement pour les amorces. Si ceux-ci ne sont pas satisfaisants pour tous les cas, ils en règlent néanmoins certains autres.

La première catégorie, simple à traiter, concerne les rares cas (14) d'amorces complétées sans reprise à l'initiale de la forme amorcée. Nous avons vu que ces deux segments ne constituaient pas des lexèmes analysables, alors que regroupés ils pouvaient être soumis à la reconnaissance par les dictionnaires. Nous reformons donc le lexème complet, en gardant dans le texte la trace de la construction initiale.

La deuxième catégorie que nous distinguons concerne les amorces que nous pourrions appeler bégayantes : il s'agit de la sous-catégorie d'amorces complétées sans reprise d'éléments antérieurs (750 cas). L'hypothèse, que fait d'ailleurs déjà le transcripteur lors de l'établissement des données, est bien que le locuteur corrige en répétant la forme reprise à l'initiale. Il s'agit

d'un phénomène en tout point comparable à celui de la répétition, si ce n'est que le répétable n'est pas identique au répété puisqu'il n'est pas un lexème entier.

Dans ces cas, l'amorce est entourée de balises <a> et ; cela nous permettra par la suite de repérer une portion de texte dont on considère qu'elle est ajoutée au texte à analyser et qui doit, dès lors, être ignorée.

C'est une procédure semblable que nous adoptons pour les amorces complétées avec reprise du contexte gauche de l'amorce (des pronoms dans le cas du verbe amorcé, des déterminants et de la préposition dans le cas des substantifs) et elles concernent 1094 occurrences. Ces amorces sont celles qui entrent dans la construction d'une séquence répétition, et nous avons vu que dans 96,7 % des cas, il n'y avait pas de rupture syntaxique.

Là aussi, on a donc un comportement comparable à la répétition, et le traitement est donc similaire. Ces cas sont d'ailleurs inclus dans l'analyse des répétitions et reçoivent les balises <r> et </r>.

Pour ce qui est des amorces corrigées, deux types sont possibles à repérer automatiquement : 1) celles dont nous venons de parler et qui s'intègrent dans une séquence répétition : le *repair* reprend le contexte gauche de l'amorce ; 2) celles qui relèvent d'une des formes d'autocorrection que nous avons analysées, avec une amorce entre la forme de départ et la forme corrigée (déterminant, pronom sujet, etc.).

Pour ce qui est des amorces corrigées contigües, celles-ci sont impossibles à différencier des amorces laissées inachevées. Nous considérons donc que tous les autres cas constituent potentiellement une rupture syntaxique de l'énoncé et que le traitement s'arrête donc après l'amorce. Ces cas sont laissés tels quels et recevront une balise dans le traitement avec UNITEX.

5.7.4 Les *euh*

Nous avons vu que le *euh* marque un piétinement sur l'axe syntaxique avec, dans la très grande majorité des cas, reprise du déroulement normal ensuite. L'analyse des marques mixtes ne nous a pas permis de dégager des régularités qui permettraient de repérer automatiquement les cas où le *euh* est le lieu d'une rupture syntaxique. Dans cette phase de prétraitement, nous les laissons tels quels dans le corpus, sauf s'ils entrent dans une séquence répétition ou d'autocorrection, auxquels cas ils sont balisés dans ce cadre.

5.8 Les parenthèses : éléments para-verbaux et passages incompréhensibles⁵⁶

Les parenthèses notent une série d'indications « hors-texte », qui figurent dans la transcription mais qui n'appartiennent pas au discours des locuteurs à proprement parler. Ces informations peuvent être de types très divers. Elles sont des indications para-verbales sur le déroulement de la conversation (déplacements, actions ou gestes des locuteurs : *iljVM1 se rapproche du micro, iljNJI travaille à nouveau sur ordinateur* ; indication, dans les conversations à plus de deux locuteurs, de l'interlocuteur : *à souGJ2* ; etc.), sur la manière de parler du locuteur (*en riant, en chuchotant, chuchoté, ileGF0 s'éclaircit la voix*, etc.) ou encore sur des événements extérieurs à la conversation (*cris d'enfant, bruits de trafic, aboiements*, etc.).

En fait, on constate que peu de systématisme existe sur le type de phénomènes à transcrire, et sur la manière d'insérer ces indications dans le corps du texte. Le problème principal en ce qui nous concerne est que très souvent, on peut difficilement déterminer à la lecture du texte transcrit si ces indications sont ajoutées, à côté la parole du locuteur, ou si elles se superposent à celles-ci, en les masquant. Ainsi, il est évident que les indications (*interruption de l'enregistrement*) ou encore (*xx*) qui note une portion de parole incompréhensible, et par conséquent non transcrite, marquent clairement une rupture – et donc sans doute également une rupture syntaxique – entre ce qui précède et ce qui suit dans la transcription. Mais avec des indications telles que (*rire*) ou (*toux*), par exemple, il est parfois difficile de déterminer si l'énoncé se poursuit normalement, comme si l'indication n'était pas présente (dans ce cas, celle-ci serait un simple ajout aux données langagières, qui ne perturberait pas le déroulement syntagmatique de celles-ci)⁵⁷.

À l'analyse des occurrences de parenthèses dans notre corpus, et plus largement dans la banque de données VALIBEL, nous pouvons considérer que ces indications constituent, dans la très grande majorité des cas, une marque de rupture dans l'énoncé.

⁵⁶ Ne sont pas concernés ici les *silences*, eux aussi notés entre parenthèses.

⁵⁷ On objectera à ce problème que toute étude sur la langue parlée nécessite le recours aux données primaires, c'est-à-dire au son, et qu'on ne peut se contenter d'un travail qui analyse l'oral uniquement sur base du texte transcrit. Sans remettre en cause cette position de bon sens, on voit que cela est impossible pour une tâche comme la nôtre. Cela étant, les problèmes de non-homogénéisation que nous avons rencontrés dans le cadre de ce travail ont été l'occasion pour l'équipe VALIBEL de revoir de manière rigoureuse les conventions de transcription sur ces points qui pouvaient a priori sembler mineurs au regard d'autres points.

5.9 Conclusions sur les marques

Les marques que nous avons détaillées ci-dessus impliquent deux conséquences différentes sur la linéarité du texte : certaines indiquent une coupure dans la linéarité syntaxique de l'énoncé (les marques entre parenthèses, les crochets de notation phonétique entre balises, les balises de l'alternance de code) ; d'autres, si elles indiquent une coupure dans la linéarité de la lecture du texte, ne marquent pas une rupture de la construction syntaxique de l'énoncé (phonétique entre crochets sans les balises, les répétitions et autocorrections, les amorces que nous avons balisées).

Nous allons maintenant voir comment nous traitons ces deux grands cas de figure, ainsi que les pauses, grâce à des graphes construits avec UNITEX.

6 LE PRÉTRAITEMENT DU CORPUS AVEC UNITEX

Toutes les marques que nous venons d'aborder dans ce chapitre, qu'elles soient présentes dans les données de départ (les pauses) ou ajoutées dans une première phase de prétraitement (les balises des disfluences) demandent à être traitées avant l'application des ressources lexicales (chap. 5) et des grammaires de levée d'ambiguïtés (chap. 6). Nous avons formalisé les conclusions auxquelles nous avons abouti lors de l'analyse de ces marques, et développées ici, pour le prétraitement.

Le prétraitement avec UNITEX va concerner deux types de modifications apportées au texte : l'insertion de balises *{S}* afin de scinder le texte en unités plus petites ; l'insertion d'étiquettes pour certaines formes ou certaines portions de textes, notamment l'étiquette *<IGN>* pour les segments à ignorer lors de l'analyse morphosyntaxique.

Les modifications apportées au texte sont décrites dans des grammaires locales représentées sous forme de graphes.

6.1 Les graphes de prétraitement

Les graphes construits avec UNITEX⁵⁸ sont basés sur le formalisme des automates à état fini (Roche et Schabès 1997, Silberztein 1993, 1997 et 1998a, Paumier 2003 et 2006).

Un graphe est une représentation graphique d'un automate. Il contient des nœuds (ou boîtes), dont le nœud initial (symbolisé ci-dessous par une flèche) et le nœud terminal (symbolisé par un carré dans un rond), et des

⁵⁸ UNITEX possède un éditeur de graphes.

transitions. Une séquence est reconnue par un automate si, en partant du nœud initial, on arrive au nœud terminal en ayant lu toute la séquence. Un graphe se lit donc de gauche à droite. Prenons l'exemple de l'automate suivant :

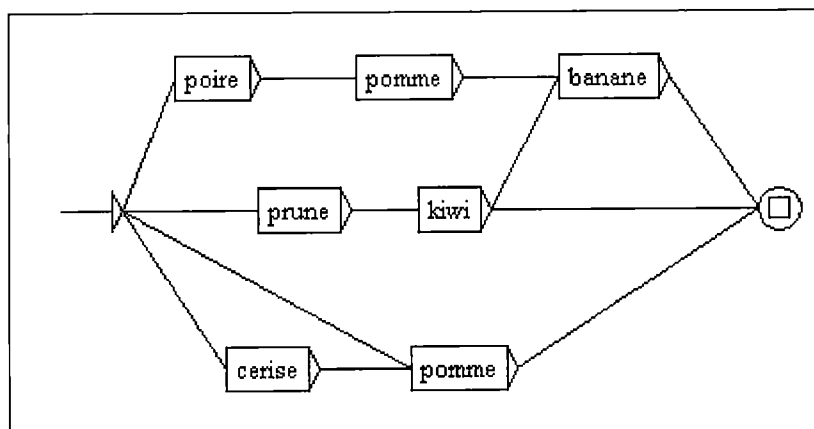


Figure 12. Exemple d'automate

Cet automate reconnaît les séquences

- *poire pomme banane* ;
- *prune kiwi* ;
- *prune kiwi banane* ;
- *cerise pomme* ;
- *pomme*.

Toutes ces séquences permettent en effet d'aboutir au nœud terminal en ayant parcouru un chemin dans l'automate. Par contre, il ne reconnaît pas des séquences telles que *cerise pomme banane* ou *poire pomme kiwi* qui ne sont pas décrites par un chemin. Il ne reconnaît pas non plus *poire pomme* puisque cette séquence, si elle est pourtant représentée dans l'automate, ne permet pas d'arriver à l'état final (elle ne concerne qu'une partie du chemin).

Un transducteur est un automate à deux bandes : une bande de lecture qui reconnaît une séquence (représentée dans les boîtes), et une bande d'écriture qui associe à cette séquence préalablement reconnue les indications présentes sous les boîtes.

Prenons l'exemple simple de la forme *euh*, à laquelle nous souhaitons associer une étiquette⁵⁹ de catégorie grammaticale (*IGN*) et de sous-catégorie (*Euh*) :

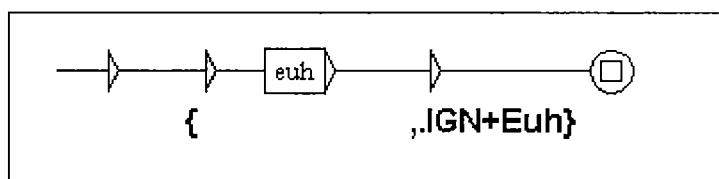


Figure 13. Étiquetage du euh

Notre graphe reconnaît dans le texte la forme indiquée dans la boîte : *euh*. Il associe à cette forme dans le texte les indications présentes sous les boîtes :

- avant le *euh* : une accolade ouvrante ;
- après le *euh* : une virgule ; un point ; le symbole *IGN* qui correspond à la catégorie *Ignore* ; le signe + ; la sous-catégorie *Euh* ; une accolade fermante.

Après application de ce graphe, la portion de texte

ileFN1 bon mon frère par exemple a toujours détesté
euh lire (...) [ileFN1r]

devient

ileFN1 bon mon frère par exemple a toujours détesté
{euh,.IGN+Euh} lire (...) [ileFN1r]

Le but du prétraitement est d'ajouter des informations au texte que l'on pourra prendre en compte lors de la construction de grammaires de levée d'ambiguïtés. Dans le cas présent, nous ajoutons le code *IGN* : celui-ci étiquette des séquences qui seront ignorées dans l'analyse morphosyntaxique, un peu comme si elles étaient absentes du texte (à la différence qu'elles sont toujours bien présentes).

Dans la phase de prétraitement de nos données, nous appliquons successivement trois graphes au texte : *Sentence*, *Replace* et *Locate*. Ces trois graphes appellent eux-mêmes des sous-graphes.

⁵⁹ Le format des étiquettes est décrit en détail au chapitre 5.

6.2 Le graphe *Sentence*

Le graphe *Sentence* appelle deux sous-graphes : *phrase* et *baliseIGN*.

6.2.1 Le graphe *phrase*

Le graphe *phrase*⁶⁰ insère des balises $\{S\}$ avant et après certaines marques : le point d'interrogation, les marques d'alternance de code (sous-graphe *alternanceCode*), les codes des locuteurs (sous-graphe *baliseLocuteurChev*), les codes des enregistrements (sous-graphe *baliseEnregistrement*), les balises de son (sous-graphe *baliseSon*) et les amorces qui n'ont pas été balisées auparavant (sous-graphe *amorces*). Ces étiquettes $\{S\}$ sont insérées parce que les portions de texte dont il est ici question constituent soit une rupture syntaxique dans la linéarité de l'énoncé (le cas des amorces ou de l'alternance de code) soit une marque claire de début d'énoncé et donc de fin d'un autre énoncé (les balises qui séparent chaque enregistrement par exemple).

Alors que l'analyse que nous en avons faite ne donnait pas des résultats très tranchés, nous avons décidé de considérer la pause longue comme une marque qui isole des segments. Cette décision a été prise pour des raisons méthodologiques que nous développerons au chapitre 6.

Voici le graphe *Phrase*. Les boîtes grisées indiquent qu'il s'agit de sous-graphes.

⁶⁰ Nous reprenons ce nom par commodité, même s'il est entendu qu'après les développements que nous avons faits dans ce chapitre, il n'est pas question de parler de « phrases » mais d'unités de traitement.

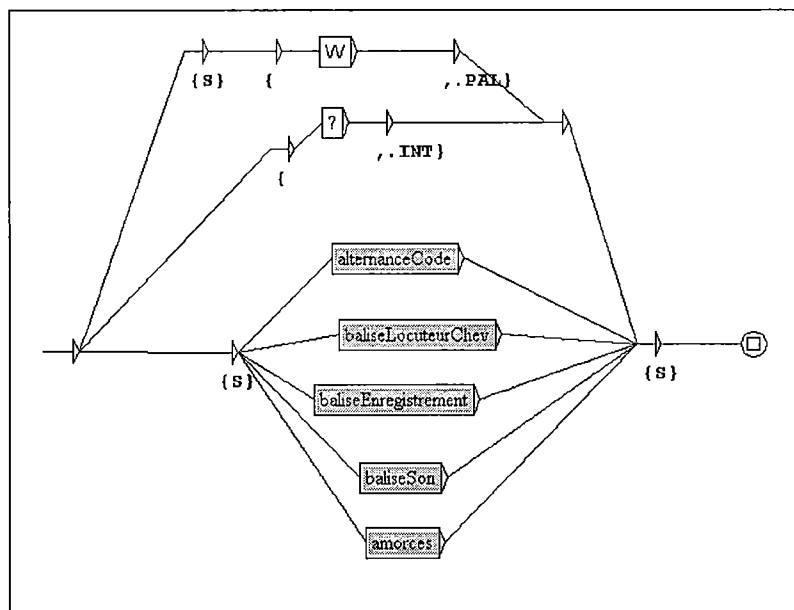
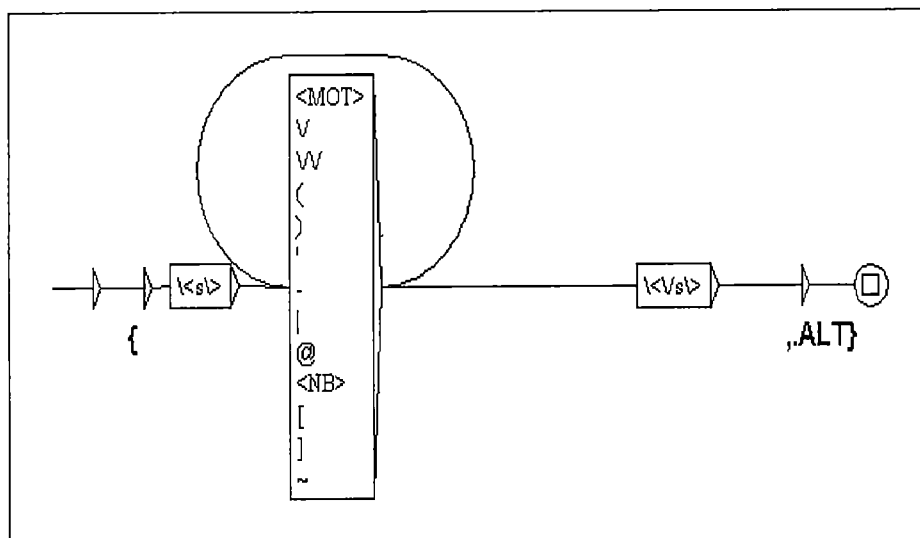


Figure 14. Graphe Phrase.grf

En fait, le traitement est double ici : non seulement on insère les signes $\{S\}$, mais en plus on insère une étiquette dans le texte, qui est fonction de la portion de texte concernée. La pause longue reçoit le code *PAL* (PAuse Longue) et le point d'interrogation le code *INT*. Les séquences d'alternance de code reçoivent l'étiquette *ALT*, comme le montre le sous-graphe suivant qui reconnaît toute la séquence encadrée par les balises ouvrantes ($<s>$) et fermantes ($</s>$ ⁶¹) d'alternance de code :

⁶¹ La barre oblique inversée ($\backslash$) protège certains caractères : dans le graphe *alternanceCode.grf* (figure 15), on protège les angles ($<$ et $>$) et la barre oblique ($/$). Ce qui ressemble à un W dans le graphe représenté en figure 14 est en fait la double barre oblique de la pause ($//$) protégée par les barres obliques inversées ($\backslash$), ce qui donne : $\backslash W$

Figure 15. Graphe *alternanceCode.grf*

L'intérêt est que toute la séquence ainsi reconnue est insérée dans une étiquette, ce qui évite lors des traitements ultérieurs de reconnaître chaque chaîne de caractères qui la compose comme devant être analysée individuellement.

6.2.2 Le graphe *baliseIGN*

Le graphe *baliseIGN* étiquette dans le texte des séquences qui doivent être ignorées lors de l'application des grammaires. Pour ce faire, on ajoute à ces segments reconnus l'étiquette *IGN* (pour « ignore »). L'idée est que ces marques ou ces séquences brisent « artificiellement » la linéarité de l'énoncé. En les supprimant (ou, en l'occurrence, en les ignorant), on a un texte qu'il est plus aisé de soumettre à l'étiquetage morphosyntaxique. Les phénomènes pris en considération sont les suivants : les marques de chevauchement, le *euh* et les balises de multitranscription pour la première solution proposée par le transcripateur (sous-graphe *simplesIGN*), la trace que nous laissons des chevauchements internes après extraction de ceux-ci (sous-graphe *LocIntIGN*), les balises que nous avons insérées pour les répétitions, certaines amorces, les multitranscriptions et les autocorrections (sous-graphe *balises*) et les indications de phonétique entre crochets (sous-graphe *phonetique*).

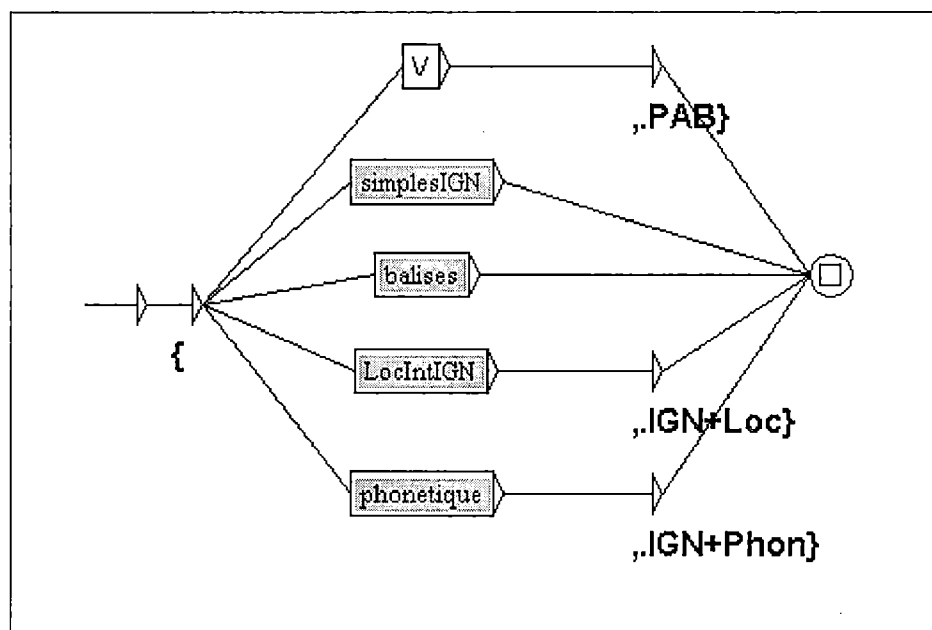


Figure 16. Graphe baliseIGN.grf

La pause brève est balisée *PAB* pour *PAuse Brève*.

C'est dans ce graphe *baliseIGN* que sont traités les cas de multitranscription. En effet, à l'heure actuelle, il n'y a pas de solution qui permette de tenir compte des multitranscriptions dans l'automate du texte, en les insérant dans des chemins parallèles par exemple. Ainsi, on pourrait imaginer que les trois solutions alternatives de l'exemple suivant soient représentées dans des chemins parallèles de l'automate :

famRM1 <m1> ben c'est ça </m1> <m2> bah c'est
sûr </m2> <m3> pas de souci </m3> [famRM1r]

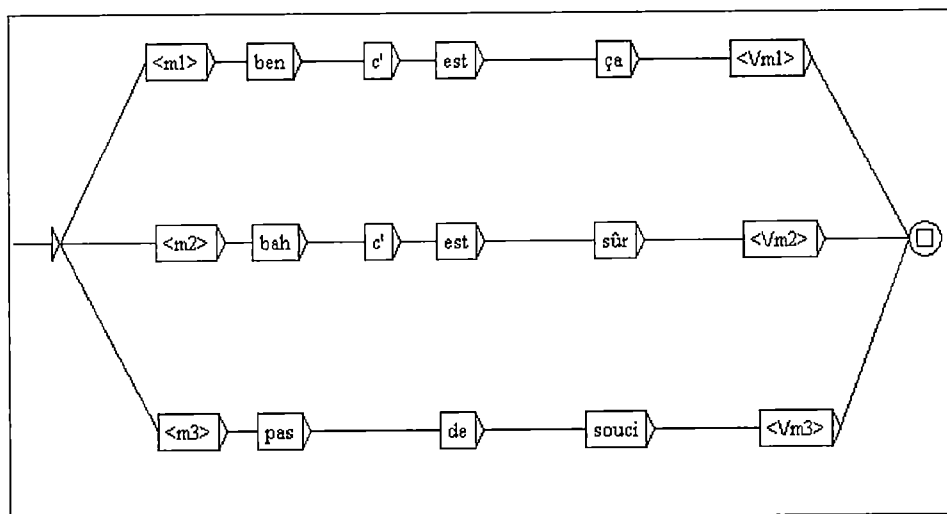


Figure 17. Automate idéal des multitranscriptions

Les chemins parallèles indiqueraient des hypothèses concurrentes, comme c'est le cas pour les hypothèses de catégories grammaticales lorsqu'il y a ambiguïté (cf. chap. 6).

Cette possibilité n'est pas offerte actuellement par UNITEX. Quand une seule solution est proposée par le transcrip-teur, il n'y a pas de problème et nous étiquetons simplement les balises *<m>* et *</m>* par l'étiquette *IGN* (sous-graphe *simplesIGN*). Ainsi, seule la portion de texte, même si elle est soumise à l'hésitation du transcrip-teur, est analysée par les grammaires. Néanmoins, la trace de l'hésitation du transcrip-teur n'est pas perdue puisque l'on conserve les balises *<m>* et *</m>*. Quand deux solutions concurrentes sont proposées dans la transcription, nous avons décidé de ne garder que la première, qui est la plus plausible aux yeux du transcrip-teur. Nous étiquetons donc les marques *<m1>* et *</m1>* par le code *IGN* (sous-graphe *simplesIGN*), ainsi que toute la séquence balisée par *<m2>* et *</m2>* (sous-graphe *balises*), qui ne sera donc pas étiquetée morphosyntaxiquement. On garde néanmoins ces séquences dans le texte, mais elles devront être étiquetées manuellement lors de la correction de l'étiquetage.

Voici le sous-graphe *balises* qui étiquette les portions du texte balisées lors du premier prétraitement comme des répétitions, des amorces, des autocorrections et des multitranscriptions. Chacun de ces segments est reconnu par le graphe qui insère respectivement les étiquettes suivantes : *IGN+Rep*, *IGN+Amo*, *IGN+Cor* et *IGN+Mult*.

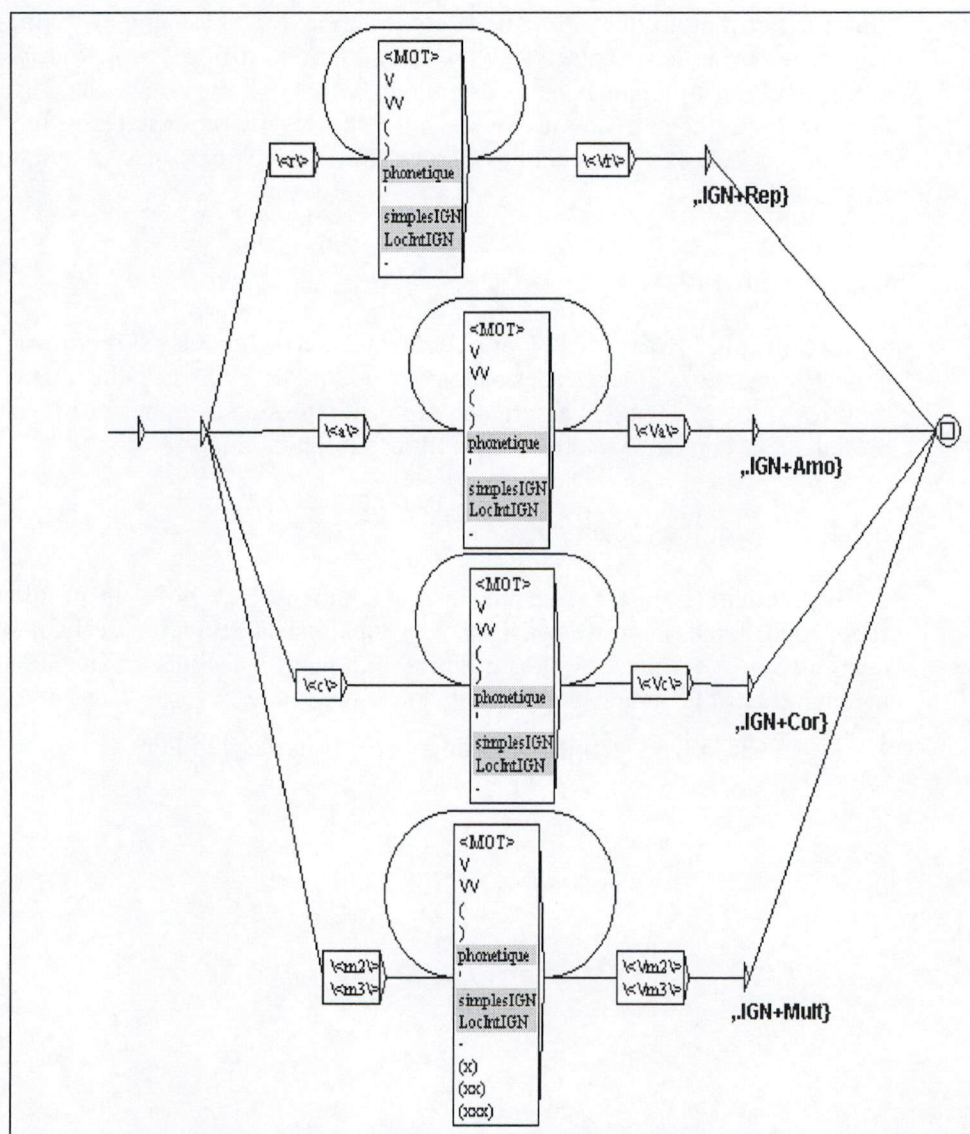


Figure 18. Graphe balises.grf

Suite à l'application du graphe *phrase*, il arrive que deux symboles de fin de phrase se suivent, comme dans l'exemple suivant, où les symboles insérés après le point d'interrogation et avant la pause longue cooccurrent :

{S} {accCT0,.NLoc} {S} ça va (toux) / tu m'as dit
l'accent c'est quelque chose qui atteint tout le monde ?
{S} {S} // {S} {|-,.IGN} {S} (x) {S}

Le but de l'insertion de ces symboles est d'isoler des unités plus petites que le texte entier. Or ici, ce qu'on isole entre deux segments, c'est un chemin qui sera vide dans l'automate du texte (cf. chap. 6). Dans ces cas,

afin d'éviter tout problème lors de la construction de l'automate, on applique le graphe en mode « replace » (*Replace.grf*). À la différence des graphes précédents qui ajoutaient les informations sous les boîtes au texte, ici, on modifie le segment reconnu par ce qui est sous les boîtes. Dans le cas présent, on remplace plusieurs symboles $\{S\}$ contigus par un seul symbole $\{S\}^{62}$.

6.3 Le graphe *LocateOral*

Le graphe *LocateOral*⁶³ appelle trois sous-graphes, *Parentheses*, *NomsPropres* et *Locuteurs*. Ces sous-graphes sont un peu particuliers dans la mesure où ils utilisent des fonctionnalités différentes des graphes vus jusqu'à présents : les contextes interdits et les filtres morphologiques.

6.3.1 Le graphe *Parentheses*

La reconnaissance des séquences entre parenthèses pose un problème spécifique dans la mesure où l'on ne souhaite pas étiqueter de la même manière les parenthèses selon qu'elles indiquent un silence, un passage incompréhensible ou une marque d'un autre type.

Le graphe suivant permet de distinguer ces trois cas de figure :

⁶² C'est en fait un peu plus compliqué parce qu'il faut tenir compte de la fin du paragraphe également.

⁶³ Nous appelons ce graphe *LocateOral* parce qu'il est appliqué avec la fonction *Locate* et non dans la fenêtre de preprocessing d'UNITEX. En fait, les graphes morphologiques et/ou contenant des contextes interdits ne peuvent être appliqués lors de la phase de prétraitement telle que décrite dans UNITEX (voir Paumier 2006). Nous appliquons donc ces automates avec le programme *Locate* en mode *merge*, sauvons le texte modifié et réappliquons le graphe *Replace*, puisque *LocateOral* a de nouveau réinséré des symboles $\{S\}$ qui sont parfois contigus à des symboles insérés précédemment.

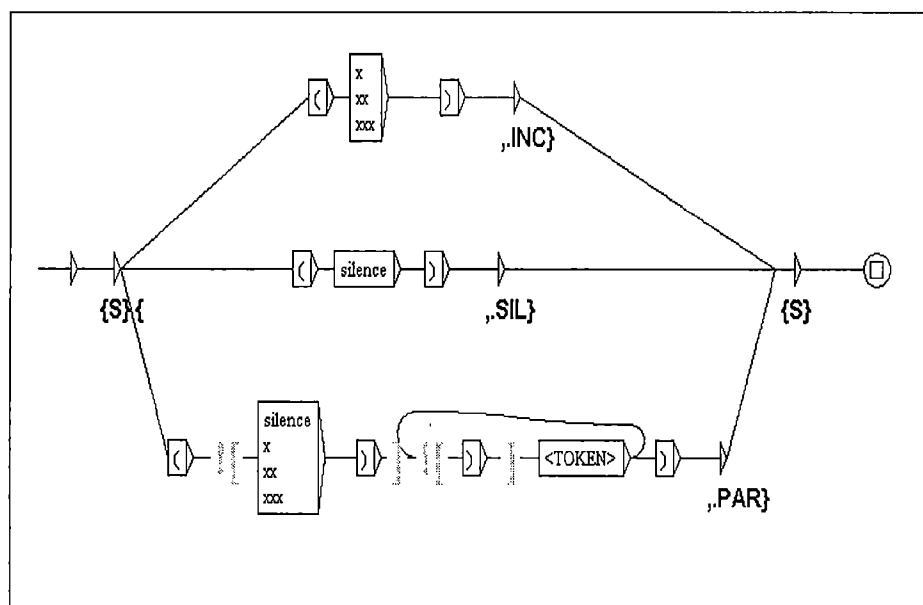


Figure 19. Graphe Parentheses.grf

Le premier chemin étiquette *INC* (« incompréhensible ») les séquences avec des *x* dans les parenthèses ; le deuxième chemin indique *SIL* pour les silences. Quant au troisième chemin, il interdit les deux contextes précédemment décrits pour que l'automate puisse s'appliquer au texte. Il reconnaît donc tout ce qui est dans les parenthèses hormis les silences et les *x*, et étiquette la séquence ainsi reconnue par *PAR* (pour « paraverbal »)⁶⁴.

Ici, en même temps que l'on balise la séquence reconnue par des accolades et un code, on insère les symboles de fin d'unité *{S}*.

6.3.2 Le graphe *NomsPropres*

Étant données les conventions de transcription appliquées à notre corpus, on peut poser l'hypothèse suivante quant aux noms propres : « Tout mot commençant par une majuscule est un nom propre ». En fait, cette hypothèse est correcte, sauf qu'elle inclut les lettres de l'alphabet parmi les noms propres, puisque celles-ci sont également notées en majuscules.

S'ils commencent tous les deux par une majuscule, lettre et nom propre peuvent néanmoins être formellement différenciés : alors que la lettre est une suite d'un seul caractère en majuscule, le nom propre contient au moins deux caractères. Ce type de conditions peut être décrit dans des graphes que l'on

⁶⁴ Pour plus de détails sur les contextes interdits, voir Paumier (2006 : 95 et sv.).

appelle *graphes morphologiques* (cf. Paumier 2006 : 57 et sv.)⁶⁵. Voici notre graphe :

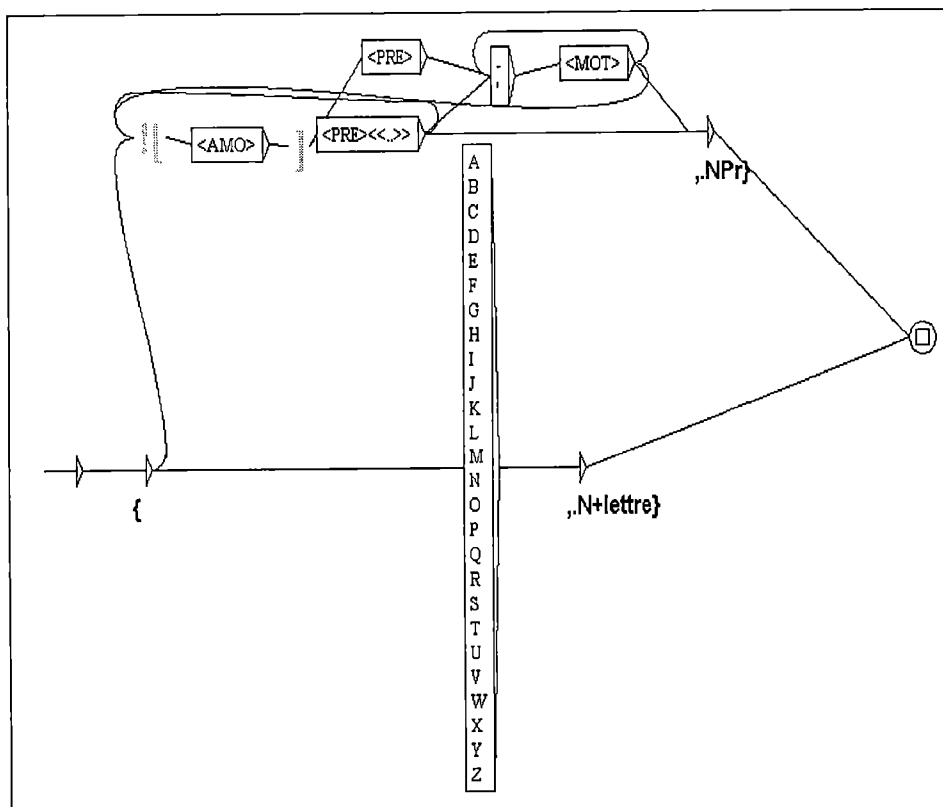


Figure 20. Graphe NomsPropres.grf

La condition pour reconnaître un nom propre est formalisée comme ceci : il s'agit d'un mot qui commence par une majuscule (<PRE>) **ET** qui comporte au moins deux lettres (symbolisées par les deux points de : <.>).

Il est des cas où le nom propre commence par une seule lettre ; il est alors obligatoirement suivi par un trait d'union ou une apostrophe, comme dans le titre *A-l'est-d'Eden*.

On voit que l'on utilise aussi ici les contextes interdits. En effet, on veut éviter de mettre l'étiquette *Npr* à des formes déjà étiquetées *AMO* par le

⁶⁵ À l'heure actuelle, ceux-ci ne peuvent être appliqués au prétraitement d'UNITEX. Nous les appliquons donc grâce au menu Locate, mais nous considérons qu'il s'agit de notre point de vue toujours de la phase de prétraitement de notre corpus.

graphe *amorces*. Sans ce contexte interdit, ce serait le cas des amorces de morphèmes de noms propres comme dans l'exemple suivant :

blaJD1 (...) c'est le docteur qui a téléphoné alors qu'on
vienn le chercher d'urgence // pour aller au F/ à Furnes
(...) [blaJD1]

Nous avons préféré garder l'étiquette AMO, qui est plus pertinente pour l'application des grammaires locales.

Les noms propres reçoivent le code *NPr* et les lettres de l'alphabet le code *N+lettre*. Ce graphe évite aussi de reconnaître des séquences de lettres contigües séparées par des blancs comme un nom propre⁶⁶.

6.3.3 Le graphe *Locuteurs*

Nous avons déjà étiqueté grâce aux graphes présentés ci-dessus les locuteurs selon qu'il s'agit de code de locuteurs extraits des chevauchements (*NlocChev*) ou de la trace gardée dans l'énoncé de la place du chevauchement (*IGN+Loc*).

On balise ici au moyen d'un graphe morphologique le code des locuteurs, qui est décrit très précisément comme commençant par trois lettres minuscules, suivies de deux lettres majuscules (on force le respect de la casse avec *_f_*) et directement d'un chiffre. On balise la séquence par l'étiquette *Nloc*, et l'on insère les signes *{S}* avant et après l'étiquette :

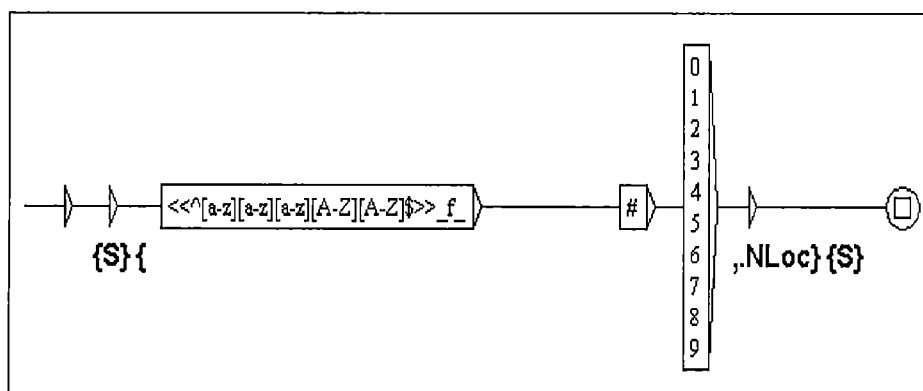


Figure 21. Graphe Locuteurs.grf

⁶⁶ Nous reviendrons sur les noms propres au chapitre 5 lorsque nous développons nos choix linguistiques.

7 DU TEXTE « BRUT » AU TEXTE PRÉTRAITÉ

Nous présentons maintenant un court extrait de notre corpus, dans trois versions différentes. La première version est la version originelle des transcriptions de VALIBEL ; la deuxième est celle que l'on obtient après le premier prétraitement, qui concerne principalement la normalisation des conventions, la codage des locuteurs, la gestion des chevauchements internes et le balisage des disfluences ; la troisième version est la version prête à être soumise à l'étiquetage morphosyntaxique.

version 1 (version originale)

L1 à la mémoire à la à la à la en tête mais enfin i-y-en-a certainement d'autres mais enfin ce sont des choses en tous les cas où je sais que lorsque je suis en France on me dit tiens / à-tantôt mais non on ne dit pas à-tantôt chez nous ça c'est c'est c'est pas ça et qui sont justifiées parce-que ça nE: bon si je dis septante et nonante à-la-place-de quatre-vingt-dix ou soixante-dix j'estime que je suis plus près-de la réalité grammaticale et vocabulaire que la France mais euh ça c'est une (rire) c'est une question grammaticale par-contre si je dis une aubette je peux comprendre qu'i ne comprennent pas une trémie euh c'est pas tout-à-fait le sens non plus si je dis à-tantôt / bè pour moi à-tantôt c'est c'est tout-à-l'heurE: en France ça c'est une c'est l'après-midi |- c'est <L0> mm -| c'est une chose donc voilà quelques expressions

L0 une une trémie / ça veut dire quoi

L1 une trémie justement une trémie i/ |- c'est une <L0> oui -| un tunnel une trémie chez nous c'est: / c'est le c'est c'est ce qu'on appelle un tunnel

L0 ah d'accord

L1 hein |- mais <L0> mm -| une pet/ un petit tunnel qui n'est pas très long

L0 mm

L1 or une trémie euh grammaticalement c' est une chose qui s' en/ qui s' enfonce plutôt dans la terre

L0 mm

version 2 (après le premier prétraitement)

ilePA2 à la mémoire <r> à la à la </r> à la en tête mais enfin il y en a certainement d'autres mais enfin ce sont des choses en tous les cas où je sais que lorsque je suis en France on me dit tiens / à tantôt mais non on ne dit pas à tantôt chez nous ça <r> c'est c'est </r> c'est pas ça et qui sont justifiées parce que ça ne bon si je dis septante et nonante à la place de quatre-vingt-dix ou soixante-dix j'estime que je suis plus près de la réalité grammaticale et vocabulaire que la France mais euh ça

<r> c'est une (rire) </r> c'est une question grammaticale par contre si je dis une aubette je peux comprendre qu'ils ne comprennent pas une trémie euh c'est pas tout à fait le sens non plus si je dis à tantôt / bè pour moi à tantôt <r> c'est </r> c'est tout à l'heure en France ça c'est une c'est l'après-midi |- <r> c'est @767 -| </r> c'est une chose donc voilà quelques expressions

@767 ileGF0 mm

ileGF0 <r> une </r> une trémie / ça veut dire quoi

ilePA2 une trémie justement une trémie i/ |- c'est <c> une @768 -| </c> un tunnel une trémie chez nous <r> c'est / </r> c'est le <r> c'est </r> c'est ce qu'on appelle un tunnel

@768 ileGF0 oui

ileGF0 ah d'accord

ilePA2 hein |- mais @769 -| <c> une pet/ </c> un petit tunnel qui n'est pas très long

@769 ileGF0 mm

ileGF0 mm

ilePA2 or une trémie euh grammaticalement c'est une chose <r> qui s'en/ </r> qui s'enfonce plutôt dans la terre

Version 3 (en fin de prétraitement, version prête pour l'annotation morphosyntaxique)

{S} {ilePA2,.NLoc} {S} à la mémoire {<r> à la à la </r>,.IGN+Rep} à la en tête mais enfin il y en a certainement d'autres mais enfin ce sont des choses en tous les cas où je sais que lorsque je suis en {France,.NPr} on me dit tiens {/,PAB} à tantôt mais non on ne dit pas à tantôt chez nous ça {<r> c'est c'est </r>,.IGN+Rep} c'est pas ça et qui sont justifiées parce que ça ne bon si je dis septante et nonante à la place de quatre-vingt-dix ou soixante-dix j'estime que je suis plus près de la réalité grammaticale et vocabulaire que la {France,.NPr} mais {euh,.IGN+Euh} ça {<r> c'est une (rire) </r>,.IGN+Rep} c'est une question grammaticale par contre si je dis une aubette je peux comprendre qu'ils ne comprennent pas une trémie {euh,.IGN+Euh} c'est pas tout à fait le sens non plus si je dis à tantôt {/,PAB} bè pour moi à tantôt {<r> c'est </r>,.IGN+Rep} c'est tout à l'heure en {France,.NPr} ça c'est une c'est l'après-midi {|-,.IGN+ChO} {<r> c'est @767 -| </r>,.IGN+Rep} c'est une chose donc voilà quelques expressions

{S} {@767 ileGF0,.NLocChev} {S} mm

{S} {ileGF0,.NLoc} {S} {<r> une </r>,.IGN+Rep} une
 trémie {/, .PAB} ça veut dire quoi
 {S} {ilePA2,.NLoc} {S} une trémie justement une trémie
 {S} {i/, .AMO} {S} {|-,.IGN+ChO} c'est {<c> une @768
 -| </c>,.IGN+Cor} un tunnel une trémie chez nous
 {<r> c'est / </r>,.IGN+Rep} c'est le {<r> c'est
 </r>,.IGN+Rep} c'est ce qu'on appelle un tunnel
 {S} {@768 ileGF0,.NLocChev} {S} oui
 {S} {ileGF0,.NLoc} {S} ah d'accord
 {S} {ilePA2,.NLoc} {S} hein {|-,.IGN+ChO} mais
 {@769,.IGN+Loc} {-|,.IGN+ChF} {<c> une pet/
 </c>,.IGN+Cor} un petit tunnel qui n'est pas très long
 {S}{@769 ileGF0,.NLocChev}{S} mm
 {S} {ileGF0,.NLoc} {S} mm
 {S} {ilePA2,.NLoc} {S} or une trémie {euh,.IGN+Euh}
 grammaticalement c'est une chose {<r> qui s'en/
 </r>,.IGN+Rep} qui s'enfonce plutôt dans la terre

Le prétraitement des textes que nous venons de détailler dans ce chapitre se fait en vue d'un objectif précis : l'étiquetage morphosyntaxique du corpus. Nous aborderons cet étiquetage dans notre dernier chapitre, mais auparavant, nous allons détailler au chapitre 5 les ressources lexicales dont nous disposons, ainsi que nos choix linguistiques.

CHAPITRE 5

LES RESSOURCES LINGUISTIQUES

Sans le savoir, Jojo s'exprimait dans le jargon en vogue cette année-là sur les campus, le patois fuck. Le mot pouvait être une simple interjection (« Fuck ! » ou « Fuck... ») en réaction à une surprise désagréable, une épithète de dénigrement (« fucking baobab »), un adverbe venant intensifier un adjectif (« c'est pourtant fucking clair ! »), un substantif (« ce grand fuck de merde »), un verbe qui, dans ses formes modifiées, pouvait signifier « Tire-toi de là », ou « foutre sur la gueule de quelqu'un », ou « rater », ou « se pinter », et, à la forme impérative, une formule de mépris péremptoire : « Fuck you ! ».

(*Moi, Charlotte Simmons*, Tom Wolfe, Paris, Robert Laffont, 2006, p. 40)

1 INTRODUCTION

Le système UNITEK (Paumier 2003 et 2006) que nous utilisons pour l'annotation morphosyntaxique est étroitement lié à des ressources linguistiques qui sont consignées dans des dictionnaires électroniques. En effet, nous le développerons davantage au chapitre 6, UNITEK fournit un étiquetage a priori des formes du texte. Cet étiquetage est basé sur des données linguistiques recensées dans des dictionnaires électroniques. Une forme sera donc considérée comme ambiguë si elle reçoit plusieurs analyses dans le dictionnaire.

Dans ce chapitre, nous allons donc détailler ces ressources linguistiques, aux sources du système de levée d'ambiguïtés.

2 DICTIONNAIRES PAPIER, DICTIONNAIRES SUR SUPPORT ÉLECTRONIQUE ET DICTIONNAIRES ÉLECTRONIQUES

Les ressources lexicales dont nous disposons sont organisées en *dictionnaires électroniques*. D'emblée, il nous faut préciser que ces dictionnaires électroniques ont peu de rapport avec les dictionnaires disponibles actuellement sur support informatique. En effet, des divergences fondamentales existent : alors que les dictionnaires électroniques adoptent un codage strict de leurs données, de manière à ce que celles-ci puissent être utilisées par des programmes informatiques, les informations des dictionnaires sur support informatique sont semblables à celles des dictionnaires papier, seules changent les modalités de consultation¹. En d'autres termes, il s'agit de dictionnaires destinés à être manipulés par des humains et dont les informations, de ce fait, n'ont pas toujours besoin d'être explicites. Ainsi les mots vedettes suivants (noms et/ou adjectifs) issus du *Petit Robert* (2000) :

doux, douce
froid, froide
mongol, e
fortifiant, iante
égal, ale, aux

Pour chaque entrée, on a l'indication de la forme au féminin : pour *douce* et *froide*, la forme complète figure en entier dans l'entrée, tandis que *mongole*, *fortifiante* et *égale* doivent être déduites de la forme au masculin. Pourtant, pour obtenir le féminin, dans les cinq cas, seul un *-e* doit être ajouté à la finale. C'est d'ailleurs le choix de notation posé pour *mongole*.

Dans les deux derniers cas, il est entendu que par cette présentation, le lecteur humain comprend que ce qui est noté avant la virgule est la forme au masculin, et que pour obtenir le féminin, il substitue le *iante* au *iant* du masculin (ou le *ale* au *al*). Il s'agit en fait d'effectuer la substitution à partir de la (ou des) lettre(s) commune(s). Il comprend aussi que dans les deux premiers cas, c'est la substitution de la forme entière qui doit être faite, là où pour *mongole*, il ajoute simplement, sans substitution.

De plus, on voit que pour *égal*, on a l'indication d'une forme au pluriel (cette indication prenant seulement la forme de la finale *aux*), alors que cette

¹ On pense notamment à celles induites par l'ordre alphabétique, convention la plus habituellement suivie comme moyen d'entrée dans un dictionnaire sur support papier.

Notons tout de même une autre divergence entre les dictionnaires usuels selon qu'ils sont sur support papier ou sur support informatique : les limites de taille que connaissent les dictionnaires papier – limites liées notamment au cout – peuvent être dépassées dans les dictionnaires sur support électronique.

indication est absente des autres entrées reprises ici, et notamment de *doux*, où l'on pourrait estimer qu'elle aurait aussi pu figurer.

2.1 Systématisation de l'information donnée

Non seulement les informations fournies dans les dictionnaires usuels ne sont pas explicites, mais elles ne sont pas non plus systématiques. On vient de le voir notamment ci-dessus pour l'indication du pluriel.

Maurice Gross (1989 : 6-7) a choisi dans le *Lexis* certains noms de profession pour lesquels il montre que les informations données dans les articles sont d'ordres très différents : parfois le terme est défini par un synonyme, parfois par une définition complète ; parfois on a l'indication textuelle qu'il s'agit d'une profession, parfois il faut deviner cette information à travers l'expression verbale *faire commerce de*, etc. Ainsi, pas de systématisation dans l'information donnée et seul un lecteur humain a la capacité de déceler dans ces divers articles le signifié pourtant identique « nom de profession ».

Toujours en ce qui concerne les noms de profession, mais du point de vue de la morphosyntaxe cette fois, Louise-L. Larivière (2004) a montré l'incohérence de plusieurs dictionnaires usuels en ce qui concerne la mention des formes au féminin de ces noms : parfois absentes, parfois dans le même article (et là aussi, la manière de les présenter diverge fortement d'un terme à l'autre), parfois dans une entrée séparée (et dans ce cas, soit avec lien, soit sans lien avec la forme au masculin), etc.

On pourrait multiplier les exemples – le nom des animaux dans le *Petit Robert* (2000), tant du point de vue sémantique que des indications de morphosyntaxe –, qu'ils relèvent de la sémantique, de la syntaxe ou de la morphosyntaxe : les informations fournies par les dictionnaires usuels sont loin de l'être de façon explicite, homogène et systématique. Elles sont en tout cas, présentées de la sorte, inexploitable dans des traitements automatisés des langues naturelles².

² Des tentatives ont vu le jour, qui consistaient à extraire des bandes de photocomposition des versions papier des dictionnaires courants certaines informations aisément formalisables. Voir notamment le projet DÉFI (Désambiguïsation par filtrage), ARC 95/00-192 de l'Université de Liège (1995-2000). On voit aussi la création de réseaux sémantiques à partir de dictionnaires explicatifs, dans le cadre d'exploitations statistiques (Kozima et Furugori 1993, Senellart et Blondel 2003).

2.2 Nombre d'entrées

Dictionnaires électroniques et dictionnaires sur support électronique diffèrent également par le nombre de mots qu'ils répertorient³.

Ainsi, le principal dictionnaire électronique que nous utilisons pour les mots simples contient 682 457 entrées. Ce nombre très important d'entrées est dû au fait que, à la différence des dictionnaires traditionnels, celui-ci recense toutes les formes fléchies des mots, et pas uniquement ce que l'on nomme classiquement le *lemme* (ou encore *forme canonique*, *forme de base*, *base* ou *mot-vedette*).

Si cette manière de présenter les données n'a pas d'impact pour des catégories grammaticales telles que les conjonctions ou encore les adverbes, dont la forme est invariable (à l'exception de *tout*, cf. § 4.1.5), elle accroît considérablement le nombre d'entrées des noms et des adjectifs par exemple. Ainsi, quand le *Petit Robert* présente une seule entrée notée

ami, ie

un dictionnaire électronique se doit de recenser les quatre formes distinctes

ami (masculin singulier)

amis (masculin pluriel)

amie (féminin singulier)

amies (féminin pluriel).

Là où la différence entre le nombre d'entrées d'un dictionnaire classique et d'un dictionnaire informatique est la plus grande, c'est évidemment dans le cas des verbes, catégorie qui en français connaît un grand nombre de formes fléchies. Ainsi, l'exemple des verbes du 1^{er} groupe tels que *manger*, pour lesquels à la seule forme canonique à l'infinitif sont associées 38 formes fléchies (classées ici alphabétiquement) : *mange*, *mangé*, *mangea*, *mangeai*, *mangeaient*, *mangeais*, *mangeait*, *mangeâmes*, *mangeant*, *mangeas*, *mangeasse*, *mangeassent*, *mangeasses*, *mangeassiez*, *mangeassions*, *mangeât*, *mangeâtes*, *mangée*, *mangées*, *mangent*, *mangeons*, *mangera*, *mangerai*, *mangeraient*, *mangerais*, *mangerait*, *mangeras*, *mangèrent*, *mangerez*, *mangeriez*, *mangerions*, *mangerons*, *mangeront*, *manges*, *mangés*, *mangez*, *mangiez* et *mangions*.

On le voit, c'est principalement la flexion des noms, des adjectifs et des verbes qui augmente considérablement le nombre d'entrées d'un dictionnaire électronique.

³ Bien entendu, les dictionnaires traditionnels divergent également de ce point de vue (que l'on pense à l'introduction de néologismes ou à la distinction homonymie/polysémie).

3 LE DICTIONNAIRE ÉLECTRONIQUE DES FORMES FLÉCHIES DU FRANÇAIS : DELAF

Les dictionnaires électroniques que nous utilisons dans le cadre de cette recherche sont ceux qui ont été développés au LADL⁴ (Laboratoire d'Automatique Documentaire et Linguistique), sous la direction de Maurice Gross (Courtois 1994).

Le LADL a développé trois grands types de dictionnaires électroniques :

- **DELAP : dictionnaire électronique pour la phonémique**

Dans ce dictionnaire, à chaque forme simple sont associés une transcription phonémique et un code morphologique. Ces informations permettent d'engendrer automatiquement le DELAP-FV (F : fléchi, V : variante phonétique), dictionnaire de près de 600 000 entrées qui comportent chacune une transcription phonétique ainsi qu'une étiquette grammaticale.

- **DELAS : dictionnaire électronique des formes simples**

Dans le DELAS, chaque mot en entrée se voit attribuer un code morphologique qui permet d'obtenir de manière totalement automatique toutes les formes fléchies associées à ce mot. (cf. § 4.9.2). Il faut donc entendre ici par « formes simples » l'ensemble des lemmes.

- **DELAF : dictionnaire électronique des formes fléchies**

Le DELAF est le dictionnaire qui résulte de la flexion automatique du dictionnaire des formes simples. Il en existe deux différents, l'un pour les mots simples, l'autre pour les mots composés, au sens graphique du terme (Courtois 1990 et Silberstein 1993).

Ces dictionnaires sont des dictionnaires de la langue générale ; des dictionnaires spécialisés ont également été constitués : dictionnaire des noms propres, dictionnaire des professions, dictionnaire des sigles, etc.

C'est ce format de dictionnaire, où apparaissent les formes fléchies, que nous utilisons dans le cadre de l'étiquetage automatique de textes.

3.1 Principes de présentation des données

Voici les premières entrées du dictionnaire DELAF :

a,.N+z1:ms:mp
a,.XI+z1

⁴ Actuellement, toutes les ressources linguistiques du LADL continuent d'être développées à l'Institut Gaspard-Monge (Université de Marne-la-Vallée), sous la direction d'Éric Laporte.

a,avoir.V+z1:P3s
 à,.PREP+z1
 aa,.N+z3:ms:mp
 aabam,.N+z3:ms
 aalénien,.A+z3:ms
 aalénien,.N+z3:ms
 aalénienne,aalénien.A+z3:fs
 aaléniennes,aalénien.A+z3:fp
 aaléniens,aalénien.A+z3:mp

Comme nous l'avons mentionné ci-dessus, toutes les informations présentes dans un dictionnaire électronique doivent être codées explicitement, de manière univoque.

Dans le DELAF, une ligne constitue une entrée du dictionnaire, qui adopte le formalisme suivant :

forme,lemme.catégorie+couche_du_vocabulaire+informations_sémantiques_ou_syntaxiques:informations_flexionnelles/commentaires

Voici le détail des informations :

- *forme* : la forme telle qu'on peut la trouver dans un texte ;
- une virgule ;
- *lemme* : le lemme est la forme du lexème choisie conventionnellement et arbitrairement, et à laquelle vont se rattacher toutes les formes fléchies du lexème. Dans le DELAF, ce lemme correspond au mot-vedette d'un dictionnaire traditionnel, c'est-à-dire donc, *grosso modo*, à la forme au masculin singulier pour les noms et les adjectifs, à l'infinitif pour les verbes⁵ ;
- un point ;
- *catégorie grammaticale* : la catégorie grammaticale est la partie du discours (ou encore la classe grammaticale) associée à chaque lexème (cf. § 4) ;
- le signe + ;
- la couche du vocabulaire : elle indique si le mot est un mot courant de la langue, un mot rare ou s'il se situe entre ces deux pôles (Garrigues 1993) ;
- le signe du double point ;

⁵ Notons que, puisqu'il s'agit d'une convention, d'autres choix sont possibles. Ainsi, dans une étude sur la féminisation des noms de professions, titres et grades lors des campagnes électorales, nous avons choisi comme lemme la forme au féminin (Dister et Moreau 2006).

- selon la catégorie grammaticale, des informations flexionnelles : genre, nombre, temps, mode, etc.

La version du dictionnaire que nous avons choisie pour cette recherche est exempte de codes syntaxiques (+*t* pour les verbes transitifs, +*d* pour les adjectifs placés à droite du nom, par exemple) ou sémantiques (de type +*Hum* pour les noms humains, +*Conc* pour les noms concrets, etc.). Nous ne tiendrons pas non plus compte de la couche du vocabulaire.

Seules les informations notées en gras ci-dessus sont obligatoires. Ainsi, pour une catégorie comme l'adverbe ou la préposition, aucune information flexionnelle n'est présente (cf. § 4). Par ailleurs, lorsque la graphie du lemme est identique à celle de la forme en entrée, elle n'a pas besoin d'être notée. C'est le cas pour les noms au masculin singulier ou pour les verbes à l'infinitif puisque, nous l'avons dit, ce sont ces formes qui ont été choisies comme lemmes.

Ainsi, la première entrée du dictionnaire présentée ci-dessus est à lire de la manière suivante : la forme *a*, dont le lemme possède une graphie identique *a*, appartient à la catégorie des mots les plus fréquents (+z1) et est un nom masculin (m), au singulier (s) ou au pluriel (p).

La seconde entrée pour la forme *a* possède également un lemme dont la graphie est identique et appartient à la catégorie des constituants de composés non autonomes (*XI*), comme dans *a priori*.

Pour la troisième entrée, on constate que le lemme est différent de la forme, puisqu'il s'agit de *avoir*, verbe de la couche 1 du vocabulaire (+z1), conjugué à la 3^e personne (3) du singulier (s) de l'indicatif présent (P).

Le nombre de 682 457 entrées donné ci-dessus correspond à une présentation des données où chaque graphie (chaîne de caractères) est séparée en plusieurs formes distinctes (et donc en entrées distinctes) selon que varie : a) soit sa catégorie grammaticale (*devant*, nom ou préposition ; *les*, pronom ou déterminant) ; b) soit son lemme (*suis*, de *suivre* ou de *être* ; *vis* de *vivre*, de *vis* ou de *voir*). Ainsi, lorsqu'à une même graphie sont associées des analyses qui ne varient que par l'information flexionnelle (et non par le lemme et/ou la catégorie grammaticale), ces différentes analyses sont notées dans la même entrée.

Prenons l'exemple des formes verbales. Dans le dictionnaire tel que nous l'utilisons (Courtois 1990), toutes les informations sur la conjugaison d'une forme verbale sont regroupées sur la même ligne, constituant ainsi une seule entrée pour la forme⁶ :

donne,donner.V:P1s:P3s:S1s:S3s:Y2s

⁶ Nous parlons ici uniquement de la forme verbale. Selon le formalisme ici adopté, il y a une seconde entrée pour le nom *donne*.

Cette entrée se lit comme suit : *donne*, dont le lemme est *donner*, est un verbe à la première personne du singulier (1s) de l'indicatif présent (P), à la troisième personne du singulier (3s) de l'indicatif présent, à la première personne du singulier du subjonctif présent (S), à la troisième personne du singulier du subjonctif présent ou à la deuxième personne du singulier (2s) de l'impératif (Y).

Mais on aurait pu imaginer, lors de la construction du dictionnaire, de n'attribuer qu'une seule information de conjugaison pour chaque entrée. Ainsi, la forme *donne* serait associée aux cinq entrées suivantes⁷ :

donne,donner.V:P1s
 donne,donner.V:P3s
 donne,donner.V:S1s
 donne,donner.V:S3s
 donne,donner.V:Y2s

Selon ce principe de codage des entrées, tous les verbes du premier groupe verraient leurs entrées multipliées par cinq pour la terminaison *-e* ; et par deux pour les finales *-es*, *-ais* et *-ent*. Pour les verbes des autres groupes, les dédoublements seraient différents mais nombreux aussi.

Par ailleurs, la forme *donne* est associée au lemme *donner*, sans aucun autre type d'informations. Or, dans le cadre théorique du lexique-grammaire, il existe plusieurs verbes *donner*, qui ont chacun des comportements syntaxiques différents. Par exemple : *Max donne un lit à Ida* (table 36DT du lexique-grammaire) ; *la fenêtre donne sur la cour* (table 35ST) ; etc. La prise en compte de ces informations syntaxiques multiplie alors les entrées de *donne*, pour l'associer à 12 verbes *donner* différents.

Si en plus on assigne à chacun de ces verbes une seule entrée par information de conjugaison, comme nous l'avons fait ci-dessus, nous obtenons alors 5 X 12, soit 60 entrées pour la seule forme verbale *donne*.

Dans la mesure où nous nous intéresserons uniquement aux ambiguïtés morphologiques, nous ne tiendrons pas compte de ces ambiguïtés qui relèvent de la syntaxe.

Pour la flexion des adjectifs, différentes possibilités de codage se présentent également. Voici les entrées pour les quatre flexions du lemme <intelligent> :

intelligent,intelligent.A:ms
 intelligente,intelligent.A:fs

⁷ C'est d'ailleurs avec cette présentation — une ligne pour chaque information de conjugaison — qu'est construit le dictionnaire électronique des mots simples de l'italien (voir Vietri et Elia 2000).

intelligentes,intelligent.A:fp

intelligents,intelligent.A:mp

Quatre formes différentes qui sont assez logiquement réparties dans quatre entrées lexicales.

Pour <docile> par contre, on ne rencontre plus que deux formes, l'une pour le singulier l'autre pour le pluriel, aucune marque de genre n'apparaissant dans la graphie. Le DELAF a choisi d'en faire deux entrées et non quatre :

docile,docile.A:ms:fs

dociles,docile.A:mp:fp

De la même manière, l'adjectif *orange*, qui n'a lui aucune marque au pluriel, possède une seule entrée :

orange,orange.A:ms:fs:mp:fp

Ici aussi, on a choisi de rassembler toutes les informations morphologiques dans une seule entrée, et non d'engendrer quatre entrées différentes. Il en va donc de même pour un grand nombre d'adjectifs de couleur (*abricot*, *acajou*, *ambre*, *amarante*, etc.), mais aussi pour : *afro*, *antibruit*, *antichoc*, *rasoir*, *pop*, *puce*, *punk*, etc. En fait, tous les adjectifs invariables du dictionnaire DELAF sont codés de cette façon. Choisir la solution de codage qui démultiplie les entrées donnerait, au lieu des 12 352 entrées actuelles, 25 824 entrées.

Le DELAF, avec ses 682 457 entrées, constitue une ressource lexicale très complète pour l'analyse automatique de textes. Ses descriptions lexicales très fines mettent en évidence des analyses ignorées — volontairement ou non — de la plupart des systèmes de traitement automatique. Citons entre autres :

- *par*, qui n'est pas seulement la préposition mais également un nom (terme de golf).

par,par.N:ms

par,par.PREP

- *nouvelles* qui est quatre fois ambigu :

nouvelles,nouveau.A:fp

nouvelles,nouveau.N:fp

nouvelles,nouvelle.N:fp

nouvelles,nouvelles.N:fp

On voit comment le fait de tenir compte de ces entrées supplémentaires augmente le nombre d'ambiguïtés potentielles.

Dans le dictionnaire des mots simples DELAF, tel que nous venons de le décrire, plus de 12 % des entrées sont ambiguës, c'est-à-dire qu'elles sont homographes avec au moins une autre entrée (Courtois 1996).

Blandine Courtois (1996)⁸ s'est intéressée également aux ambiguïtés, dans le dictionnaire, en termes de catégorie grammaticale. Parmi les combinaisons d'ambiguïtés les plus fréquentes dans le DELAF, elle donne, par ordre décroissant : nom Humain / adjectif ; verbe au participe passé / adjectif ; nom masculin / adjectif ; verbe à l'indicatif présent / nom féminin ; nom féminin / adjectif ; verbe au participe passé / nom humain / adjectif ; verbe au participe présent / adjectif ; verbe à l'indicatif présent / nom masculin.

Lorsque nous parlons de texte désambiguïté dans ce travail, nous entendons par là qu'une information unique en termes de lemme, catégorie grammaticale et information flexionnelle a été donnée. Nous ne référons donc pas à une entrée du dictionnaire telle que présentée ci-dessus (où plusieurs informations flexionnelles peuvent être associées à un lemme).

3.2 Comparaison de nos ressources avec d'autres lexiques

Pionnières en la matière pour le domaine francophone, les ressources lexicales à large couverture du LADL ont été récemment rejointes par d'autres lexiques mis à la disposition de la communauté scientifique.

Le plus intéressant à notre sens est MORPHALOU⁹, lexique ouvert et évolutif des formes fléchies du français élaboré à l'ATILF (Analyse et Traitement Informatique de la Langue Française, anciennement INALF).

Le lexique qui compose MORPHALOU provient de la nomenclature du *Trésor de la Langue Française (TLFnome)* (Romary *et al.* 2004) qui a fourni 68 075 lemmes auxquels sont associées 539 413 formes fléchies. Il s'agit donc, comme le DELAF, d'un lexique à large couverture, amené à être progressivement augmenté au fil des contributions des usagers. Distribué depuis le début de l'année 2005, MORPHALOU est en accès libre à des fins de recherche et d'enseignement.

⁸ On trouvera encore dans cet article de nombreux exemples des différents types d'ambiguïtés, répartis dans les quatre grandes classes que l'auteure isole : ambiguïtés entre deux formes verbales ; ambiguïtés entre une forme verbale et non verbale ; ambiguïtés entre une forme nominale et une forme non verbale ; ambiguïtés restantes.

⁹ Morphalou est téléchargeable à l'adresse suivante : <http://actarus.atilf.fr/morphalou>

3.2.1 Parallèles avec le DELAF

Tout comme dans le DELAF, des lemmes homographes sont répartis dans des entrées lexicales séparées lorsqu'ils diffèrent 1) par leur catégorie grammaticale (*devoir* verbe versus *devoir* nom) ou 2) pour les noms, par le genre sans qu'il y ait dérivation suffixale (*la voile* versus *le voile*).

De plus, il s'agit également d'un lexique morphologique.

3.2.2 Divergences avec le DELAF

Nous pouvons dégager cinq grandes différences dans la présentation des données entre MORPHALOU et le DELAF. Les deux premières sont des divergences qui relèvent du format, de la notation générale et n'ont guère d'incidences théoriques ou méthodologiques, tandis que les trois dernières nous paraissent plus fondamentales.

1) La notion d'entrée lexicale diverge. Dans MORPHALOU, « entrée lexicale » a un sens plus large, puisqu'une entrée regroupe toutes les formes fléchies d'un même lemme. Ainsi, la composante /inflection/ apparaît quatre fois pour *portatif*, *portatifs*, *portative* et *portatives* dans l'entrée lexicale de l'adjectif *portatif*.

Dans le DELAF, on dira qu'il s'agit de quatre entrées distinctes :

```
portatif,.A+z1:ms
portatifs,portatif.A+z1:mp
portative,portatif.A+z1:fs
portatives,portatif.A+z1:fp
```

2) Une même forme orthographique (de même lemme et de même catégorie grammaticale et, pour les noms, de même genre) qui varie par ses descripteurs linguistiques (les traits flexionnels) est répétée dans MORPHALOU, alors qu'elle ne l'est pas dans le DELAF où tous les traits flexionnels sont notés sur la même ligne et constituent donc une seule et même entrée. Voici l'exemple de la forme de l'adjectif *homogène* :

Dans MORPHALOU :

```
<lexicalEntry lemma="homogène" grammaticalCategory="adjective">
  <inflectionGroup>
    <inflection orthography="homogène"
grammaticalGender="masculine" grammaticalNumber="singular"/>
    <inflection orthography="homogènes"
grammaticalGender="masculine" grammaticalNumber="plural"/>
```

```

      <inflection orthography="homogène"
grammaticalGender="feminine" grammaticalNumber="singular"/>
      <inflection orthography="homogènes"
grammaticalGender="feminine" grammaticalNumber="plural"/>
    </inflectionGroup>
  </lexicalEntry>

```

Dans le DELAF :

homogène,.A:ms:mp:fs:fp

3) Dans MORPHALOU, le lemme des noms féminins obtenus par dérivation suffixale est la forme féminine du nom. Le rapport entre les formes masculines et féminines est mentionné par la composante */feminineOf/*. Voici l'exemple de l'entrée lexicale des lemmes *chat* et *chatte*.

```

  </lexicalEntry>
  <lexicalEntry lemma="chatte" grammaticalCategory="commonNoun"
grammaticalGender="feminine" feminineOf="chat">
    <inflectionGroup>
      <inflection orthography="chatte"
grammaticalNumber="singular"/>
      <inflection orthography="chattes"
grammaticalNumber="plural"/>
    </inflectionGroup>
  </lexicalEntry>
  <lexicalEntry lemma="chat" grammaticalCategory="commonNoun"
grammaticalGender="masculine">
    <inflectionGroup>
      <inflection orthography="chat"
grammaticalNumber="singular"/>
      <inflection orthography="chats"
grammaticalNumber="plural"/>
    </inflectionGroup>
  </lexicalEntry>

```

4) Aucune forme composée, qu'elle soit graphiquement marquée par le trait d'union (*prie-Dieu*) ou non (*parce que*) n'est répertoriée dans le lexique de MORPHALOU.

5) La couverture linguistique de MORPHALOU se répartit comme suit :

- noms¹⁰ : 42 981 lemmes / 85 552 formes fléchies ;
- adjectifs : 15 211 lemmes / 60 498 formes fléchies ;
- verbes : 7 387 lemmes / 37 6851 formes fléchies ;
- adverbes : 1 579 ;
- mots grammaticaux : 176 ;
- interjections : 139 ;
- onomatopées : 41.

La catégorie des « mots grammaticaux » de MORPHALOU comprend les pronoms, les déterminants, les conjonctions (de subordination et de coordination), les prépositions ou encore quelques adverbes. Les mots grammaticaux ne sont en fait pas sous-catégorisés en ces différentes classes. Les interjections et les onomatopées sont rangées dans deux autres catégories spécifiques.

Le code de ces entrées est *functionWord*. Les informations très lacunaires associées à cette métacatégorie ne nous permettent guère de tirer des conclusions. Elles les rendent en l'état inutilisables pour l'étiquetage morphosyntaxique.

Parmi les lexiques morphologiques, nous pouvons encore citer les ressources développées dans le cadre de MULTEXT (Ide et Véronis 1994) ou la base de données CELEX (Burnage 1990).

À côtés de ces lexiques morphologiques, on voit se développer actuellement des lexiques syntaxiques ou sémantiques, dans lesquels les informations codées sont plus complexes (Sagot 2006 ; voir aussi le projet LexSynt¹¹). Nous ne nous intéressons pas dans ce travail à ce type de codage, non présent dans nos ressources lexicales.

4 LA NORME LEXICOGRAPHIQUE : NOS CHOIX LINGUISTIQUES DE CATÉGORIES GRAMMATICALES

Dans cette section, nous décrivons les choix que nous avons faits en ce qui concerne les catégories grammaticales, les sous-catégories et les informations associées. Il s'agit bien de choix puisque cette classification ne s'impose pas d'elle-même. Celle proposée par la grammaire traditionnelle n'a d'ailleurs jamais cessé de faire l'objet de nombreuses critiques (Wilmet 1997a notamment) et de contre-propositions, dont aucune n'a jamais rallié

¹⁰ Il s'agit de la catégorie dite des noms communs, aucun nom propre ne figurant dans le lexique.

¹¹ <http://lexsynt.inria.fr/index.php>.

l'ensemble de la communauté des linguistes et grammairiens. Les décisions que nous avons prises quant à la catégorisation sont aussi liées à la question : « Quels types d'étiquettes souhaitons-nous au final ? » Le choix des étiquettes est fonction des buts poursuivis. En effet, nous l'avons déjà dit, il n'y a pas d'étiquetage « en soi » et cette tâche procède toujours d'une série de décisions qui sont à poser dès le début de l'entreprise. Nos choix linguistiques tiennent compte de trois paramètres principaux :

- l'utilisation ultérieure du corpus étiqueté. Celui-ci sera manipulé par des linguistes, en vue de recherches concernant principalement la syntaxe et la morphosyntaxe, mais aussi dans le cadre de l'analyse du discours. Les linguistes qui interrogeront le corpus ne s'inscrivent pas dans une théorie syntaxique particulière, qui ne doit donc pas guider le choix de nos étiquettes. Ainsi par exemple, les codes utilisés doivent, moyennant un apprentissage peu coûteux, être relativement limpides ;
- la particularité de notre corpus, composé de transcriptions de l'oral spontané. Ainsi, les données particulières que constituent les transcriptions rendent inopérantes certaines règles de levée d'ambiguïtés. De plus, certaines distinctions ne doivent pas être prises en compte, car les ambiguïtés ainsi générées sont impossibles à lever (cf. ci-dessous le cas de la distinction des pronoms toniques et sujet pour le *elle* par exemple) ;
- la particularité du système de levée d'ambiguïtés. En effet, certaines ambiguïtés sont impossibles à lever automatiquement, parce que les informations dont on devrait disposer pour le faire ne sont pas disponibles (la distinction pronom relatif / pronom interrogatif par exemple).

Nous détaillons les choix que nous faisons, liés au fait qu'il ne faut pas rendre la tâche de l'étiquetage impossible avant de l'entreprendre. On oscille donc entre le souhaitable et le faisable, entre ce qui est pertinent pour la théorie et ce qui est praticable (cf. § 4.5 pour la prise en compte des mots composés).

On a donc dans ce travail un jeu d'étiquettes moins fin que celui utilisé dans les grammaires que nous avons construites, avec le même système de levée d'ambiguïtés, pour le français écrit standard (cf. Blanc et Dister 2003, Dister 2004).

Nous séparons dans la description ci-dessous, comme il est classique de le faire, les mots lexicaux et les mots grammaticaux, appelés aussi mots vides (Tesnière 1969). Si les premiers constituent une classe ouverte, impossible à définir en extension, les seconds forment une classe pratiquement fermée, dans laquelle les nouveaux items sont très rares.

Nous verrons que nous nous éloignons finalement assez peu de la classification établie par la grammaire traditionnelle, principalement pour la raison évoquée ci-avant, qui tient compte des futurs utilisateurs du corpus étiqueté.

4.1 Les mots lexicaux

Parmi les mots lexicaux, nous avons les noms (divisés en deux catégories : les noms propres et les noms communs), les adjectifs, les verbes et les adverbes.

4.1.1 Les noms propres : NPr

Étant donné les choix de transcription adoptés pour notre corpus, seront considérées comme des noms propres toutes les graphies commençant par une majuscule, à l'exception des lettres majuscules isolées les unes des autres qui transcrivent effectivement la lettre elle-même (et sont donc des noms communs).

Ainsi, à la différence des corpus standard pour lesquels une désambiguïsation est nécessaire, la forme *Pierre* dans notre corpus sera toujours effectivement le prénom et non le substantif en début de phrase. Nous avons vu au chapitre 4 comment nous reconnaissons ces formes lors du prétraitement, grâce à un graphe morphologique.

De ce fait, les classes de mots suivantes seront considérées comme des noms propres :

- les prénoms ;
- les noms de famille ;
- les toponymes et les gentilés ;
- les titres d'œuvres, de journaux, de films ;
- les noms de journaux, de chaînes télévisées, d'émissions, etc. ;
- les sigles, notés totalement en majuscules ;
- les acronymes pour lesquels seule la première lettre est en majuscules.

L'application du graphe *NomsPropres* (décrit au chap. 4, § 6.3.2) sur notre corpus reconnaît et étiquette les séquences suivantes (ordre alphabétique) :

Mots simples :

Aarschot, Académiciens, Académie, Académies, Acadi, Acropole, Adeps, Adidas, AFP, Africains, Afrique, Aix, Albion, Alfred, Alger, Algérie, Algériennes, Algériens, Aline, etc.

Mots composés :

ALE, A-l'Ouest-rien-de-nouveau, A-propos, Académie Française, Académie-Française, Académie-française, Aimé Quernol, Aire-le-Romain, Alain Destreel, Alain Fournier, Alain Guallon, Alain Guillaou, Alain Lerond, Albert Desmed, Albert Jacquard, Alexandre Jardin, Alfred Hitchcock, etc.

Cette méthode de reconnaissance des noms propres appelle cinq remarques :

1. Les conventions de transcription de VALIBEL préconisent l'utilisation des tirets dans les noms d'œuvres, les titres, etc. Cela a pour conséquence que ces séquences, telles *La-Chartreuse-de-Parme* et *Le-Soir*, seront analysées comme une seule unité. C'est le cas dans les exemples donnés ci-dessus pour le journal *À propos* ou pour le livre, dont on a fait un film, *À l'Ouest rien de nouveau*. Cela est évidemment un avantage lors de l'étiquetage de ne pas décomposer ces séquences en unités plus petites ;
2. On ne distingue pas ici de sous-catégorie de noms propres, puisque prénoms, œuvres, toponymes, gentilés, etc. reçoivent le code unique *Npr*. Or, ils ont des comportements syntaxiques différents. Ainsi par exemple, linguistiquement, les gentilés sont considérés comme des noms communs puisqu'ils sont accompagnés d'un déterminant ;
3. On n'a pas les indications de genre et de nombre traditionnellement associées aux noms, puisque seul le code *NPr* est associé aux formes reconnues ;
4. Le lemme est automatiquement le même que la forme en entrée, puisque les formes ont dans le texte le format suivant : *{Aarschot,.NPr}*, *{Académiciens,.NPr}*, *{Académie,.NPr}*, *{Académies,.NPr}*, etc. Ainsi, on n'a pas de lien pour les gentilés entre les formes masculines et le féminin, comme on le voit dans les exemples ci-dessus avec *Algérien* et *Algérienne*. Ce lien morphologique serait classiquement noté si les entrées avaient été encodées manuellement. De la même manière, *Académie* et *Académies* sont considérés comme deux formes séparées, qui seraient deux entrées séparées dans le dictionnaire, avec deux lemmes différents.

5. Une dernière remarque, liée à nos données. On constate ici la variation graphique présente dans les transcriptions : on a ainsi les trois noms propres reconnus : *Académie Française* (deux majuscules, pas de tiret), *Académie-Française* (deux majuscules, tiret), et *Académie-française* (une majuscule, une minuscule et tiret). Là où une constitution manuelle des données aurait établi un lien entre ces trois variantes¹² (par le choix d'un lemme commun), celles-ci sont ici présentées sans lien les unes avec les autres.

Néanmoins, malgré ces remarques, cette méthode de reconnaissance nous semble intéressante dans le cas de nos transcriptions. En effet, elle évite un recensement laborieux préalable de ces formes, puisque l'heuristique de la majuscule permet de les reconnaître sans ambiguïté¹³.

4.1.2 Les noms (communs) : N

Tout comme pour les verbes, les adjectifs et les adverbes, nous avons repris les données du dictionnaire sans guère les modifier. Dans le corpus étiqueté, un nom doit être accompagné de son lemme, de sa catégorie grammaticale et des informations flexionnelles de genre et de nombre.

Certaines incohérences peuvent subsister dans le corpus étiqueté ; celles-ci sont en général liées aux choix de transcription. Par exemple, la variation graphique que nous citons ci-dessus pour *Académicien/académicien* et qui se rencontre pour un certain nombre de formes¹⁴.

4.1.3 Les verbes : V

Aux verbes sont associées des informations quant au lemme, au temps et au mode, ainsi qu'à la personne. Les participes passés sont également codés comme des verbes (*V:K*) mais possèdent en plus des informations de genre et de nombre (mais non de personne).

¹² On a aussi dans le corpus une occurrence de *Académie française* (majuscule, minuscule, pas de tiret), qui n'est pas reprise dans les noms composés puisque la recherche avec le graphe s'arrête s'il y a un blanc et que le mot qui suit commence par une minuscule. Cf. graphe décrit au chap. 4, § 6.3.2.

¹³ Dans le graphe dictionnaire distribué avec UNITEX, l'heuristique est la suivante : est étiqueté comme nom propre tout mot commençant par une majuscule et non reconnu par l'application préalable des dictionnaires. (cf. Paumier 2006 : 47)

¹⁴ Il serait souhaitable que ces incohérences soient corrigées lors de la validation de l'étiquetage. Il faudrait aussi sans doute veiller à reformuler ce que l'on entend par « nom propre » dans les conventions de transcription de VALIBEL.

Nous n'avons pas fait de distinction pour les verbes *être* et *avoir* selon qu'ils sont auxiliaires ou formes pleines. Ces verbes sont donc considérés comme non ambigus et codés *V*. Cela a comme corollaire que, à l'heure actuelle, nous ne tenons pas compte des temps composés. Ainsi, il n'y a pas de code dans le texte étiqueté, comme c'est le cas dans les dictionnaires, pour le passé composé ou le plus-que-parfait. La difficulté est que ces formes composées peuvent être discontinues ; on peut les rencontrer avec des insertions, comme le montre l'exemple suivant :

blaAD1 mm -| ben tu sais elle **a jamais vraiment**
cherché à savoir ce qu'il avait hein [blaNB1]

Cette discontinuité des séquences fait qu'on ne peut avoir une étiquette qui regrouperait les deux formes, alors que ce serait envisageable dans les cas où elles sont contiguës. Il faudrait les déplacer dans le texte, et c'est sans doute la solution que nous envisageons à terme, afin de prendre en considération les temps composés¹⁵.

4.1.4 Les adjectifs : A

Comme le nom, l'adjectif est accompagné de son lemme et d'informations flexionnelles de genre et de nombre. Nous ne faisons pas de sous-catégorisation des adjectifs, entre qualificatif, possessif, etc.

Beaucoup des adjectifs codés dans le dictionnaire sont ambigus avec le participe passé. Ainsi, *abandonné* est codé comme suit :

abandonné,.A:ms
abandonné,abandonner.V:Kms

La distinction entre adjectif et participe passé est souvent difficile à établir, et il semble qu'il y ait un continuum entre ces deux classes. Ainsi, dans

un enfant abandonné

abandonné est un adjectif (le fameux « participe passé employé seul » de l'école primaire), là où le complément d'agent en fait un participe passé dans

un enfant abandonné par ses parents.

Des critères existent pour distinguer les deux emplois (utilisation de *très*, paraphrase avec une relative, etc.), mais de nombreux cas sont « à la limite ». Clément (2001 : 84), pour la validation de l'étiquetage, parle en dernier recours dans ces cas de « l'intuition de l'annotateur ».

¹⁵ Voir Gross (1999) pour la lemmatisation des temps composés en anglais.

Dans le DELAF, cette ambiguïté adjectif/ participe affecte 7759 formes. Nous avons décidé de neutraliser cette ambiguïté en ajoutant un code supplémentaire : *B*. L'entrée de *abandonné* devient alors :

abandonné,abandonner.V:Bms

On a alors trois types d'entrées : celles où l'adjectif n'est pas ambigu avec un participe passé (le lemme est l'adjectif et le code est *A*), celles où le participe passé n'est pas ambigu avec un adjectif (le lemme est le verbe et le code est *V:K*) et enfin celles où l'adjectif et le participe sont ambigus (le lemme est le verbe et le code est *V:B*). Les entrées suivantes illustrent ces trois cas de figure :

abordables,abordable.A:mp:fp
reconverti,reconvertir.V:Kms
reconnues,reconnaître.V:Bfp

Cette étiquette indifférenciée est également proposée dans le cadre de GRACE, campagne d'évaluation des catégoriseurs morphosyntaxiques (cf. chap. 6, § 7).

4.1.5 Les adverbes : ADV

Les adverbes ne connaissent pas de flexion : le lemme est toujours identique à la forme en entrée dans le dictionnaire.

La seule exception est *tout*, qui prend une forme féminine devant un mot féminin commençant par une consonne (ou un h aspiré).

4.2 Les mots grammaticaux

Dans le DELAF, les mots grammaticaux se répartissent en sept catégories distinctes : les conjonctions de coordinations (CONJC), les conjonctions de subordination (CONJS), les déterminants (DET), les interjections (INTJ), les préfixes (PFX), les prépositions (PREP) et les pronoms (PRO).

Nous avons procédé à un recodage des données, que nous détaillons ci-dessous pour chaque catégorie.

4.2.1 Les préfixes : PFX

Un préfixe est une forme graphique, non autonome, qui précède un mot – dans le sens d'unité graphique telle que nous l'avons définie plus haut – et est unie à celui-ci par un trait d'union. Le préfixe est un élément de formation de mots nouveaux.

Nous englobons également dans la catégorie des préfixes des formes qui, dans les grammaires traditionnelles, relèvent de la composition, dans la mesure où les termes unis ont une existence indépendante antérieurement à leur union (BU § 176). Ainsi, les formes de composition savante en *-o* seront-elles notées également préfixes. Ceci se justifie dans la mesure où ces termes n'ont en fait pas d'existence autonome sous la forme en *-o* (celle-ci dérive par suffixation de la forme autonome). Ces composés relèvent fréquemment de diverses terminologies (souvent scientifique : *alcalino-*) et servent notamment à désigner des communautés ou des langues (*italo-belges, anglo-américain*).

La différence entre dérivation préfixale et composition avec les formes ici concernées n'étant pas très nette, nous avons décidé de nous en passer, et ce d'autant plus que la combinatoire est très productive (*belgo-polonais, italo-turc, italo-français, franco-italien, turco-polonais, etc.*).

Dans de nombreux cas, le préfixe est directement soudé à la base : *a* (*apesanteur, apolitique*), *in* (*inégal, innommable, incompetent*), *pré* (*prévente, prépension, prérequis*), *re* (*redécorer, reconduire, repenser*), etc. Sans exclure de la catégorie des préfixes ces formes *a, in, pré* ou *re*, celles-ci ne nous intéressent cependant pas en tant que telles dans les mots repris ci-dessus. En effet, la composition interne de ces mots, graphiquement soudés, ne nous occupe pas ; ceux-ci, lexicalisés, doivent figurer tels quels dans une entrée du dictionnaire.

Néanmoins, pour des créations plus récentes, certaines variantes graphiques peuvent se rencontrer dans les textes. Ainsi, lorsque la dérivation suffixale se fait avec des morphèmes grammaticaux classiquement soudés à la base, on peut trouver les variantes où celles-ci ne sont pas soudées. Dans le journal *Le Soir* de 2001, on a : *re-rugit, re-zapper, re-conçus, re-hommage, re-révélation, re-arrêt, re-régulation, etc.*, à côté de *rescolarisation, repolitisation, repliage, rephasage, rephaser, rephrasé, repigmentation, etc.*

Les cas où le préfixe est soudé et ceux où il ne l'est pas ne sont pas toujours clairs et les dictionnaires et grammaires ne s'entendent pas toujours selon les cas.

Ainsi, dans les transcriptions de VALIBEL, nous trouvons les trois variantes graphiques pour *hyper*¹⁶ :

- soudure : *hypercorrectisme* [accFJ1], *hypertension* [norPM1] ;

¹⁶ Ceci est typiquement le cas de conventions à ajouter dans le référentiel de transcription, afin d'assurer une harmonisation des transcriptions, et cela d'autant plus que de nombreuses formations ne sont pas reprises dans le dictionnaire de référence proposé aux transcrip-teurs.

- trait d'union : *hyper-spécialisation* [iljPF1], *hyper-riche* [ilpMJ1] ;
- suite de deux mots : *hyper accentué* [accCP1], *hyper joli* [famVV1].

Le désavantage de la soudure dans les néologismes est que ceux-ci doivent être introduits dans les dictionnaires, tandis que la formation d'une unité par un trait d'union ou la simple juxtaposition permet de reconnaître et d'analyser les formes¹⁷.

Parmi les préfixes les plus productifs en français contemporain nous pouvons citer : *archi*, *hyper*, *super*, *extra*, *anti*, *ultra*, *ex*, *mi*, etc.

Notons d'emblée qu'à la différence de la suffixation, la préfixation n'entraîne pas de changement de catégorie grammaticale du mot concerné. Celui-ci peut être un nom (*anti-mondialiste*), un adjectif (*hyper-attentif*), un verbe (*re-zapper*), etc.

4.2.2 Les pronoms : PRO

Étymologiquement, les pronoms sont des formes qui sont mises « pour le nom ». Ils ont donc en général une fonction de *représentant* (BU § 626).

Classiquement, les pronoms sont subdivisés en six sous-catégories, selon leurs propriétés syntaxiques et sémantiques. Ils varient en nombre et en genre et, pour les personnels, les relatifs et les interrogatifs, également en fonction.

4.2.2.1 Les pronoms dits personnels : PRO+Ppers

La caractéristique des pronoms personnels est de marquer la personne grammaticale des êtres ou des choses qu'ils remplacent (BU § 961). Les pronoms personnels varient donc non seulement en genre et en nombre mais également en personne.

Cet héritage des cas latins se marque dans la forme du pronom qui, tout en ayant le même genre, le même nombre et la même personne, varie selon la fonction syntaxique qu'il exerce dans l'énoncé.

Suivant ce critère de la fonction syntaxique, il est possible de distinguer trois catégories parmi les pronoms personnels, selon qu'ils remplissent la fonction de sujet, d'objet direct ou d'objet indirect.

Par ailleurs, il faut noter également que le pronom peut changer de forme selon la place qu'il occupe dans l'énoncé, et cela en ayant toujours les

¹⁷ Cette limite est purement informatique et devrait bientôt être réglée pour le français grâce au programme Polylex d'UNITEX (Paumier et Ulland 2005).

mêmes genre, nombre et personne et en exerçant la même fonction. On a ainsi :

- a. Je leur explique.
- b. J'explique à eux.
- c. Je te parle.
- d. Je parle à toi.

En *a* et *c*, les pronoms *leur* et *te* sont dits *conjoint*s (ou *atones* ou *faibles* ou *préverbaux* ou *clitiques*), alors qu'en *b* et *d*, *eux* et *toi* sont *disjoint*s (ou *toniques* ou *forts*). Les termes *conjoint* et *disjoint* s'emploient pour situer le pronom par rapport au verbe. Les pronoms conjoints se caractérisent en effet par une distribution particulière et une place stable dans l'environnement du verbe. Celle-ci, sans être figée, est relativement contrainte.

Ainsi, par exemple, entre le clitique sujet et le verbe, on ne peut trouver qu'un (plusieurs) autre(s) pronom(s) clitique(s) et / ou une particule négative. Par ailleurs, l'ordre de ces autres pronoms clitics est lui aussi contraint :

- Il rendra.
- Il le rendra.
- *Le il rendra.
- Il le lui rendra.
- *Il lui le rendra.
- Il ne rendra pas.
- *Ne il rendra pas.
- Il ne le rendra pas.
- *Il le ne rendra pas.
- Il ne le lui rendra pas.
- *Il ne lui le rendra pas.

À cette catégorie des conjoints, on ajoute également les pronoms parfois qualifiés de pronoms adverbiaux *en* et *y* (Pinchon, 1972). Leur particularité est de remplacer des groupes prépositionnels.

Nous verrons que cette distribution particulièrement contrainte permet de construire des grammaires qui lèvent l'ambiguïté avec un très bon rendement (et un taux d'erreur nul pour les corpus écrits standard).

Le tableau ci-dessous, repris du *Bon Usage*, synthétise les emplois et les formes respectives des pronoms personnels.

	Formes conjointes						Formes disjointes			
	Sujet		Autres fonctions				Non réfléchi		Réfl.	
			Objet direct		Obj.					
	Masc.	Fém.	Masc.	Fém.	indir.		Réfl.	Masc.		Fém.
1 ^{re} pers. sing.	je		me				moi			
2 ^e pers. sing.	tu		te				toi			
3 ^e pers. sing.	il/on	elle	le	la	lui	se	lui	elle	soi	
1 ^{re} pers. pl.	nous									
2 ^e pers. pl.	vous									
3 ^e pers. pl.	ils	elles	les		leur	se	eux	elles	soi	

Tableau 38. Formes du pronom personnel (BU § 634)

Ainsi, selon ce classement, différents codes caractérisent les pronoms, qui reflètent les sous-catégories établies ici :

- les pronoms préverbaux sujets (PRO+PpvIL) : le pronom est sujet du verbe : *il, elle, ils, elles, on, je, tu, nous, vous, l'on*¹⁸ ;
- les pronoms préverbaux compléments directs (PRO+PpvLE) : le pronom est objet direct du verbe : *le, la, les, me, te, se, moi, toi, nous, vous* ;
- pronoms préverbaux compléments indirects (PRO+PpvLUI) : le pronom est objet indirect du verbe : *lui, leur, me, te, se, moi, toi, nous, vous* ;
- les pronoms adverbiaux (PRO+PpvPR) : *en, y* ;
- les pronoms disjoints¹⁹ (PRO+Ton) : *elle, cézigue, elle-même, lui-même, eux-mêmes, lui, moi, nous autres, soi, soi-même, etc.*

Ce classement appelle certaines remarques. En effet, les pronoms disjoints peuvent exercer les trois fonctions des pronoms conjoints :

¹⁸ Nous classons *ce* et *cela* parmi les pronoms démonstratifs.

¹⁹ Dans les grammaires de levées d'ambiguïtés construites pour l'écrit standard (Dister 2004), ils sont appelés toniques, d'où leur symbole *PRO+Ton*. Cette appellation nous semble peu adéquate car des *PRO+PpvLE*, en position postverbale, deviennent également toniques : donne-le ! Il nous semble donc plus judicieux de garder la distinction conjoint/disjoint, afin de ne pas rendre le terme de *tonique* ambigu.

- sujet : *eux aussi rêvaient*
- objet direct : *je les regarde, eux et personne d'autre*
- objet indirect : *je donne les clés à eux et à mon frère*

Ainsi, un étiquetage qui tiendrait compte des distinctions entre les pronoms selon leur fonction syntaxique manquerait d'une certaine manière de cohérence puisque dans le cas des formes disjointes, cette fonction syntaxique ne figurerait pas dans l'annotation, alors qu'elle le serait pour les formes conjointes.

Néanmoins, ce choix de catégorisation a été fait pour l'étiquetage de corpus de français standard avec ELAG, dans la mesure où la prise en compte de telles distinctions s'avère « rentable » pour l'écriture de grammaires : la distinction des fonctions syntaxiques permet de lever certaines ambiguïtés, la distribution des clitiques étant très contrainte.

Le *Corpus annoté de Paris 7* (Clément 2001 : 99) fait également la distinction entre clitiques et pronoms forts.

Si nous l'abandonnons dans le cadre de notre recherche, c'est parce qu'elle génère de nombreuses ambiguïtés. On le voit, les formes *elle* et *elles* peuvent être sujet (PRO+PpvIL), objet (PRO+PpvLUI) et disjoint (PRO+Ton). Une seule forme et trois analyses possibles, parce que l'on tient compte de la sous-catégorisation.

De plus, la plupart des ambiguïtés engendrées sont impossibles à lever dans notre corpus, non seulement de manière automatique, mais également lors de la lecture du texte par un humain.

C'est notamment le cas pour des séquences où se suivent plusieurs occurrences du pronom. Les exemples suivants serviront à éclairer notre choix.

accGP1 vous auriez dû poser des questions à ma fille
parce que elle elle est // elle est fort // pointilleuse /
sur sur cet accent (en riant) [accGP 1r]

iljPF1 et elle elle parle assez bien / franchement (rire)
[iljPF1r]

Dans un corpus de français écrit standard, on aurait l'analyse suivante : le premier *elle* est un pronom disjoint (tonique), tandis que le second est le pronom sujet. Or, dans le français parlé où abondent les disfluences, transcrites dans les données que nous exploitons (cf. chap. 1 et chap. 3), l'analyse peut être une répétition du pronom sujet. Dans la plupart de ces cas, le recours à la bande son sera nécessaire pour permettre l'annotation,

seule la prosodie permettant dans de nombreux cas de lever l'ambiguïté (cf. chap. 3).

Nous avons donc choisi, pour des raisons de faisabilité, et parce que l'aspect de la fonction syntaxique ne nous semble pas relever de l'étiquetage grammatical, de ne garder qu'une seule métacatégorie « pronom personnel », dont les items recevront le code *PRO+Pers*.

Dans cette classe,

1. nous ne faisons donc pas de distinction entre pronom réfléchi et pronom réciproque ;
2. nous annotons les formes composées : *eux-mêmes*, *l'on*, *vous autres*, etc.
3. nous ne distinguons pas le genre des pronoms quand ceux-ci sont ambigus de cet unique point de vue. Ainsi, *il* sera bien étiqueté masculin singulier et *elles* féminin pluriel, mais *je* et *tu* seront uniquement singulier et *nous* et *vous* seulement pluriel, sans indication de genre. En effet, l'ambiguïté entre masculin et féminin est impossible à lever dans la majorité des occurrences de ces formes pour lesquelles on n'a pas d'indice morphosyntaxique dans le texte (on peut lever cette ambiguïté, par exemple, dans un temps composé où le participe passé révèle le genre du pronom sujet) ;
4. nous ne distinguons pas les emplois impersonnels de *il* des emplois personnels (*il pleut* vs *il est boucher*).

4.2.2.2 Les pronoms relatifs : *PRO+Prel*

Le pronom relatif introduit une proposition dans laquelle il exerce une fonction. Comme le pronom personnel, le relatif a un genre, un nombre et une personne. Ces informations flexionnelles peuvent parfois être déduites de la seule forme du relatif (*lesquelles*, *auquel*), parfois non (*dont*, *qui*). Dans ces cas, lors de l'annotation automatique, la détermination du genre est particulièrement complexe. Nous ne codons donc pas le genre de ces pronoms dans le dictionnaire.

Les pronoms relatifs ont des formes graphiquement simples (*qui*, *que*, *quoi*, *dont*, *où*, *lequel*, *laquelle*, *lesquels* et *lesquelles*), dont des formes contractées (*auquel*, *auxquels*, *auxquelles*). Comme nous avons décidé de ne pas décontracter les formes contractées lors du prétraitement (chap. 4), notre codage sépare donc de cette série la forme *à laquelle* qui se présente dans les textes de manière linéaire et non contractée.

Toutes ces formes reçoivent le même code *PRO+Prel*, mais les formes contractées reçoivent en plus le code *Preppro* (pour *préposition* + *pronom*).

Ce code permet de tenir compte de la distribution particulière de ces formes par rapport aux formes non contractées.

Nous n'intégrons pas les prépositions dans les formes composées du pronom (*vers lesquels, dans lequel, etc.*).

4.2.2.3 Les pronoms démonstratifs : PRO+Pdem

« Les pronoms démonstratifs désignent un être ou une chose en les situant dans l'espace (fonction *déictique*) ; ils peuvent aussi renvoyer à un terme qui précède (fonction *anaphorique*) ou qui suit (fonction *cataphorique*) dans le contexte. » (BU § 665). Nous ne distinguerons pas ces trois emplois dans l'annotation.

Comme les pronoms personnels, les pronoms démonstratifs varient en genre et en nombre. Aux formes simples de ces pronoms correspondent des formes composées par l'adjonction de l'adverbe de lieu *ci* ou *là* précédé d'un trait d'union, adverbe qui vise à renforcer la valeur démonstrative : *celle, celle-ci, celle-là*.

Les formes *icelui, icelle, iceux* et *icelles* sont des formes sorties de l'usage, auxquelles on préfère actuellement *celui, celle, ceux* et *celles*. Si elles ont peu de chances de figurer dans un corpus de français contemporain, sauf peut-être à titre de plaisanterie (cf. mention du *Petit Robert*), nous les répertorions néanmoins dans la mesure où nos données sont ainsi complètes et où, de toute façon, du fait qu'elles ne sont pas ambiguës avec d'autres formes, leur présence dans le dictionnaire ne compliquera en rien le traitement.

Parmi les pronoms démonstratifs figurent deux formes élidées : *c'* et *ç'*. Nous leur avons associé respectivement les lemmes *cela* et *ça*.

4.2.2.4 Les pronoms possessifs : PRO+Ppos

« Les pronoms possessifs sont des représentants qui indiquent que l'être ou la chose dont il s'agit sont en rapport avec une personne grammaticale : celui qui parle ; celui à qui l'on parle, celui ou ce dont on parle. » (BU § 661)

Nous avons 21 formes de cette catégorie, notée PRO+Ppos. Toutes les formes de pronoms possessifs sont des formes graphiquement composées :

- *le mien, le tien, le sien, le nôtre, le vôtre, le leur*, au masculin singulier ;
- *la mienne, la tienne, la sienne, la nôtre, la vôtre, la leur*, au féminin singulier ;

- *les miens, les tiens, les siens*, au masculin pluriel ;
- *les miennes, les tiennes, les siennes*, au féminin pluriel ;
- *les nôtres, les vôtres, les leurs*, au masculin et au féminin pluriel.

Historiquement, le pronom possessif est composé d'un déterminant défini et d'un adjectif possessif. Quand la personne grammaticale est la 1^{re} personne du singulier, les formes se déclinent sur *mien* ; pour la 2^e personne du singulier, l'adjectif est *tien*, etc. Nous codons cette personne grammaticale dans les dictionnaires, en plus de la flexion en nombre ou en genre. Pour le féminin singulier, cela donne les entrées suivantes :

la mienne, le mien. PRO+Ppos+Ppos1s:fs
 la tienne, le tien. PRO+Ppos+Ppos2s:fs
 la sienne, le sien. PRO+Ppos+Ppos3s:fs
 la nôtre, le nôtre. PRO+Ppos+Ppos1p:fs
 la vôtre, le vôtre. PRO+Ppos+Ppos2p:fs
 la leur, le leur. PRO+Ppos+Ppos3p:fs

La sous-catégorisation selon le nombre et la personne du « possesseur » donne la répartition suivante :

Pos1s : *le mien, la mienne, les miens, les miennes* ;

Pos2s : *le tien, la tienne, les tiens, les tiennes* ;

Pos3s : *le sien, la sienne, les siens, les siennes* ;

Pos1p : *le nôtre, la nôtre, les nôtres* ;

Pos2p : *le vôtre, la vôtre, les vôtres* ;

Pos3p : *le leur, les leurs*.

Garder cette sous-catégorisation n'engendre pas d'ambiguïtés supplémentaires dans la mesure où chaque forme n'est attribuée qu'à un seul possesseur. Cette indication permettra éventuellement, une fois le texte annoté, de faire des recherches sur la sous-catégorie sans devoir spécifier les formes concernées.

En ce qui concerne les formes ambiguës du seul point de vue du genre (*les nôtres, les vôtres*, etc.), nous avons seulement codé le nombre.

Toutes les formes qui composent les possessifs sont ambiguës non seulement avec la forme du déterminant et de l'adjectif qui s'emploient toujours indépendamment de leur composition dans le pronom possessif, mais également avec d'autres formes très fréquentes :

- un déterminant défini : *le, la, les, leur, leurs* ;
- un adjectif possessif : *mien, miens, mienne, miennes, nôtre, nôtres, vôtre, vôtres* ;
- un pronom personnel : *le, la, les, leur, lui*.

Néanmoins, ces formes composées sont non ambiguës dans le sens où elles ne sont pas en concurrence avec la suite des formes simples qui les composent, sauf pour *les tiens*, *la tienne* et *les tiennes* où la seconde forme est ambiguë avec une forme du verbe *tenir*, dans le premier cas à l'indicatif et dans les deux autres au subjonctif et *les tiens* :

Je les tiens enfermés chez moi.
Je veux qu'il la tienne en laisse.
Je veux que tu les tiennes en laisse.

À l'exception de ces trois cas précis, on peut faire des pronoms possessifs composés des formes non ambiguës.

Aux formes citées ci-dessus s'ajoutent 16 formes contractées (*du mien, des tiennes*, etc.). Nous avons procédé comme pour les pronoms relatifs : les formes sont conservées telles quelles, c'est-à-dire sans être décomposées, et le code *Preppro* (pour préposition + pronom) est ajouté à ces entrées.

4.2.2.5 Les pronoms indéfinis : *PRO+Pind*

La catégorie des pronoms indéfinis apparaît dans les grammaires comme une catégorie un peu fourre-tout ; elle est sans doute celle pour laquelle le consensus est le moins établi.

Nous avons gardé parmi les pronoms indéfinis les formes suivantes : *aucun, d'aucuns, autrui, beaucoup, certain, chacun, la plupart, l'un, nul, personne, peu, quelqu'un, quelques-uns, quiconque, rien, tout, tout un chacun*.

Sont considérées également les formes fléchies (*quelques-unes*) et contractées (*des uns*), avec pour ces dernières le code *Preppro* ajouté.

Quelque chose n'est pas repris ici car nous le considérons comme un syntagme nominal, tout comme *autre chose, peu de chose*, etc.

4.2.2.6 Les pronoms interrogatifs

Les pronoms interrogatifs sont des pronoms utilisés afin de poser une question, dans des tournures directes (*Qui vient ?*) et indirectes (*Je me demande qui vient*). Les formes des pronoms interrogatifs sont les mêmes que celles des pronoms relatifs, hormis le *dont* qui n'est que relatif.

Traditionnellement la forme *où* est également classée parmi les adverbes interrogatifs ; les grammaires renseignent également *combien* parmi les pronoms interrogatifs.

Ce serait le cas dans

je me demande combien tu gagnes

où le *combien* sert à marquer une question, mais pas dans

je sais combien tu gagnes

où il s'agit d'une affirmation. On voit combien (!) la frontière est mince ici.

Ainsi, cette classe des interrogatifs augmente le nombre d'ambiguïtés (exception faite de *dont*, chaque forme de relatif est ambiguë) et ces ambiguïtés sont très difficiles à lever. En effet, il faut des informations contextuelles et syntaxiques dont nous ne disposons pas actuellement afin de faire le départ entre pronom relatif et interrogatif (cf. le développement de Clément 2001 : 32).

4.2.3 Les déterminants : DET

La question des déterminants est sans doute l'une des plus complexes dans la catégorisation linguistique. La catégorie des déterminants reprend les catégories classiques des articles (définis, indéfinis, et partitifs), des « adjectifs » possessifs, des « adjectifs » démonstratifs, des « adjectifs » indéfinis, des « adjectifs » numériques, des « adjectifs » interrogatifs, des « adjectifs » exclamatifs et des « adjectifs » relatifs.

Cette nomenclature traditionnelle issue de la grammaire scolaire (en tout cas pour ce qui est d'y inclure les adjectifs) pose un certain nombre de problèmes, identifiés depuis longtemps. Selon Chevalier (1966 : 241) cette catégorie est un

assemblage dont la disparate est célèbre.

Gross (1976 : 9) va dans le même sens :

ces classifications sont défectueuses de divers points de vue.

Wilmet (1997a) la remet également en cause.

Les grammaires traditionnelles fournissent des listes de déterminants et de prédéterminants, listes établies en fonction de critères souvent divers.

Toujours selon Gross (1976 : 10), les grammaires donnent des

détails de la combinatoire superficielle du déterminant et du nom. (...) En général, deux (pré)déterminants

quelconques n'ont pas le même ensemble de propriétés syntaxiques.

On est donc face à une grande redondance dans les descriptions et les classifications qui ne sont jamais systématiques.

Ce n'est évidemment pas ici que nous réglerons ce problème particulièrement complexe. Nous avons donc fait des choix de classification critiquables à plus d'un titre, mais la question est épineuse et aucune solution totalement satisfaisante ne peut être envisagée.

Nous avons distingué six sous-classes de déterminants : les déterminants définis, les déterminants indéfinis, les déterminants possessifs, les déterminants démonstratifs, les déterminants numéraux et enfin la sous-classe groupée des déterminants interrogatifs et exclamatifs.

4.2.3.1 Les déterminants définis : *DET+Ddef*

Les déterminants définis sont *le, la, les* et la forme élidée *l'*, qu'on trouve devant un mot commençant par une voyelle ou un *h* dit aspiré (*l'enfant, l'hippopotame*).

Le déterminant défini se rencontre également sous des formes contractées avec les prépositions *à* et *de*²⁰ :

au < *à le* : *je vais au cinéma* ;

aux < *à les* : *je donne à manger aux poules* ;

du < *de le* : *je parle du voisin* ;

des < *de les* : *je parle des voisins*.

La contraction en *au* et en *du* se fait, au singulier, uniquement pour le déterminant masculin. Lorsque le déterminant est féminin, on garde la suite préposition + déterminant :

je vais au cinéma vs *je vais à la salle de sports*

je parle du voisin vs *je parle de la voisine*

Nous avons décidé de ne pas décontracter ces formes pour les raisons expliquées dans le chapitre 4. Néanmoins, nous ajoutons au code *DET+Ddef* le code *Prepdet* afin de pouvoir identifier ces formes et leur comportement particulier par rapport aux formes non contractées.

²⁰ La forme contractée *ès* ne figure pas ici car, du point de vue synchronique, c'est une préposition.

4.2.3.2 Les déterminants indéfinis : DET+Dind

Nous recensons parmi les déterminants indéfinis les formes simples suivantes : *un, une, des, de, d', aucun, certain, chaque, différent, divers, force, maint, nul, plusieurs, quelque, tel et tout*.

Nous ne classons pas parmi les déterminants indéfinis, suivant en cela la classification du *Bon Usage*, *autre, même* et *quelconque*.

En effet, ces formes n'ont pas la même distribution que les déterminants repris ici, dans la mesure où ils ne peuvent apparaître seuls devant le nom :

*autres majorités sont possibles.

Ils doivent être combinés avec un autre déterminant :

D'autres majorités sont possibles.

Maintes autres majorités sont possibles.

Tout fait exception dans la liste donnée ci-dessus, puisqu'il a le comportement d'un prédéterminant (*toutes les filles*). Nous le classons néanmoins parmi les déterminants indéfinis et nous gérerons cette variation de distribution lors de l'écriture des grammaires locales. Nous avons préféré cette solution à celle qui consiste à adopter les codes grammaticaux utilisés dans les grammaires de désambiguïsation que nous avons écrites pour le français standard (Dister 2004), codes qui sont moins intuitifs et nécessitent un apprentissage parfois un peu laborieux de la part de l'utilisateur. L'idée étant de rendre le corpus étiqueté accessible à des personnes non au fait d'une théorie linguistique particulière (en l'occurrence, celle du lexique-grammaire), nous avons adopté pour cette catégorie un codage plus consensuel (qui correspond donc *grosso modo* à celui des grammaires traditionnelles).

La distinction traditionnellement établie entre indéfinis et partitifs est liée au caractère non comptable (non nombrable, non quantifiable) du substantif qui suit le partitif : *manger du cheval, de la viande, de l'autruche, des vitamines*. On rencontre aussi le partitif devant des noms abstraits (*crever de peur*) ou devant des noms concrets mais à valeur d'espèce (*du lapin*) (BU § 567).

En fait, les partitifs ne sont qu'une sous-catégorie d'indéfinis, que nous ne distinguons pas dans l'étiquetage.

À la liste des déterminants indéfinis donnée ci-dessus s'ajoutent donc les formes suivantes, dont certaines sont graphiquement composées :

-
- *de la* : suivi d'un nom au féminin singulier qui commence par une consonne : *vendre de la camelote* ;
 - *du* : suivi d'un nom au masculin singulier qui commence par une consonne : *acheter du diesel* ;
 - *de l'* : suivi d'un nom au masculin ou au féminin singulier qui commence par une voyelle : *acheter de l'essence, de l'uranium* ;
 - *des* : suivi d'un nom au pluriel : *manger des épinards, des pâtes* ;
 - *de* : (allomorphe de *des*) suivi d'un adjectif et d'un nom au pluriel : *manger de bons épinards*. On trouve *de* partitif également avec des noms singuliers dans des tournures négatives : *je n'ai pas de temps à perdre, je ne veux pas de vin* (BU § 569C) ;
 - *d'* : suivi d'un adjectif commençant par une voyelle et d'un nom au pluriel : *manger d'excellents épinards*. On trouve *d'* partitif également avec des noms singuliers commençant par une voyelle dans des tournures négatives : *je n'ai pas d'argent*.

Outre les formes composées du partitif, on a aussi des déterminants indéfinis composés : *la plupart de, beaucoup de, etc.*

Nous ne tiendrons pas compte des prédéterminants composés dans lesquels la base est un substantif (cf. les déterminants nominaux de Buvet (1994) ou de Leroy (2005)) : *une armada de, une foule de, une quantité invraisemblable de, etc.* Ils seront donc analysés à ce stade comme une suite de mots simples.

4.2.3.3 Les déterminants possessifs : DET+Dpos

À l'inverse de ce que nous avons pour les pronoms possessifs, la catégorie des déterminants possessifs est uniquement composée de formes simples. Les formes concernées sont les suivantes :

- *mon, ton, son*, au masculin singulier ;
- *ma, ta, sa*, au féminin singulier ;
- *notre, votre, leur*, au masculin et au féminin singulier ;
- *mes, tes, ses, nos, vos, leurs*, au masculin et au féminin pluriel.

Outre l'accord classique en genre et en nombre avec le nom qu'ils déterminent, les possessifs portent la marque de nombre et de personne du « possesseur » (mais non son genre). Ainsi, cette sous-catégorisation donne le regroupement suivant :

- Dpos1s : *ma, mes, mon* ;
- Dpos2s : *ta, tes, ton* ;

- Dpos3s : *sa, ses, son* ;
- Dpos1p : *notre, nos* ;
- Dpos2p : *votre, vos* ;
- Dpos3p : *leur, leurs*.

Nous ajoutons ces codes dans les entrées du dictionnaire afin de permettre aussi des recherches spécifiques sur le possesseur.

4.2.3.4 Les déterminants démonstratifs : DET+Ddem

Dans cette sous-classe ne figurent également que des formes simples. Sont concernées les quatre formes suivantes :

ce, cet, cette, ces

Seule la première est ambiguë, avec le pronom démonstratif.

4.2.3.5 Les déterminants numéraux : DET+Dnum

Les adjectifs numéraux sont classiquement scindés en numéraux *cardinaux* et numéraux *ordinaux*. Les cardinaux indiquent le nombre des choses ou des êtres actualisés par le déterminant : *quatre consonnes et trois voyelles, trente-six chandelles, les cent et un dalmatiens*. Les ordinaux, quant à eux, indiquent le rang : *le troisième homme, la quatrième dimension*.

Les ordinaux, qui entretiennent des liens morphologiques réguliers avec les cardinaux (adjonction du suffixe *-ième*), sont classés parmi les adjectifs. En effet, d'un point de vue syntaxique, leur comportement est celui des adjectifs classiques et non des déterminants : ils ne peuvent jamais être utilisés seuls devant un nom, la présence d'un autre déterminant étant obligatoire dans le contexte gauche de l'ordinal²¹ :

* *septième victoire* vs *sa septième victoire*

Cette caractéristique distributionnelle nous fait ranger ces termes dans la catégorie des adjectifs (cf. ci-dessus pour *autre* classé aussi parmi les adjectifs et non parmi les déterminants indéfinis).

Voici les formes simples des déterminants numéraux :

un, une, deux, trois, quatre, cinq, six, sept, huit, neuf, dix, onze, douze, treize, quatorze, quinze, seize, vingt, trente, quarante, cinquante, soixante, septante, nonante, huitante, octante, cent, cents et mille.

²¹ On pourrait imaginer l'absence du déterminant dans des titres : *Septième victoire d'affilée pour Justine Hennin*.

Étant donné le nombre d'ambiguïtés générées par la prise en compte des formes *un* et *une* comme des déterminants numéraux, nous ne les prendrons pas en compte et considérerons qu'elles sont toujours des déterminants indéfinis.

Les formes composées des déterminants numéraux le sont à partir de ces formes simples, avec ou sans trait d'union selon les règles de l'orthographe traditionnelle, avec trait d'union selon les recommandations orthographiques de 1990 (Goosse 1991). Sans pour autant appliquer ces recommandations dans les transcriptions, les conventions de transcription de VALIBEL préconisent l'usage systématique du trait d'union dans les nombres cardinaux (que ceux-ci soient déterminants ou substantifs).

Afin de reconnaître ces formes composées, nous appliquons un graphe appelé *graphe dictionnaire*. Un graphe dictionnaire est un graphe qui, appliqué grâce au programme Locate en mode Merge (Paumier 2006 : 46 et sv., voir aussi chap. 4), vise à produire des séquences qui sont des entrées de dictionnaire.

Nous avons modifié le graphe de reconnaissance construit par Éric Laporte afin de le rendre compatible avec nos données qui systématisent le trait d'union du début à la fin du composé. Ainsi, à côté des formes déjà présentes dans le dictionnaire (les formes simples et quelques composés comme *dix-sept*), ce graphe de reconnaissance identifie les déterminants composés présents dans le texte et construit automatiquement les entrées correspondantes dans le dictionnaire du texte.

Sur notre corpus, le graphe identifie les formes qui correspondent aux premières entrées suivantes (classées par ordre alphabétique) :

cent-cinquante,cent-cinquante.DET+Dnum:mp:fp
cent-cinquante-mille,cent-cinquante-
mille.DET+Dnum:mp:fp
cent-mille,cent-mille.DET+Dnum:mp:fp
cent-septante,cent-septante.DET+Dnum:mp:fp
cent-vingt,cent-vingt.DET+Dnum:mp:fp
cinq-cents,cinq-cents.DET+Dnum:mp:fp
cinq-mille,cinq-mille.DET+Dnum:mp:fp
cinquante-cinq,cinquante-cinq.DET+Dnum:mp:fp
cinquante-deux,cinquante-deux.DET+Dnum:mp:fp
cinquante-et-un,cinquante-et-un.DET+Dnum:mp
cinquante-mille,cinquante-mille.DET+Dnum:mp:fp
cinquante-neuf,cinquante-neuf.DET+Dnum:mp:fp
cinquante-trois,cinquante-trois.DET+Dnum:mp:fp
deux-cent-mille,deux-cent-mille.DET+Dnum:mp:fp
deux-cent-quatre-vingt-un,deux-cent-quatre-vingt-
un.DET+Dnum:mp
deux-cents,deux-cents.DET+Dnum:mp:fp

deux-cents-dix-neuf,deux-cents-dix-
neuf.DET+Dnum:mp:fp

Dans notre graphe, nous avons permis toutes les variations graphiques possibles (à la différence de celui d'Éric Laporte qui décrivait la norme) en ce qui concerne le pluriel : *quatre-vingts* vs *quatre-vingt*, *cent* vs *cents*, etc., indépendamment de la place du mot dans le composé. On voit l'intérêt de procéder ainsi, puisque dans la liste ci-dessus on identifie *deux-cent-quatre-vingt-un* (sans *s* à *cent*) mais aussi *deux-cents-dix-neuf* (avec *s* à *cent*).

4.2.3.6 Les déterminants relatifs

Les déterminants relatifs « n'appartiennent qu'à la langue écrite, surtout juridique et parfois littéraire » (BU § 600). Selon toute vraisemblance, on a donc peu de chance de les rencontrer dans des productions orales spontanées.

Les formes des déterminants relatifs sont les mêmes que celles des pronoms relatifs construits avec *quel* : *lequel*, *auquel*, *duquel*, etc.

On a donc ainsi pour chacune de ces formes une ambiguïté entre la catégorie déterminant et la catégorie pronom. Étant donnée la probabilité très faible que l'on a de les rencontrer en tant que déterminants dans notre corpus, nous avons préféré ne pas multiplier les ambiguïtés et n'avons donc pas inséré ces formes comme déterminants dans le dictionnaire.

4.2.3.7 Les déterminants exclamatifs et interrogatifs : DET+Dexi

Les formes concernées par ces deux catégories de déterminants sont *quel*, *quelle*, *quels* et *quelles*. Considérées comme interrogatives dans le cas de questions, et comme exclamatives « quand on exprime un sentiment vif » (BU § 603).

Nous avons créé une catégorie mixte pour ces formes, qui pourra être scindée éventuellement lors de la relecture du corpus étiqueté.

4.2.3.8 Les conjonctions

Les conjonctions sont des mots invariables qui se répartissent en deux sous-catégories aux propriétés syntaxiques distinctes : les conjonctions de coordination et les conjonctions de subordination.

4.2.4 Les conjonctions de subordination : CONJS

Les conjonctions de subordination servent « à unir deux éléments de fonctions différentes, dont l'un est une proposition (sujet ou complément) » (BU § 1023).

On distingue classiquement les propositions complétives (introduites par *que*) et les propositions circonstancielles.

On a quelques formes simples de conjonctions (*comme, lorsque, puisque, quand, que, quoique, si, sinon*), mais la plupart des conjonctions de subordination sont des formes composées. La composition a en général la forme suivante (BU § 1025) :

- syntagme nominal + *que* : *à mesure que, à condition que, de crainte que*, etc.
- adverbe + *que* : *bien que, encore que, alors que*, etc.
- préposition + *que* : *pour que, dès que, après que*, etc.
- forme verbale + *que* : *vu que, attendu que*, etc.

Certaines de ces conjonctions sont discontinues : *tout...que, si...que, quelque...que*, etc. D'autres, dont les composants sont la plupart du temps contigus, peuvent se rencontrer avec une insertion entre les composants et se présenter également de façon discontinue, comme le montrent les exemples suivants issus du journal *Le Soir* (2001) :

Mais les avis sont partagés sur la question. Introduire des caméras équivaldrait à suspecter la mère de maltraitance, **alors même que** la confiance est le fondement de la relation au pédiatre.

Footballeur n°1 dans son pays pour l'année 2000, quelques jours **avant seulement que** tout frais équipier de Dortmund, Jan Koller, ne soit couronné Soulier d'or en Belgique, Tomas Rosicky sera le véritable chef d'orchestre de la Tchéquie, ce mercredi soir.

À l'heure actuelle, nous n'avons pas de procédure qui nous permettrait de tenir compte des formes de conjonctions discontinues. Chaque membre est donc analysé de façon séparée (cf. ci-dessous).

4.2.5 Les conjonctions de coordination : CONJC

À la différence des conjonctions de subordination, les conjonctions de coordination sont « chargées d'unir des éléments de même statut : soit des phrases ou des sous-phrases, – soit, à l'intérieur d'une phrase, des éléments de même fonction » (BU § 1029).

On recense traditionnellement six conjonctions de coordination : *car, et, mais, ni, or* et *ou*.

Conjonctions de coordination, de subordination et adverbes conjonctifs (*donc, pourtant, néanmoins, cependant*) sont également regroupés sous le terme générique de *connecteurs* (Piot 1993), et l'on rencontre souvent dans les typologies des grammaires des glissements d'une catégorie à l'autre (le *donc*, souvent classé parmi les conjonctions de coordination).

Nous maintenons les trois catégories distinctes, sans séparer les adverbes conjonctifs des autres adverbes.

4.2.6 Les prépositions : PREP

Les prépositions sont des mots invariables qui régissent des syntagmes nominaux (ou équivalents, tels des pronoms), ou verbaux (verbe à l'infinitif ou au participe).

Certaines prépositions font exception à cette règle distributionnelle et peuvent n'être pas suivies d'un syntagme nominal. C'est le cas des prépositions *après, avant, avec, contre, depuis, derrière, devant, pour* et *sans*, qui sont dans ce cas traditionnellement classées parmi les adverbes. Nous ne considérons pas cette recatégorisation en adverbe mais conservons ces formes uniquement parmi les prépositions. Elles reçoivent néanmoins un code supplémentaire dans le dictionnaire (*eff*, pour *effacement*) qui nous permet de traiter leur distribution différemment lors de l'écriture de grammaires de levée d'ambiguïtés.

Mise à part cette distinction, qui nous évite de considérer ces formes comme grammaticalement ambiguës, nous ne faisons pas de sous-catégorisation parmi les prépositions.

Comme pour les conjonctions, outre les quelques formes simples, un grand nombre de prépositions sont composées et nous les avons intégrées dans le dictionnaire (*autour de, en train de, face à*, etc.). Pour certaines d'entre elles, il est possible qu'il y ait contraction de la préposition et du déterminant. Le code *Prepdet* est alors ajouté (cf. ci-dessus pour les déterminants définis), en plus du code *PREP*. Pour ces exemples, la préposition peut donc recevoir un code flexionnel, comme le montrent les entrées suivantes :

grâce à,.PREP
grâce au,grâce à le.PREP+Prepdet:ms
grâce aux,grâce à le.PREP+Prepdet:mp:fp

Comme les conjonctions, certaines prépositions nous confrontent au problème de la discontinuité (*grâce seulement à, autour même de*, etc.), pour

laquelle, nous l'avons dit, nous n'avons pas de solution satisfaisante. Les composants seront donc étiquetés individuellement.

4.2.7 Les interjections : INTJ

La catégorie des interjections comprend des formes invariables, qui ont la particularité d'être syntaxiquement autonomes au sein de la phrase.

Parmi les interjections classiquement retenues dans les grammaires et les dictionnaires, on a *ouste*, *parbleu*, *pardi*, *hélas*, *saperlipopette*, *hein*, etc. Dans des textes écrits standard, ces formes qui notent une émotion (surprise, joie, colère, etc.) sont en général suivies d'un point d'exclamation. Si les interjections que nous venons de citer sont classiques, toute forme peut potentiellement devenir une interjection. On assiste alors à une recatégorisation grammaticale, et celle-ci peut concerner aussi bien des noms ou des syntagmes nominaux (*ma parole ! mon dieu ! respect !*), des adjectifs (*joli !*), des adverbes (*comment !*), des verbes (*gare !, tiens tiens !, soit !*) que des formes mixtes (*ça mon dieu !, dis donc !, tu parles !, mon œil !, sans blague !*), pour reprendre des exemples lexicalisés.

Nous avons conservé les interjections classiques, mais y avons néanmoins ajouté certaines, propres à la production orale. En effet, au vu des exemples suivants, on peut difficilement analyser *allez* comme un verbe ou *bon* comme un adjectif :

chaHI1 ah parce que les familles étaient plus liées // toutes les familles étaient euh / allez / le fils le beau-fils la fille la (x) les nièces et tout le monde était à l'enterrement / que maintenant / peut-être aussi mais moins quand même (silence)[chaHI1r]

souTJ1 ah ben ouais hein / bon allez au revoir [souVC1r]

ilrLD3 c'est lié déjà aux parents / ça aussi ça fait beaucoup / bon si on a des parents / ça existe c'est malheureux à dire bon moi j' ai des copains que ça / leurs parents bon c'est des alcoolos ou des trucs comme ça quoi bon euh chez eux c'est toujours des bagarres euh entre ses parents euh toujours des mots grossiers et tout bon ils n'entendent que ça et tout / bon ils en ont marre qu'est-ce qu' ils font bon moi je me dispute tout le temps avec mes parents mais mes parents ils ne je n'ai rien à leur reprocher quoi bon / ils m'ont bien élevé ils m'ont toujours appris à être poli dire bonjour au revoir vraiment euh bon // toujours bon euh les règles de politesse euh primaires ils me les ont vraiment a/ bien éduquées et tout ce qu' il faut quoi enfin bon moi je me dispute [ilrLD3r]

On le voit dans l'extrait ci-dessus, le locuteur utilise *bon* comme un véritable ponctuant : celui-ci vient rythmer l'énoncé et le scinde en segments cohérents, syntaxiquement et sémantiquement. Dans l'ensemble du corpus, on a en fait très peu d'occurrences de *bon* qui sont des adjectifs, la plupart servent de ponctuant.

Nous regroupons dans la catégorie des interjections les onomatopées, formes particulièrement fréquentes à l'oral et prisées notamment par les illustrateurs de bandes dessinées. Parmi celles non répertoriées par le DELAF et présentes dans notre corpus, nous relevons entre autres : *bè, gnagnagna, m, mm, oula, pf, pt, waf, wouf, t, grr, hops, miam, ouh, vroum, oh la la, toup*. Dans ce cas, une multitude de formes différentes existent, non prédictibles, qui doivent être enregistrées dans le dictionnaire.

Mouais et *moui*, absents du DELAF, ont été catégorisés, au même titre que *oui* dont ils sont des variantes, comme des adverbes. Nous avons procédé de même pour *ouais*, qui était classé comme interjection dans le DELAF, et que nous considérons comme un adverbe.

4.3 La recatégorisation

Par recatégorisation (Kerleroux 1996), on entend le phénomène par lequel un mot ayant une classe grammaticale dans le lexique peut, en discours, changer de classe.

Ainsi, le déterminant a la faculté de recatégoriser n'importe quel mot qui le suit en substantif.

On l'a vu ci-dessus, notamment dans la section consacrée aux interjections, n'importe quel mot peut potentiellement, en discours, être recatégorisé. Cela est fréquent dans les usages autonymiques, comme l'illustre l'exemple suivant où la locutrice s'excuse de ses « bafouillages » :

iljDV1 voilà désolée -| pour les euh les ah les euh
euh [iljDV1r]

Si l'on pouvait prendre en compte la recatégorisation grammaticale, les interjections *euh* et *ah* seraient étiquetées « nom ».

Ces occurrences sont en contradiction avec le lexique. Nous n'avons pas tenu compte de tels phénomènes dans ce travail, qui se base étroitement sur la description du lexique faite dans les dictionnaires électroniques.

4.4 Tableau synoptique du jeu d'étiquettes

Le tableau suivant synthétise les catégories et sous-catégories que nous prenons en compte dans ce travail, ainsi que les informations flexionnelles

associées à chacune d'elles. Il s'agit donc du jeu d'étiquettes qui seront présentes dans le texte annoté.

La première colonne reprend la catégorie (classement alphabétique), la deuxième le code, la troisième colonne indique les sous-catégories éventuelles, dont le code est donné également dans la deuxième colonne ; la dernière colonne reprend les codes flexionnels associés à ces catégories et sous-catégories. les codes flexionnels sont détaillés dans le tableau de la section suivante.

catégorie grammaticale	code de catégorie grammaticale	sous-catégorie grammaticale	code de sous-catégorie grammaticale	codes associés à une sous-catégorie	codes flexionnels
adjectif	A				m, f, s, p
adverbe	ADV				
conjonction de coordination	CONJC				
conjonction de subordination	CONJS				
déterminant	DET				m, f, s, p
		déterminant défini	Ddef	Prepdet	m, f, s, p
		déterminant indéfini	Dind		m, f, s, p
		déterminant démonstratif	Ddem		m, f, s, p
		déterminant possessif	Dpos	Dpos1s, Dpos2s, Dpos3s, Dpos1p, Dpos2p, Dpos3p	m, f, s, p
		déterminant numéral	Dnum		m, f, s, p
		déterminant interrogatif ou exclamatif	Dexi		m, f, s, p
interjection	INTJ				
nom commun	N				m, f, s, p
nom propre	NPr				
préfixe	PFX				
préposition	PREP				

			PREP+eff		
			PREP+Prepdet		m, f, s, p
			PREP+Preppro		m, f, s, p
pronom	PRO				m, f, s, p
		pronom personnel	Ppers		m, f, s, p
		pronom relatif	Prel	Preppro	m, f, s, p
		pronom démonstratif	Pdem		m, f, s, p
		pronom possessif	Ppos	Ppos1s, Ppos2s, Ppos3s, Ppos1p, Ppos2p, Ppos3p Preppro	m, f, s, p
		pronom indéfini	Pind	Preppro	m, f, s, p
verbe	V				1, 2, 3 s, p B, C, F, G, I, J, K, P, S, T, Y, W
constituant non autonome	XI				

Tableau 39. Tableau synoptique du jeu d'étiquettes

Les codes flexionnels

Nous présentons dans le tableau ci-dessous les codes flexionnels présents dans nos données. Comme nous l'avons vu dans le tableau ci-dessus, ces codes ne sont pas présents pour toutes les catégories grammaticales.

Les codes de genre *m* et *f* accompagnent les catégories grammaticales du nom, de l'adjectif, du déterminant et du pronom, ainsi que les prépositions contractées.

L'information flexionnelle de nombre (*s* et *p*) est associée, comme celle du genre (*m* et *f*), au nom, à l'adjectif, au déterminant, au pronom et aux prépositions contractées, mais également à la personne du verbe tensé à un temps fini.

Les codes de personne *1*, *2* et *3* sont associés aux entrées des pronoms et des verbes.

B, *C*, *F*, *G*, *I*, *J*, *K*, *P*, *S*, *T*, *Y* et *W* sont les douze codes flexionnels se rapportant exclusivement au verbe²².

²² *B* note l'ambiguïté entre le participe passé et l'adjectif, mais ce code est associé à celui du verbe et non à celui de l'adjectif.

code flexionnel	signifié
m	masculin
f	féminin
s	singulier
p	pluriel
1	1 ^{re} personne
2	2 ^e personne
3	3 ^e personne
B	participe passé ou adjectif
C	conditionnel présent
F	futur simple
G	participe présent
I	indicatif imparfait
J	passé simple
K	participe passé
P	indicatif présent
S	subjonctif présent
T	subjonctif passé
Y	impératif présent
W	infinitif présent

Tableau 40. Codes flexionnels

4.5 Mots simples / mots composés

Le question de la composition et du figement est un problème vaste, au cœur des recherches actuelles en linguistique et en traitement automatique des langues, où l'on a mis en évidence le continuum qui existe entre suite libre et suite figée (Bosredon et Tamba 1991, Danlos 1988, Dister *et al.* 2007a, Gaston Gross 1996a et 1996b, Maurice Gross 1982, Klein et Lamiroy 1995a et 1995b, Lamiroy *et al.* 2003, Laporte 1988, Mejri 2000, 2002 et 2003, Moon 1998, Silberztein 1990).

Néanmoins, cette pierre d'achoppement du TAL est souvent passée sous silence, sauf dans les cas où elle constitue le sujet principal de la recherche. Nous ne ferons guère exception à cette règle, même si cette question a été au cœur de notre réflexion tout au long du choix de nos ressources linguistiques.

Pour ce qui est des mots graphiquement composés, dont l'un des composés est un constituant non autonome (c'est-à-dire qu'il n'apparaît pas indépendamment de la composition : le cas de *parce* dans *parce que*, *priori* dans *a priori*, etc.), il va sans dire qu'il est nécessaire de les recenser. Ils ne posent d'ailleurs pas de problème dans la mesure où l'on n'hésite pas entre la suite libre et le composé.

Car c'est véritablement là que réside le problème : l'ambiguïté entre suite libre et suite composée. À l'heure actuelle, bien que constituant une ressource très riche, les données du LADL sont incomplètes en ce qui concerne les mots composés. En les utilisant lors de l'étiquetage, nous ne pouvons donc prétendre à l'exhaustivité. De plus, les choix multiples de découpage générés par la prise en compte des formes composées multiplient considérablement le nombre des ambiguïtés, puisqu'à côté de la suite libre, on a l'hypothèse d'une analyse de mot composé. Pour prendre l'exemple franco-français²³ *carte bleue* (« carte de crédit » vs « carte de couleur bleue »), on a deux chemins dans l'automate du texte :

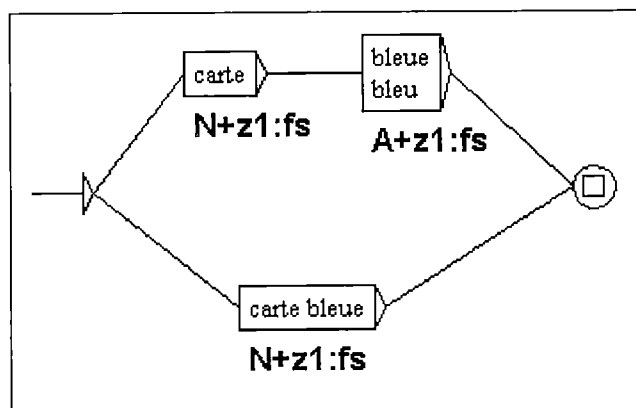


Figure 22. Automate de carte bleue

Le chemin du dessus représente la suite libre tandis que le chemin du dessous est l'analyse de la séquence en un mot composé.

²³ Le composé n'est pas utilisé en Belgique francophone.

Dans un cas de ce type, il est impossible de résoudre cette ambiguïté de manière automatique, d'autres informations que celles qui relèvent de la morphosyntaxe étant nécessaires.

Ainsi, incomplétude des données et difficulté de les traiter font que nous avons, dans un premier temps, décidé de nous passer des noms et des adjectifs composés.

En fait, le fait de reporter l'intégration de ces formes composées dans le système, lorsque leur recensement sera complet, ne remet pas en cause, dans l'immédiat, le travail effectué : si le *cordon bleu*, lorsqu'il est attesté dans le corpus avec le sens de « bon cuisinier », n'est pas un « cordon de couleur bleue », il n'en reste pas moins que cette suite, même figée, est composée d'un nom et d'un adjectif. Tout comme le composé prototypique *pomme de terre* est bien, dans sa composition interne, la suite nom + préposition + nom. Ainsi, en attendant la mise à disposition de ressources lexicales complètes (autant que faire se peut), nous ne considérerons pas ces formes composées.

En revanche, nous l'avons vu ci-dessus, nous avons choisi d'intégrer les composés pour les catégories relevant des mots grammaticaux. Ainsi, nous avons des dictionnaires de prépositions composées, de conjonctions, etc. Ceux-ci demanderont sans doute à être complétés, et nous profiterons de la correction des résultats de la désambiguïsation pour le faire.

En fait, la prise en compte des formes composées nous confronte directement à deux problèmes :

1. l'augmentation du nombre des ambiguïtés. Dans les systèmes de levée d'ambiguïtés qui tiennent compte des mots composés, il est fréquent de privilégier l'hypothèse de la composition face à une suite libre (cf. Clément et le *Corpus de Paris* 7 : 40),
2. la discontinuité de certains composés, qui, dans la procédure actuelle, ne sont pas reconnus. Nous avons eu l'occasion de l'illustrer avec des exemples issus du français écrit, mais le phénomène est encore accentué dans nos transcriptions. Ces insertions peuvent être liées à l'oralité du corpus. Ainsi, l'énoncé suivant qui concerne la préposition *par rapport aux* :

ileGF0 euh à propos des des journalistes et et **par rapport euh aux** journalistes français euh est-ce que vous vous estimez que les journalistes français s'expriment mieux / que les journalistes belges / ou bien pas [ileFN1r]

On le voit, le morphème *euh* est inséré à l'intérieur de la préposition composée. Mais outre la possibilité d'insertion de ce morphème propre à l'oral, on a aussi des insertions liées à nos conventions de transcription : la

marque de la pause, d'un début ou d'une fin de chevauchement, etc., peuvent en effet découper graphiquement le composé dans le texte transcrit.

C'est le cas de cette conversation entre amies où les chevauchements sont très nombreux. Le problème ici est que la reconnaissance des déterminants numéraux échoue à cause de l'insertion de la marque de chevauchement :

blaAD1 oui mais -| attends / je l'ai trouvée j'ai pas dit
que je l'achetais
blaNB1 ah oui / oui / |- oui
blaAD1 quatre-cent -| soixante-mille euros
blaNB1 attends ça fait c/
blaAD1 dix-huit millions et une chique [blaNB1]

Dans ce cas, notre graphe dictionnaire reconnaît une suite de deux déterminants : *quatre-cents* et *soixante-mille*.

Certaines de ces marques peuvent également avoir été introduites lors du prétraitement du corpus. C'est le cas dans l'exemple suivant, où les balises de répétitions *<r>* et *</r>* empêchent la reconnaissance de la conjonction composée à cause de (cf. chap. 3 et 4)²⁴ :

blaJV1 (...) et je suis un consommateur et ça me fait
chier quoi parce que / à cause <r> de </r> de cons
qui ont abusé / ben moi je ne peux plus euh (silence)
fumer mon stick euh pénard après une réunion de staff
ou quoi (...) [blaJV1]

Baliser le deuxième *de* et non le premier *de* permettrait de reconnaître le composé dans ce cas-ci. Ce choix de conserver dans le texte le dernier répété a été fait en fonction du processus de la répétition qui vise à garder le *repair* (cf. chap. 3). Si l'autre solution – qui consisterait ici à garder le répétable, c'est-à-dire le premier *de*, et à baliser le deuxième *de* – aurait permis ici de reconnaître le composé, elle aurait posé le même problème dans d'autres cas, comme le montre l'exemple suivant :

ileDC1 (...) et alors le je pense que de plus en plus <r>
à </r> à cause des médias notamment à cause de la
télévision / on arrive de plus en plus à je dirais à
travailler non plus alors son accent / mais à travailler la
façon de communiquer / hein [ileDC1r]

²⁴ Il ne s'agit en fait que d'un problème technique et on pourrait imaginer que le programme de consultation des dictionnaires puisse reconnaître des formes composées dans le texte en ignorant certaines insertions, comme le font les grammaires d'ELAG pour les étiquettes *IGN* (cf. chap. 4 et 6).

Dans cet énoncé, on reconnaît le composé à *cause des*, ce qui ne serait pas le cas si l'on avait balisé le deuxième *de*. On voit ici que face à ce problème, il n'y a pas de solution idéale. Dans ces cas, chaque forme sera analysée séparément.

4.6 Conclusions sur nos ressources lexicales

Les ressources lexicales que nous utilisons sont formalisées dans des dictionnaires à large couverture. Dans notre processus d'annotation, nous sommes tributaire de ces données puisque ce sont elles qui établiront si une forme est ambiguë, c'est-à-dire si elle reçoit plus d'une analyse dans le dictionnaire. Nous avons choisi la version de dictionnaire la plus « simple », celle qui ne présente pas de code syntaxique pour les emplois différents des verbes. Procédant de la sorte, on diminue déjà le nombre d'ambiguïtés. Nous avons fait de même en supprimant l'ambiguïté adjectif *vs* participe passé qui affectait un grand nombre d'entrées.

Dans un processus inverse, nous avons affiné les données qui concernent les mots grammaticaux en ajoutant de nouveaux codes qui nous permettent de tenir compte des sous-catégories (de pronoms, de déterminants). Si nous ne pouvons tenir compte des changements de catégorie en contexte, nous avons tout de même augmenté le nombre d'interjections dans le dictionnaire, afin de mieux représenter ces marques de l'oral présentes dans notre corpus.

Les choix que nous avons faits pour la classification lexicale peuvent être critiqués à plus d'un égard : très souvent, ils suivent la classification traditionnelle de la grammaire scolaire. C'est que les résultats de l'étiquetage doivent être réutilisables par des chercheurs qui ne sont pas engagés dans une théorie linguistique particulière.

Bien que très complètes, les ressources linguistiques que nous venons de décrire s'avèrent insatisfaisantes pour analyser des corpus ayant certaines particularités que nous pourrions qualifier de « non standard ». C'est le cas du français *parlé*, que nous avons déjà abordé, mais aussi des textes du français hors de France. En effet, les particularités lexicales du français en Belgique ne sont pas répertoriées dans ces dictionnaires élaborés pour le français hexagonal (Fairon 1998, Dister et Fairon 2004). Il s'est donc avéré nécessaire de construire un dictionnaire électronique du français en Belgique (DELFRABEL), que nous allons brièvement décrire.

4.7 Le Delfrabel : le dictionnaire électronique du français en Belgique

Dans cette section, nous ne tenterons pas de donner une définition stricte des *belgicisms* (Piron 1973, Goosse 1977), ni de classer ces particularités

lexicales selon leur statut (les statalismes de Pohl 1984), leur origine (Goosse 1995), leurs référents (Piron 1979, Klein et Lenoble-Pinson 1997) ou encore de les comparer avec celles d'autres variétés de français (Lamiroy *et al.* 2003). Les publications sur le sujet ne manquent pas.

Nous allons décrire l'origine de nos données, et la formalisation que nous en avons faite dans un dictionnaire électronique de type DELAF.

4.8 Les particularités lexicales du français de Belgique : enquête de vitalité et de diffusion

Si l'intérêt pour ce que l'on appelle communément les *belgicisms* s'est manifesté depuis longtemps, la plupart des travaux publiés consistent bien souvent en des répertoires à visée prescriptive et normative de type : « Ne dites pas mais dites. » Pour leurs auteurs, il s'agit le plus souvent de faire la *Chasse aux belgicisms* (Hanse *et al.* 1971 et 1974) ou de dégager les *Belgicisms de bon aloi* (Doppagne 1979). Si les recherches les plus récentes s'écartent de cet aspect normatif, elles donnent néanmoins l'impression d'une grande homogénéité des pratiques : les belgicisms seraient répandus à l'ensemble de la Belgique francophone, terre périphérique sans variation linguistique. De la même manière que la variante géographique est évacuée, les variations sociolinguistiques sont largement ignorées : les belgicisms semblent donc non seulement répandus à l'ensemble de la Belgique francophone, mais aussi à l'ensemble de ses habitants, sans distinction aucune. Comme le constatent Francard *et al.* (2002 : 14) :

Au lieu de favoriser une perception adéquate de la variation linguistique en Belgique francophone, elles [les cacographies sur les belgicisms] contribuent à faire des « particularités » du français en Belgique un domaine dont l'apparente homogénéité n'est que le résultat d'une méthodologie sommaire.

Afin de combler ces lacunes et de caractériser de manière plus rigoureuse les particularités lexicales du français de Belgique, le centre de recherche VALIBEL a mené une vaste enquête dont le but est de dégager précisément la diffusion et la vitalité des particularités lexicales du français de Belgique. C'est sur base de cette enquête que les formes qui figurent dans notre dictionnaire ont été retenues.

L'enquête est décrite dans Francard *et al.* (2002). Elle se base sur un réseau d'informateurs répartis en 20 points d'enquête (qui correspondent à six zones), et échantillonnés selon le sexe, l'âge (trois tranches) et le profil socioculturel (quatre degrés de scolarité). L'un des buts de cette enquête est de servir de base à la nomenclature du *Dictionnaire du français en Belgique*, actuellement en cours de rédaction (Francard *et al.* 2003).

Dans la mesure où ni les résultats de l'enquête, ni le dictionnaire n'ont encore été publiés, nous ne détaillerons pas le contenu de nos données.

Aux données issues des recherches menées au Centre VALIBEL, nous avons aussi ajouté certaines formes, principalement des adjectifs issus de noms de localités belges (*namurois*, *carolo*, etc.).

Sans établir la liste des entrées du DELFRABEL, nous allons décrire son mode de constitution.

4.9 Constitution du DELFRABEL

4.9.1 Les entrées construites manuellement

Le DELFRABEL adopte les codes et le formalisme des dictionnaires du LADL (cf. § 3.1) et est compatible avec le système UNITEX.

Outre les indications standard de lemme, de catégorie grammaticale et de flexion, toutes les entrées du DELFRABEL sont accompagnées du nom du dictionnaire. Ceci permet d'identifier que la forme est reconnue par l'application de ce dictionnaire. Un équivalent « en français de France » est également indiqué en fin de ligne après une barre oblique. Voici l'entrée pour la forme *guindailleur* :

guindailleur,.N:ms+Hum/fêtard+Delfrabel

Notre dictionnaire comprend des mots simples et des mots composés, selon la distinction formelle établie ci-dessus (Silberztein 1990 et 1993, Gross 1990).

Voici quelques entrées du DELFRABEL :

à bras,.ADV/dans les bras+Delfrabel
 à disposition de,.PREP/à la disposition de+Delfrabel
 à dos,.ADV/sur le dos+Delfrabel
 à fond de balle,.ADV/à fond de train+Delfrabel
 à force que,.CONJS/à force de+Delfrabel
 à la toilette,.ADV/aux toilettes+Delfrabel
 bistouille,.N:fs/mélange de café et d'eau-de-vie+Delfrabel
 bistouilles,bistouille.N:fp/mélange de café et d'eau-de-vie+Delfrabel
 bulletin de virement,.N:ms/mandat de virement+Delfrabel
 bulletins de virement,bulletin de virement.N:mp/mandats de virement+Delfrabel
 guindaille,.N:fs/fête, beuverie+Delfrabel
 guindailles,guindaille.N:fp/fête, beuverie+Delfrabel
 guindailleur,guindailleur.N:ms+Hum/fêtard+Delfrabel

guindailleurs, guindailleur.N:mp+Hum/fêtard+Delfrabel
 guindailleuse, guindailleur.N:fs+Hum/fêtard+Delfrabel
 guindailleuses, guindailleur.N:fp+Hum/fêtard+Delfrabel
 satis, satisfaction.N:fs:fp/mention minimum pour une réussite aux
 examens+Delfrabel
 satisfaction, satisfaction.N:fs/mention minimum pour une réussite aux
 examens+Delfrabel
 satisfactions, satisfaction.N:fp/mention minimum pour une réussite aux
 examens+Delfrabel

Les entrées qui viennent d'être présentées ont été introduites manuellement. Nous allons voir maintenant comment il est possible de générer automatiquement, à partir d'une forme canonique, toutes les formes fléchies qui s'y rapportent.

4.9.2 La flexion automatique des verbes

En effet, s'il n'est pas très coûteux, pour un nombre limité de mots, de noter manuellement les différentes formes fléchies d'un lemme (ce que nous avons fait pour les noms et les adjectifs), ce travail s'avère beaucoup plus fastidieux lorsqu'à la forme canonique est associé un grand nombre de formes fléchies. C'est le cas des verbes.

Pour pallier cet inconvénient, UNITEX permet de générer de manière automatique les formes fléchies, au moyen de graphes de flexion.

Ainsi, le graphe suivant permet d'obtenir automatiquement le futur des verbes du type de *blinquer* « reluire, astiquer » (verbes du 1^{er} groupe).

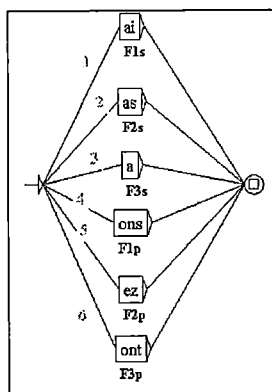


Figure 23. Graphe de flexion du futur

Il faut lire ce graphe de la manière suivante : quand on parcourt le premier chemin du graphe (chemin noté 1), on ajoute *ai* à la forme donnée

en entrée (ici donc le verbe *blinquer* qui signifie « astiquer » ou « reluire ») ainsi que l'information *F:1s*, c'est-à-dire futur à la 1^{re} personne du singulier. On obtient alors dans un fichier en sortie :

blinquerai,blinquer.V:F1s

De la même manière, on parcourt ensuite le deuxième chemin du graphe (chemin noté 2), on ajoute *as* à la forme donnée en entrée ainsi que l'information *F:2s* (futur à la 2^e personne du singulier). On obtient alors dans un fichier en sortie :

blinqueras,blinquer.V:F2s

Et ainsi de suite afin d'obtenir toute la conjugaison du futur :

blinquerai,blinquer.V:F1s
blinqueras,blinquer.V:F2s
blinquera,blinquer.V:F3s
blinquerons,blinquer.V:F1p
blinquerez,blinquer.V:F2p
blinqueront,blinquer.V:F3p

Évidemment, la conjugaison des verbes n'est pas toujours aussi simple. Si, au futur et au conditionnel présent, il s'agit simplement d'ajouter une terminaison à la forme de l'infinitif, il n'en va pas de même pour la conjugaison de l'indicatif ou du subjonctif présent.

Ainsi, pour obtenir la forme de la 1^{re} personne du présent de l'indicatif *blinque*, il faut enlever la dernière lettre du lemme. De même, la forme de la 2^e personne du présent *blinques* est obtenue par suppression de la dernière lettre du lemme et par l'ajout d'un *s* final. Il est possible de noter ces opérations de suppression grâce au symbole formel L.

Voici le graphe de flexion du présent de l'indicatif.

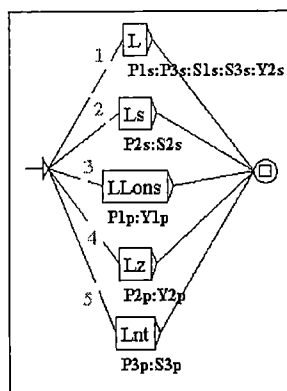


Figure 24. Graphe de flexion du présent

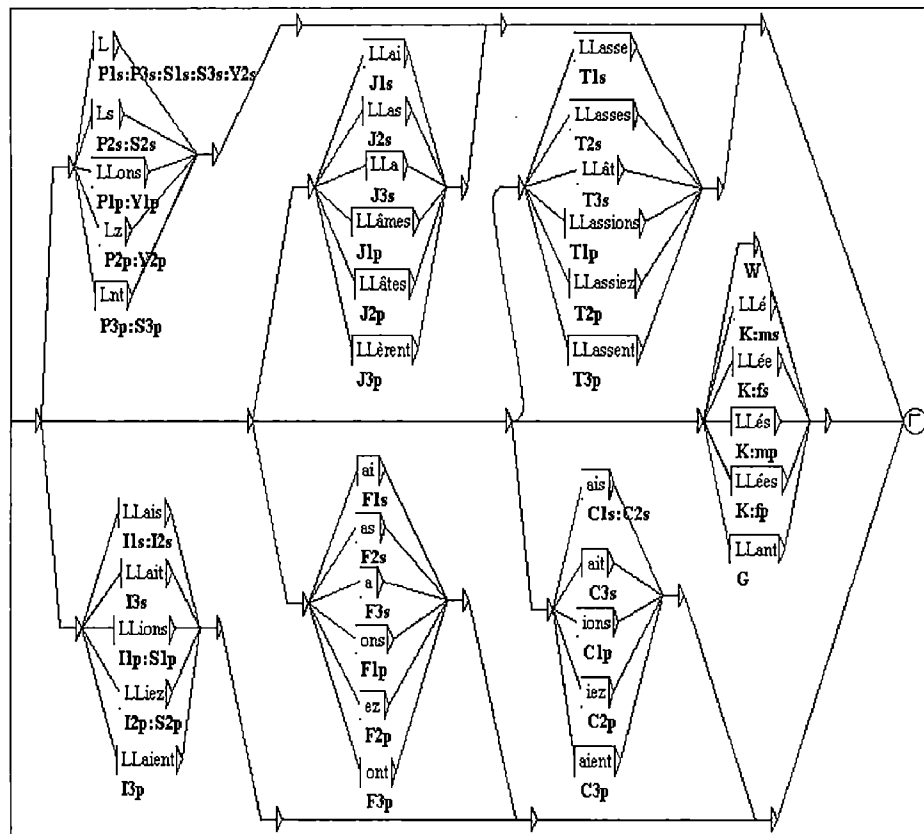
Le chemin 1 effectue l'opération d'effacement de la dernière lettre avec la commande L. On a donc la forme *blinque*, qui correspond aux 1^{re} et 3^e personnes de l'indicatif présent (P) et du subjonctif présent (S) et à la 2^e personne de l'impératif (Y).

Le chemin 3 efface les deux dernières lettres du lemme (avec LL) et ajoute ensuite la suite de lettres *ons* pour donner la forme *blinquons*, 1^{re} personne du pluriel de l'indicatif présent (P) et de l'impératif (Y).

Avec ce graphe, nous obtenons donc les formes suivantes :

blinque,blinquer.V:P1s:P3s:S1s:S3s:Y2s
 blinques,blinquer.V:P2s:S2s
 blinquons,blinquer.V:P1p:Y1p
 blinquez,blinquer.V:P2p:Y2p
 blinquent,blinquer.V:P3p:S3p

Voici le graphe général, qui permet de générer toutes les formes fléchies à partir du lemme *blinquer*.

Figure 25. Graphe de flexion des verbes du 1^{er} groupe

À partir d'une forme canonique, on génère automatiquement toutes les formes fléchies, au nombre de 39, qui correspondent à 51 analyses :

blinuai,blinquer.V:J1s/reluire ou astiquer+Delfrabel
 blinquas,blinquer.V:J2s/reluire ou astiquer+Delfrabel
 blinqua,blinquer.V:J3s/reluire ou astiquer+Delfrabel
 blinquâmes,blinquer.V:J1p/reluire ou astiquer+Delfrabel
 blinquâtes,blinquer.V:J2p/reluire ou astiquer+Delfrabel
 blinquèrent,blinquer.V:J3p/reluire ou astiquer+Delfrabel
 blinquerai,blinquer.V:F1s/reluire ou astiquer+Delfrabel
 blinqueras,blinquer.V:F2s/reluire ou astiquer+Delfrabel
 blinquera,blinquer.V:F3s/reluire ou astiquer+Delfrabel
 blinquerons,blinquer.V:F1p/reluire ou astiquer+Delfrabel
 blinquerez,blinquer.V:F2p/reluire ou astiquer+Delfrabel
 blinqueront,blinquer.V:F3p/reluire ou astiquer+Delfrabel
 blinquasse,blinquer.V:T1s/reluire ou astiquer+Delfrabel
 blinquasses,blinquer.V:T2s/reluire ou astiquer+Delfrabel
 blinquât,blinquer.V:T3s/reluire ou astiquer+Delfrabel
 blinquassions,blinquer.V:T1p/reluire ou astiquer+Delfrabel

blinquassiez,blinquer.V:T2p/reluire ou astiquer+Delfrabel
 blinquassent,blinquer.V:T3p/reluire ou astiquer+Delfrabel
 blinquerais,blinquer.V:C1s:C2s/reluire ou astiquer+Delfrabel
 blinquerait,blinquer.V:C3s/reluire ou astiquer+Delfrabel
 blinquerions,blinquer.V:C1p/reluire ou astiquer+Delfrabel
 blinqueriez,blinquer.V:C2p/reluire ou astiquer+Delfrabel
 blinqueraient,blinquer.V:C3p/reluire ou astiquer+Delfrabel
 blinquer,blinquer.V:W/reluire ou astiquer+Delfrabel
 blinqué,blinquer.V:K:ms/reluire ou astiquer+Delfrabel
 blinquée,blinquer.V:K:fs/reluire ou astiquer+Delfrabel
 blinqués,blinquer.V:K:mp/reluire ou astiquer+Delfrabel
 blinquées,blinquer.V:K:fp/reluire ou astiquer+Delfrabel
 blinquant,blinquer.V:V:G/reluire ou astiquer+Delfrabel
 blinquais,blinquer.V:I1s:I2s/reluire ou astiquer+Delfrabel
 blinquait,blinquer.V:I3s/reluire ou astiquer+Delfrabel
 blinquions,blinquer.V:I1p:S1p/reluire ou astiquer+Delfrabel
 blinquiez,blinquer.V:I2p:S2p/reluire ou astiquer+Delfrabel
 blinquaient,blinquer.V:I3p/reluire ou astiquer+Delfrabel
 blinque,blinquer.V:P1s:P3s:S1s:S3s:Y2s/reluire ou astiquer+Delfrabel
 blinques,blinquer.V:P2s:S2s/reluire ou astiquer+Delfrabel
 blinquons,blinquer.V:P1p:Y1p/reluire ou astiquer+Delfrabel
 blinquez,blinquer.V:P2p:Y2p/reluire ou astiquer+Delfrabel
 blinquent,blinquer.V:P3p:S3p/reluire ou astiquer+Delfrabel

Concrètement, la méthode est la suivante. On construit un dictionnaire de lemmes (dans notre cas, ces lemmes sont uniquement des verbes, à l'infinitif)²⁵. Chacun de ces lemmes est associé à un code de flexion qui correspond à un graphe de flexion. Le graphe que nous avons présenté ici est le graphe de flexion des verbes du 1^{er} groupe, et tous les verbes de ce groupe recevront donc, dans le dictionnaire de lemmes, le même code de flexion que le verbe *blinquer*. Il y a autant de codes de flexion (et donc de graphes de flexions associés) que de verbes ayant des flexions différentes. Ensuite, ce dictionnaire est fléchi automatiquement grâce au logiciel UNITEX. Si l'on ajoute ou modifie un lemme dans le dictionnaire de lemmes, il suffit de relancer la flexion pour recréer automatiquement un dictionnaire de formes fléchies actualisé.

4.9.3 Conclusions

La création du DELFRABEL était une nécessité pour travailler de manière satisfaisante sur des corpus du français en Belgique. En effet, bien

²⁵ Lorsque le lexique est important, on peut également construire des graphes de flexion pour les noms et les adjectifs.

qu'assurant une couverture très complète des textes, les dictionnaires du LADL, conçus pour le français de France, s'avéraient inadéquats pour analyser des corpus belges francophones. Cela était d'autant plus prégnant qu'une partie de nos textes abordent des questions linguistiques liées à la norme, à l'accent et aux particularités lexicales.

L'application du DELFRABEL, qui comprend près de 5500 entrées (la moitié générées automatiquement grâce à des graphes de flexion), permet maintenant d'augmenter la couverture lexicale des textes analysés.

À terme, les entrées du DELFRABEL contiendront les informations issues de l'enquête sociolinguistique menée au centre VALIBEL. On pourra ainsi formuler des hypothèses, devant une forme reconnue grâce à notre dictionnaire, en ce qui concerne l'origine géographique du locuteur²⁶.

²⁶ Ou du scripteur. Ainsi, cette interview dans laquelle le journaliste fait dire au réalisateur danois Lars von Trier *j'étais bleu de* (« j'étais fou de ») :

Pourquoi avoir voulu tourner un « musical » ? Enfant, **j'étais bleu de** « La mélodie du bonheur » avec Julie Andrews, explique le réalisateur [Lars von Trier]. J'ai voulu insérer un personnage en souffrance dans cet univers-là qui est généralement rose bonbon. Un peu comme Gene Kelly, que je préfère à Fred Astaire, et Stanley Donen l'avaient fait avec « Beau fixe sur New York » qui est une oeuvre sur le désenchantement de l'amitié (Luc Honorez, *Le Soir*, 22 mai 2000, p. 9).

CHAPITRE 6

L'ÉTIQUETAGE MORPHOSYNTAXIQUE

- Miss Brewis affirme que lady Stubbs lui a demandé de porter un plateau de gâteaux, accompagnés de jus de fruits à Marlène Tucker. Michaël Weyman déclare que lady Stubbs était tout à fait incapable d'y avoir pensé...

- Ah ! dit Poirot, toujours les témoignages contradictoires ! On ne peut les éviter !

- Et ils sont bien ennuyeux à démêler, reconnut Bland. Ils sont parfois importants mais, dans neuf cas sur dix, ils n'ont aucune valeur. Nous ne manquons pas de travail d'élagage !

A. Christie, *Poirot joue le jeu*, Paris, Librairie des Champs-Élysées, 1957, p. 155.

1 QU'EST-CE QU'ÉTIQUETER UN TEXTE ?

Étiqueter un texte, ou encore l'annoter (*to mark up*, *to annotate*), consiste à adjoindre au texte « brut » des étiquettes contenant des informations¹. Habert *et al.* (1997) parlent d'*enrichissement* apporté au texte.

Leech (1997b) en donne la définition suivante :

¹ Pour un panorama de l'annotation automatique de textes, voir Garside *et al.* (Éds, 1997) et Véronis (2000). Notons que Véronis (2000 : 112) considère l'annotation comme

tout apport d'informations aux données brutes originales : la transcription d'un corpus oral sera donc considérée comme un processus d'annotation.

Si l'on a vu ce que la transcription supposait comme travail d'interprétation, nous n'avons pas utilisé le terme *annotation* pour cette étape. Chaque fois que nous avons parlé d'annotation dans ce travail, il s'agissait bien de qualifier l'ajout d'étiquettes dans le texte.

But what *is* corpus annotation ? It can be defined as the practice of adding **interpretative, linguistic** information to an electronic corpus of spoken and/or written language data. 'Annotation' can also refer to the end-product of this process : the linguistic symbols which are attached to, linked with, or interspersed with the electronic **representation** of the language material itself.

Ces informations constituent donc des ajouts aux données de base, ajouts qui peuvent être de natures très diverses. En effet, en ce qui concerne des données linguistiques, les informations dans les étiquettes peuvent relever de la phonétique, de la prosodie, de la syntaxe, de la morphosyntaxe ou encore de la sémantique (Cloeren 1999, Garside *et al.* 1997). Mais elles peuvent aussi concerner des phénomènes non linguistiques. Dans le cas de corpus oraux, on peut ajouter au texte brut des étiquettes qui indiquent le temps de parole des locuteurs, etc.²

Non seulement les annotations concernent donc des phénomènes divers, mais elles sont structurées différemment : cela va des formats d'annotations très personnels aux balises standardisées préconisées pour l'échange des données par des organismes comme EAGLES³ ou la TEI (Burnard 1995, Johansson 1995, Llisterri 1996, Sinclair 1996, TEI 2002).

Les dispositions graphiques des annotations sont elles aussi variées (Cloeren 1999, Habert 2006). Certaines annotations sont ajoutées directement dans le texte de départ, d'autres figurent « à côté » de celui-ci. Le système de gestion et d'annotation de données orales [moca] (Kevers *et al.* 2006) permet aux utilisateurs de faire plusieurs niveaux d'annotations dans des couches superposées. Celles-ci sont adjointes au texte de base, sans le modifier, et sont présentées en parallèle.

Enfin, l'annotation peut se faire de manière totalement automatique, de façon semi-automatisée ou encore manuellement.

² La TEI (*Text Encoding Initiative*), qui propose des normes d'annotation des données et des métadonnées afin de faciliter leur échange et leur réutilisation, propose de séparer les métadonnées des données et de les regrouper dans un *header* (en-tête) (TEI 2002). Voir <http://www.tei-c.org/>.

³ EAGLES (Expert Advisory Group on Language Engineering Standards, 1996) est un groupe de travail dont le but est de promouvoir la création de standards pour le développement, l'exploitation et l'évaluation de vastes ressources linguistiques. Voir <http://www.ilc.cnr.it/EAGLES/home.html>.

2 NOTRE ÉTIQUETAGE : L'ÉTIQUETAGE MORPHO-SYNTAXIQUE⁴

Dans la phase de prétraitement que nous avons décrite au chapitre 4, nous avons déjà ajouté une série d'étiquettes dans le texte, qui visent à baliser des portions de celui-ci et à le découper en unités plus petites que le tour de parole. Nous avons aussi balisé automatiquement certaines disfluences, mais sans procéder à une normalisation, qui consisterait par exemple à rétablir la forme pleine des amorces, comme chez Bénard (2005) ou chez Oostdijk (2003), à la suite de Nelson (1996b).

Il s'agissait pour nous d'un prétraitement des données, même s'il est vrai que nous participions de l'enrichissement du texte de base en y ajoutant des étiquettes.

Dans le cadre de ce travail, nous entendons plus précisément par *étiquetage* (*tagging*) ou *annotation* un *étiquetage morphosyntaxique* (*part-of-speech tagging*, *word-class tagging*)⁵ du texte : il s'agit d'attribuer à chaque mot (ou groupe de mots, dans le cas des mots composés) du texte une étiquette qui contient trois types d'informations :

- le lemme (ou forme canonique),
- la catégorie grammaticale,
- des informations quant à la flexion.

Cette triple contrainte n'est pas une évidence dans le cadre de l'annotation morphosyntaxique. En effet, tous les systèmes ne donnent pas nécessairement le lemme, et les indications de flexion sont ou non présentes. On a donc pour un même niveau d'annotation des informations plus ou moins développées dans les étiquettes (cf. § 4).

L'étiquetage morphosyntaxique n'est pas une fin en soi, nous l'avons déjà dit. En général, il s'agit d'une étape préalable à d'autres traitements sur le texte. Comme le fait remarquer Laporte (2000 : 28) :

(...) pratiquement toutes les applications qui mettent en jeu du texte verraient leurs performances s'améliorer par l'intégration de meilleurs systèmes d'étiquetage lexical, y compris des applications relevant de la syntaxe ou de la parole. C'est d'ailleurs un des acquis

⁴ Le lecteur se reportera utilement à Van Halteren (1999) pour une vue très complète de la question.

⁵ On trouve aussi les termes *morphosyntactic annotation* ou *wordclass tagging*.

Le travail qui consiste à regrouper des mots en constituants et à leur attribuer une fonction syntaxique est le *parsage* (*parsing*), et il ne nous intéresse pas ici.

sur lesquels s'accorde la communauté scientifique du domaine : après une période de décantation des techniques destinées à effectuer un traitement donné, on arrive à la conclusion qu'une étape préalable d'étiquetage lexical de qualité résoudrait bien des difficultés.

Les participants à l'action d'évaluation GRACE (cf. § 7 de ce chapitre) visaient d'ailleurs une série d'applications diverses : la correction orthographique, l'accentuation de textes non accentués, la lemmatisation, l'extraction terminologique, le filtrage ou la classification de documents, la recherche documentaire, la segmentation de textes, l'analyse syntaxique, la traduction automatique ou encore l'annotation sémantique.

Si l'annotation peut être manuelle, en règle générale, l'étiquetage est obtenu grâce à des outils appelés *étiqueteurs morphosyntaxiques* ou encore *tagueurs* (*part-of-speech tagger*).

3 LES CORPUS ANNOTÉS MORPHOSYNTAXIQUEMENT

Nous avons déjà abordé au chapitre 2 la notion de corpus. Nous allons maintenant passer rapidement en revue les corpus annotés les plus célèbres.

La liste des corpus cités ici ne se veut pas une liste exhaustive. En effet, depuis plus d'une quarantaine d'années, la pratique de la constitution et de l'annotation de corpus s'est considérablement développée et répandue, principalement dans le monde anglo-saxon. De nombreux corpus ont été constitués, annotés ensuite et parfois mis à la disposition des chercheurs.

Nous ne reprenons ici que les corpus les plus connus, soit parce qu'ils font figure de corpus « historiques », soit parce qu'ils sont fréquemment cités car utilisés dans de nombreuses recherches en linguistique. Il est en effet intéressant de constater que c'est bien souvent la diffusion libre (dans le sens de « gratuite ») du corpus auprès de la communauté des chercheurs qui est à la base de sa notoriété⁶. Ainsi, si certains corpus sont devenus en quelque sorte incontournables, cela n'est pas tant dû à la qualité des données qu'à leur large diffusion parmi la communauté des linguistes et autres chercheurs.

⁶ La facilité avec laquelle certains corpus sont accessibles explique leur popularité auprès des chercheurs. Ainsi, en linguistique de corpus, on ne calcule plus les études qui prennent pour données *Le Monde* ou, pour le français en Belgique, le journal *Le Soir*.

3.1 *Brown Corpus*

Parmi les corpus les plus célèbres, le *Brown Corpus* (*A Standard Corpus of Present-Day American English*) s'avère une référence incontournable pour l'anglais américain écrit.

Constitué en 1964 à l'Université de Brown par Kucera et Francis (1967), il se compose de textes rédigés par des natifs anglophones américains et publiés durant l'année 1961. Le corpus, d'un peu plus d'un million de mots, est divisé en 500 extraits de 2000 mots chacun. On peut réellement parler d'extraits dans le cas de ce corpus puisque le texte composant une séquence s'arrête arbitrairement à la fin de la phrase qui comprend le deux millièm mot, sans tenir compte, par exemple, de la disposition du texte en paragraphes⁷.

Les auteurs ont résolument privilégié la prose écrite, écartant volontairement la poésie⁸ (qui pose des problèmes linguistiques spécifiques) ainsi que les drames « as being the imaginative recreation of spoken discourse, rather than true written discourse »⁹ (Francis et Kucera et 1979).

Leur but était d'offrir un échantillon de toutes les facettes de l'américain écrit contemporain. Ainsi, si les extraits ont été choisis de manière aléatoire, ceux-ci entrent par contre dans les 15 catégories établies par Kucera et Francis (1967), catégories qui correspondent *grosso modo* à des « genres » différents, censés représenter l'usage contemporain de l'écrit (presse, écrits scientifiques, écrits religieux, rapports du gouvernement ou de l'industrie, romans ou nouvelles différenciés selon les genres de science-fiction, d'aventure, d'amour, etc.). On voit apparaître ici l'idée de *représentativité* que nous avons abordée au chapitre 2.

Ayant été révisé et remis à jour en 1971, le *Brown corpus* doit également sa renommée actuelle au fait qu'il fait figure de pionnier parmi les corpus annotés. En effet, un étiqueteur, TAGGIT, a été développé pour annoter une partie du corpus, qui a ensuite été corrigé manuellement (Greene et Rubin 1971).

Il existe six versions du corpus qui diffèrent par le codage et le format (Francis et Kucera 1979). En 1979 est diffusée une version étiquetée semi-automatiquement du corpus, la version C, qui utilise 71 étiquettes différentes.

⁷ Notons que l'extrait ne commence pas nécessairement non plus au début d'un paragraphe ou d'une division plus grande.

⁸ Sauf lors de citations très brèves dans la prose.

⁹ Des textes de fiction ont été repris, à la condition que le dialogue ne dépasse pas 50 % du texte.

3.2 *The Lancaster-Oslo-Bergen Corpus*

Le *Lancaster-Oslo-Bergen (LOB) Corpus* est l'homologue en anglais britannique du *Brown Corpus*.

Dirigé de 1970 à 1976 par Geoffrey Leech à l'Université de Lancaster, le projet a ensuite été déplacé à Oslo où Stig Johansson, du Département d'anglais de l'Université d'Oslo, l'a pris en charge. Il fut achevé en 1978 (Johansson *et al.* 1978).

C'est de manière explicite et volontaire que ce corpus a été constitué pour être l'équivalent du *Brown Corpus*, afin de permettre des recherches comparatives :

- il comprend un million de mots, répartis en 500 échantillons de 2000 mots ;
- les extraits ont été sélectionnés selon les critères de « genre » équivalents pour l'anglais britannique ;
- les textes qui le composent ont été publiés durant l'année 1961.

Le million de mots du *LOB* a également été étiqueté en 1986 (Beale 1985).

3.3 *British National Corpus*¹⁰

Le *British National Corpus* (ou *BNC*), avec ses 100 millions de mots, est à notre connaissance le plus grand corpus étiqueté.

La constitution du corpus, entamée en 1991, a été achevée en 1994 sous la houlette d'un consortium composé de centres de recherche (des universités d'Oxford et de Lancaster associés à la British Library) et d'éditeurs de dictionnaires (OUP, Longman et Chambers-Larousse).

Le *BNC* veut enregistrer les usages contemporains de l'anglais britannique. Il s'agit donc d'une même perspective synchronique que dans le *Brown* ou le *LOB*, les textes repris dans le *BNC* n'étant pas antérieurs à 1975 (1964 pour des textes littéraires célèbres). Visant lui aussi à la représentativité, le *BNC* se compose d'échantillons variés (en général de 45 000 mots pour la partie écrite), allant d'extraits de journaux à des lettres personnelles, en passant par des mémoires d'étudiants, etc.

Trois modes différents de constitution du corpus ont prévalu : le scannage, la reprise d'un texte déjà sur support électronique (c'est le cas des journaux notamment) et l'encodage direct via le clavier d'un ordinateur.

¹⁰ <http://www.natcorp.ox.ac.uk/>.

À la différence du *Brown* et du *LOB*, qui comportent strictement des textes écrits, l'oral a été pris en compte dans le *BNC* (Crowdy 1995). En effet, 10 % des données sont des transcriptions du parlé de locuteurs anglais natifs (863 des 4 124 textes du corpus). Les situations de communication sont diversifiées, allant de l'informel à des transcriptions de débats gouvernementaux.



La partie orale du *BNC*

La partie anglais parlé du *BNC* est divisée en deux grands types de productions :

- la partie démographique du corpus « The Demographic part of the Spoken Corpus », composée de 4 211 216 mots : des enregistrements de conversations informelles, faits par 124 volontaires issus de 38 points d'enquêtes du Royaume-Uni ;
- la partie gouvernée par le contexte (« The Context-governed part of the Spoken Corpus »), qui compte 6 154 248 mots : pour les concepteurs du corpus, il s'agissait de rassembler des productions discursives, en quantité égale, qui relèvent des quatre contextes sociaux suivants : événements éducatifs (p. ex. conférences), d'affaires (p. ex. réunions de travail, ventes), institutionnels ou publics (p. ex. discours politiques) et événements liés aux loisirs (p. ex. commentaires sportifs).

Le corpus est scindé en unités phrases (« 6 and a quarter million sentences ») grâce au logiciel CLAWS4 développé par Roger Garside à l'Université de Lancaster¹¹. L'annotation du corpus, totalement automatisée, a été réalisée avec l'étiqueteur stochastique GRIFFES (Leech *et al* 1994). Elle consiste en une étiquette de catégorie grammaticale uniquement, sans lemmatisation.

Le jeu d'étiquettes de base (C5), pour les parties traditionnelles du discours, comprend 61 étiquettes (Garside 1995). La taille du corpus excluait d'emblée une correction manuelle de celui-ci. Le taux d'erreur approximatif est de 1,7 % (Leech 1997a). 4,7 % des étiquettes sont des étiquettes qui neutralisent certaines ambiguïtés (cf. § 5).

¹¹ Notons que les transcriptions de la partie orale du *BNC* sont ponctuées afin d'identifier des « sentence-like units » (Crowdy 1995 : 231).

3.4 *Penn Treebank*

Le *Penn Treebank* est un corpus de 4 millions de mots d'anglais américain, qui reprend une partie du *Brown Corpus* (1,2 millions de mots). Il est étiqueté morphosyntaxiquement (Marcus *et al.* 1993, Santorini 1990), mais également arboré. L'annotation automatique a été corrigée manuellement.

Le nombre d'étiquettes a été réduit à 48 (36 pour les parties du discours et 12 pour la ponctuation et les symboles).

Le *Penn Treebank* contient également une partie d'oral retranscrit.

3.5 *Le Corpus d'Orléans*

Le *Corpus d'Orléans* (aussi appelé *ESLO*, *Étude sociolinguistique sur Orléans*) est sans doute la première entreprise de constitution à grande échelle d'un corpus de français parlé. Réalisés entre 1968 et 1971, les enregistrements ont été faits à Orléans dans le cadre d'enquêtes sociolinguistiques, par des sociolinguistes et des professeurs de français britanniques¹².

Le corpus comprend 315 heures d'enregistrements, ce qui correspond approximativement à 4 500 000 mots. Se sont ajoutés ensuite à cette entreprise deux autres corpus de français parlé : *Le livre parlé de Tours* (1974) et *Voix d'Auvergne* (1976).

Initialement constitués par des Britanniques, une partie de ces corpus a été transcrite graphiquement dans les universités de Leuven et d'Amsterdam.

La collection de corpus ELICOP (Étude linguistique de la communication parlée, voir Mertens 2002) a rassemblé les transcriptions de ces corpus à l'Université de Leuven (approximativement 1 250 000 mots) où celles-ci ont été normalisées.

Une partie du corpus, composée de 243 000 mots, a été étiquetée automatiquement grâce au lemmatiseur MORLEX¹³ et à l'analyseur VERTEX¹⁴. L'étiquetage a ensuite été validé (Mertens 2002). Le corpus (étiqueté ou sous sa forme brute) est consultable en ligne¹⁵.

¹² Les chercheurs français qui déplorent le retard dans la constitution et l'exploitation de corpus de français parlé voient dans le *Corpus d'Orléans* tout un symbole : constitué par une équipe britannique, il est exploité hors de France.

¹³ <http://bach.arts.kuleuven.ac.be/pmertens/morlex>.

¹⁴ <http://bach.arts.kuleuven.be/pmertens/vertex/>.

¹⁵ <http://bach.arts.kuleuven.be/elicop/>.

3.6 Le Corpus Paris 7

Le plus grand corpus de français (écrit) étiqueté morphosyntaxiquement est le *Corpus de Paris 7*. Le corpus est constitué d'articles écrits entre 1989 et 1993 et issus de différentes rubriques du journal *Le Monde*. Il comprend un million de mots graphiques, ponctuation comprise¹⁶. Il est annoté morphosyntaxiquement. Le corpus a été étiqueté automatiquement par un étiqueteur mixte en deux phases : une première passe avec des règles stochastiques, puis une seconde passe avec des règles contextuelles. Les annotations ont ensuite été validées par un expert humain (Clément 2001).

On le voit, la communauté francophone manque cruellement de données : rares sont les corpus oraux de grande taille (cf. chap. 2), mais encore plus rares les corpus annotés morphosyntaxiquement. Le corpus ELICOP fait figure d'exception, mais les quelques recherches que nous avons faites montrent qu'il contient encore beaucoup d'erreurs, malgré la validation humaine¹⁷. Les bonnes intentions ne manquent pourtant pas : en 1999, Mireille Bilger¹⁸ annonçait l'étiquetage prochain du corpus du GARS ; Antoine *et al.* (2002), qui présentent comme objectif d'annoter des corpus oraux de dialogues, précisent : « L'annotation morpho-syntaxique de nos corpus n'a pas encore été réalisée ». À notre connaissance, aucune de ces deux entreprises n'a encore abouti.

Étiqueter morphosyntaxiquement une partie de la banque de données VALIBEL comblera donc une lacune unanimement reconnue dans le monde francophone.

4 GRANULARITÉ DU JEU D'ÉTIQUETTES ET CONCEPT D'AMBIGÜITÉ

Si l'étiquetage morphosyntaxique n'est pas une activité simple, c'est que beaucoup de mots sont ambigus¹⁹. On parle ici d'ambigüités *virtuelles*, et non d'ambigüités *réelles* en discours (Fuchs 1996). Comme le précise Mertens (2002 : 10) :

Étant donné la facilité apparente avec laquelle l'être humain effectue ce genre d'étiquetage, on a tendance à

¹⁶ On considère en général que la ponctuation représente 15 % des tokens (Leech et Wilson 1999 : 62).

¹⁷ La recherche sur la forme *est* comme adjectif ne fournit que des mauvaises occurrences, par exemple. La validation a été faite par des étudiants.

¹⁸ Lors du colloque « Le français de référence. Constructions et appropriations d'un concept », qui s'est tenu à Louvain-la-Neuve, du 3 au 5 novembre 1999.

¹⁹ Voir Fuchs (1996) pour une typologie des différents niveaux d'ambigüités.

sous-estimer sa complexité réelle d'un point de vue informatique.

Nous avons vu au chapitre 5 (§ 3.1) que, pour les systèmes qui utilisent des dictionnaires, le concept d'ambiguïté dépend de la manière dont ceux-ci sont conçus. Ainsi, le codage que nous avons choisi pour les mots grammaticaux réduit d'emblée certaines ambiguïtés ; de même, ne pas tenir compte des emplois des verbes selon la distinction établie dans le lexique-grammaire permet de ne pas démultiplier les entrées pour une même forme (Gross 1975).

Prenons l'exemple suivant, extrait d'un article d'Alain Guillet (1995), dans lequel l'auteur discute de la possibilité de lever automatiquement les ambiguïtés lexicales²⁰ :

La porte du car se ferme automatiquement

Selon Guillet, on a pour cette phrase 72 analyses possibles, en fonction du degré d'ambiguïtés attribué à chaque forme : $3 \times 3 \times 1 \times 2 \times 1 \times 4 \times 1$.

Lorsque la description des traits dans le dictionnaire est plus fine²¹, on n'a plus 72 analyses mais 9216 analyses : $4 \times 8 \times 2 \times 2 \times 8 \times 9 \times 1 = 9216$ analyses.

Voici les ambiguïtés de chaque forme : **la** (4 analyses) : DET:fs, PRO:3fs, N:ms, N:mp ;

- **porte** (8 analyses) : V:P1s, V:P3s, V:S1s, V:S3s, V:Y2s, A:fs, A:ms, N:fs ;
- **du** (2 analyses) : **du**.DET+Dind, **de**.PREP **le**.DET+Def ;
- **car** (2 analyses) : N:ms, CONJC ;
- **se** (8 analyses) : PRO+PpvLUI:3ms, PRO+PpvLUI:3fs, PRO+PpvLUI:3mp, PRO+PpvLUI:3fp, PRO+PpvLE:3ms, PRO+PpvLE:3fs, PRO+PpvLE:3mp, PRO+PpvLE:3fp ;
- **ferme** (9 analyses) : V:P1s, V:P3s, V:S1s, V:S3s, V:Y2s, ADV, N:fs, A:fs, A:ms ;

²⁰ Cette phrase a servi de base à une communication que nous avons présentée avec Olivier Blanc (Université de Marne-la-Vallée) au 22^e colloque international « Grammaires et lexiques comparés » qui s'est tenu à Thessalonique (Grèce) du 17 au 22 septembre 2003. Le titre de la communication était : « Levée d'ambiguïtés lexicales par contraintes locales ».

²¹ Ce calcul est effectué avec le jeu d'étiquettes utilisé pour le français écrit standard décrit au chapitre 5.

• **automatiquement** (1 analyse) : ADV. Si maintenant on tient compte de la distinction établie pour les emplois des verbes dans le lexique-grammaire, le nombre d'ambiguïtés est encore multiplié pour *porte* et *ferme*, et l'on obtient : $4 \times 71 \times 2 \times 2 \times 8 \times 59 \times 1 = 536\,192$ analyses.

On voit ici que le calcul de l'ambiguïté est fortement variable (on passe de 72 à 536 192 analyses pour la même phrase), et que ces résultats illustrent des ambiguïtés virtuelles ou artefactuelles. En effet, une seule analyse est correcte pour cette phrase, parmi toutes celles supposées possibles selon les données recensées en langue.

Si l'ambiguïté dépend de la richesse de la description lexicale, elle est aussi intimement liée au degré de précision attendu dans le résultat de l'étiquetage. Par *précision*, nous entendons ici le nombre et le type d'informations souhaitées dans les étiquettes. Ainsi, prenons le cas de différents types d'étiquettes, selon la richesse et la finesse de l'information qu'elles contiennent (on parle de *granularité*) :

1. une **entrée lexicale complète**, c'est-à-dire le choix que nous faisons dans ce travail : lemme+catégorie+flexion, telles que ces informations sont présentes dans le dictionnaire DELAF ;
2. le **lemme** et la **catégorie grammaticale**. Il n'y a donc pas d'informations de mode, temps, personne et nombre pour les verbes ; pas d'informations de nombre, de personne et de genre pour les pronoms personnels ; pas d'informations de genre et de nombre pour les autres pronoms, les noms, les adjectifs et les déterminants ;
3. la **forme** et la **catégorie**, sans les informations flexionnelles ou de conjugaison (cf. le mode précédent) ;
4. le **lemme** uniquement.

Les quelques exemples qui suivent montreront aisément que la tâche de désambiguïsation part de données différentes (même si l'on se base sur le même dictionnaire comme c'est le cas ici) en fonction du type d'étiquettes choisi.

La forme *fin* a quatre entrées distinctes dans le DELAF :

fin,fin.A:ms
 fin,fin.ADV
 fin,fin.N:fs
 fin,fin.N:ms

Si l'on se place dans le mode d'étiquetage où les étiquettes correspondent à l'entrée lexicale — les étiquettes sont les plus complètes —, *fin* est quatre fois ambigu (puisque nous avons quatre entrées). À l'inverse, si l'on désire obtenir uniquement le lemme, la forme n'est plus ambiguë. Par contre, si au lemme s'adjoint la catégorie grammaticale (mode 2), on a alors trois ambiguïtés :

fin.A
 fin.ADV
 fin.N

On obtient le même résultat si l'on désire la forme et la catégorie, en mode (c).

On pourrait imaginer que, pour ces deux derniers modes, les informations flexionnelles accompagnent la catégorie grammaticale. Puisque les deux hypothèses pour le nom se distinguent par le genre, on aurait alors de nouveau, bien que l'on choisisse uniquement le lemme ou la forme, quatre ambiguïtés.

Prenons encore l'exemple de *nouvelles*, qui a un degré d'ambiguïté 4 dans le DELAF :

nouvelles,nouveau.A:fp
 nouvelles,nouveau.N:fp
 nouvelles,nouvelle.N:fp
 nouvelles,nouvelles.N:fp

Quatre entrées, donc quatre ambiguïtés pour le mode 1. Si l'on passe dans le mode 2 (le lemme et la catégorie), on obtient trois hypothèses pour la forme :

nouveau.N
 nouvelle.N
 nouvelles.N

Dans le mode 3 (forme et catégorie), on n'a plus alors que deux ambiguïtés :

nouvelles.A
 nouvelles.N

Les deux exemples qui précèdent montrent que les modes 2 et 3 ne donnent pas nécessairement le même nombre d'étiquettes.

On pourrait encore multiplier les exemples, et l'on voit très clairement que la nature même des données lexicales de base (la finesse des informations présentes dans le dictionnaire et les choix de codage), ainsi que la précision des résultats attendus, en termes de complétude des étiquettes, interviennent dans la définition du concept d'ambiguïté.

Dans le cadre de notre projet, l'étiquetage consiste à indiquer non seulement le lemme (forme canonique), mais également la catégorie grammaticale et les informations flexionnelles (s'il y a lieu). Or, dans certaines annotations, les étiquettes ne sont pas nécessairement aussi complètes.

Ainsi, si l'étiqueteur Cordial²² étiquette le déterminant *le* DETDMS, le TreeTagger²³ propose DET:ART, sans indication du genre et du nombre ; de la même manière, le TreeTagger ne donne pas la personne de la conjugaison des verbes²⁴. On peut parler, dans le cas des étiqueteurs, d'un « éparpillement des étiquetages » (Watrin 2002 : 25).

Les jeux d'étiquettes sont très variables d'un système à l'autre. Alors que Labbé (1990) utilise seulement 16 étiquettes pour étudier les discours de François Mitterrand (Labbé 1990), le catégoriseur Winbrill pour le français en contient 100 (Lecomte 1998) et MORLEX en recense 207 (Mertens 2002). Selon Véronis et Khouri (1995), la plupart des étiqueteurs probabilistes du français utilisent un jeu d'étiquettes de l'ordre de 70 à 150 étiquettes. Le jeu d'étiquettes de l'action GRACE (postérieure à l'article de Véronis et Khouri 1995) en contenait 311. Pour les catégoriseurs de l'anglais, Leech (1997c : 29) donne des exemples d'étiqueteurs dont le jeu d'étiquettes varie de 32 à 270 étiquettes.

Notre jeu d'étiquettes est quant à lui composé de 230 étiquettes. Mais si l'on considère que le lemme fait partie de l'étiquette (ce qui est notre cas effectivement), on a alors autant d'étiquettes que d'entrées dans le dictionnaire²⁵.

²² http://www.synapse-fr.com/correcteur_orthographe_grammaire/correcteur_orthographe_grammaire.htm#cordialStandard.

²³ <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html>.

²⁴ Voir Habert (2005b) ou Watrin (2002) pour la comparaison des étiqueteurs. Salmon-Alt et Durand (2003) montrent également les représentations différentes fournies par deux étiqueteurs.

²⁵ Pour une variation sur les jeux d'étiquettes, voir Laporte (1996).

Comme le fait remarquer Laporte (1996 : 298) à propos des jeux d'étiquettes :

(...) bien que le domaine [de l'étiquetage lexical] donne lieu à une intense compétition internationale, l'abondante littérature sur la levée d'ambiguïtés est généralement laconique sur les jeux d'étiquettes : Brill (1995), par exemple, en présentant un des systèmes d'étiquetage les plus réputés, se place clairement sur un plan compétitif mais ne donne aucun élément en faveur de son choix d'un jeu de... 36 étiquettes lexicales faisant abstraction des mots composés. L'habitude, et c'est regrettable, est de ne faire porter la compétition que sur d'autres critères, purement quantitatifs (...).

Cela dit, on peut augmenter le nombre d'étiquettes sans rien ajouter à la précision de la description. Ainsi, la création d'étiquettes *ad hoc* multiplie le nombre d'étiquettes sans fournir d'information supplémentaire dans l'annotation. C'est le cas des étiquettes qui ne réfèrent qu'à une forme en particulier (étiquette uniquement pour l'adverbe de négation *ne*), ou des étiquettes *porte-manteaux*²⁶ qui neutralisent des ambiguïtés difficiles à lever (pour le français, la distinction entre le présent de l'indicatif et du subjonctif des verbes du premier groupe, par exemple, cf. Cloeren 1999 : 49, van Halteren 1999 : 87). Dans ces cas, on a bien des étiquettes supplémentaires, mais guère de valeur ajoutée aux informations fournies avec l'annotation.

Mais même avec un consensus sur le jeu d'étiquettes et sa granularité, on ne peut garantir une uniformisation de l'étiquetage fait par des systèmes différents. Comme le précisent Véronis et Khouri (1995) :

Même si l'on est d'accord avec le jeu d'étiquettes, leurs extensions (c'est-à-dire l'ensemble des formes lexicales qu'elles couvrent) peuvent être différentes. Le problème est particulièrement aigu pour les catégories fermées, déterminants, pronoms, adjectifs indéfinis, etc. où l'on rencontre de très grosses différences d'appréciation quant au placement des mots dans les catégories, et ce dans la plupart des langues.

C'est d'ailleurs la raison pour laquelle nous avons particulièrement détaillé ce point dans la section consacrée à nos choix linguistiques.

²⁶ Les étiquettes *porte-manteaux*, *portmanteau tags*, sont appelées ainsi en référence aux *portmanteau words*, c'est-à-dire aux mots-valises, créés grâce à la combinaison de deux mots (cf. les créations du lapin Humpty Dumpty de Lewis Carroll).

5 AMBIGÜITÉS FRÉQUENTES

Le taux d'ambigüités de 12 % que nous donnions au chapitre 5 (§ 3.1) a été calculé par Blandine Courtois sur les entrées du dictionnaire DELAF. Il s'agit en quelque sorte de chiffres « théoriques », calculés sur l'ensemble du vocabulaire recensé dans le dictionnaire. Une autre démarche pour comptabiliser les ambigüités consiste à se baser sur des données pragmatiques qui relèvent de comptages des occurrences de formes dans des textes.

En effet, « à partir d'un texte donné, l'homographie au niveau des formes graphiques (en tant que réalisations de lexèmes) est beaucoup plus fréquente que celle visible au niveau des vocables dans un dictionnaire » (Bolasco 1993) et cela, sans doute dans une large mesure parce que « [les ambigüités] concernent bien plus les mots fréquents que les mots rares » (Silberztein 1995).

Pour Sergio Bolasco, les homographies les plus intéressantes à lever se répartissent comme suit, en termes de catégorie grammaticale :

nom / verbe	60 %
nom / adjectif	16 %
adjectif / verbe	11 %
le reste	13 %

Tableau 41. Répartition des homographies chez Bolasco (1993)

L'auteur centre donc son objectif de levée d'ambigüité sur les mots lexicaux.

Suivant une démarche comparable qui consiste à analyser les données d'un corpus, nous avons extrait les 100 chaînes de caractères les plus fréquentes dans une année du journal *Le Monde* (LM94)²⁷. Après avoir enlevé les caractères diacritiques de cette liste, nous avons regroupé les différentes graphies sous une même forme (*et* pour *et*, *ET*, *Et*).

En termes de formes, on constate que parmi les 70 occurrences les plus fréquentes, très peu participent des types d'ambigüités de catégories grammaticales mises en avant par Bolasco. Pour l'ambigüité **nom / verbe**, nous avons *a*, *être*, *été*, *plus*, *président* et *A*. Pour l'ambigüité **nom / adjectif**, on trouve uniquement la forme *bien*, qui apparaît également dans la locution

²⁷ Sur les homographies les plus fréquentes dans un corpus, avec des données chiffrées, voir Dister (2000).

conjonctive *bien que*. On trouve seulement deux formes qui possèdent la triple ambiguïté **nom / verbe / adjectif** : *est* et *fait*.

En fait, nous avons vu au chapitre 2 que les formes les plus fréquentes sont principalement des mots grammaticaux. En tête de *LM94*, nous avons *de* (ambiguïté entre déterminant et préposition), *la* et *l* (nom, déterminant ou pronom), *le* et *les* (déterminant ou pronom). Dans notre corpus, on a 92 % de mots grammaticaux et d'adverbes parmi les 25 formes les plus fréquentes.

On peut se demander dans quelle mesure il est intéressant de s'appliquer à lever l'ambiguïté de ces mots grammaticaux, considérés comme vides de sens. Cette tâche ne nous semble pas inutile car lever l'ambiguïté sur un mot grammatical permet bien souvent de lever l'ambiguïté sur la ou les forme(s) pleine(s) qui l'entourent. En effet, celles-ci marquent fréquemment le début d'un constituant minimal (Abney 1991, Vergne 2001).

Lorsque l'on parle des ambiguïtés les plus fréquentes, il est possible d'aborder le problème sous deux angles différents : celui de la catégorie grammaticale en général – comme c'est le cas de Bolasco –, ou celui des formes graphiques en particulier.

Dans les grammaires de désambiguïsation que nous avons construites, nous avons abordé simultanément ces deux approches, l'une qui consiste à se baser sur le lexique (dans des grammaires dites *lexicalisées*), et l'autre, traitant plutôt des catégories grammaticales, en tenant compte ou non des informations flexionnelles.

6 LES MÉTHODES D'ÉTIQUETAGE

Non seulement les jeux d'étiquettes et la finesse des informations varient d'un étiqueteur à l'autre, mais les méthodes utilisées également.

Notre processus d'étiquetage se déroule en deux temps. Dans un premier temps, on procède à un *étiquetage a priori* (Paroubek et Rajman 2000) : à chaque mot ou groupe de mots du texte, on associe une ou plusieurs étiquettes après consultation du dictionnaire. Dans une seconde étape, on choisit parmi ces étiquettes celle qui est pertinente en contexte (cf. § 11). C'est cette deuxième phase, la phase de *désambiguïsation*, qui produira l'étiquetage final.

Notre système d'étiquetage est un système que l'on peut qualifier de totalement linguistique : la recherche de toutes les étiquettes possibles se base sur des ressources lexicales consignées dans des dictionnaires électroniques par des linguistes ; la désambiguïsation se fait grâce à des descriptions linguistiques fines représentées formellement dans des grammaires locales construites par des experts humains (*hand-crafted rules*, Voutilainen 1999).

Dans notre méthode d'étiquetage, nous partons de descriptions lexicales fournies *a priori*. L'étiquetage est vu comme un problème de *classification supervisée* (Paroubek et Rajman 2000). Mais tous les systèmes d'étiquetage ne fonctionnent pas de la sorte.

Ainsi, de nombreux systèmes sont basés sur des apprentissages automatiques. L'étiqueteur sans doute le plus célèbre basé sur des règles avec apprentissage est celui mis au point par Eric Brill (2003) pour l'anglais. Il a connu de nombreuses adaptations pour d'autres langues, notamment le français (Lecomte 1998, Thibeault 2004). Il s'agit d'un étiqueteur *supervisé*. L'idée est de partir d'un corpus étiqueté manuellement par un expert humain. Ensuite, tout le processus d'étiquetage se fait par apprentissage automatique sur cet échantillon étiqueté qui fonctionne comme un corpus d'entraînement (voir Brill 1999b pour les règles apprises sur corpus, *corpus-based rules*).

Afin d'éviter cette couteuse tâche au départ d'un corpus étiqueté manuellement, des systèmes ont vu le jour qui se passent de cette étape dans laquelle intervient le linguiste (Atwell 1983). Ces approches statistiques ont connu un véritable essor (Merialdo 1994, Roche et Schabès 1995), ainsi que les systèmes mixtes, qui combinent les approches probabilistes et linguistiques (El-Beze et Spriet 1995, Clément 2001).

Parce que notre travail se situe dans une perspective résolument linguistique, qui tient compte de représentations formalisées sur la base de descriptions linguistiques fines, nous ne développerons pas davantage ces méthodes. Il ne s'agit pas pour nous de juger de la pertinence de l'une ou l'autre méthode, mais bien de séparer des approches qui nous paraissent radicalement différentes²⁸.

Comme d'autres projets avant le nôtre – corpus *Penn TreeBank* (Marcus *et al.* 1993) ou le *Corpus annoté de Paris 7* (Clément 2001) –, le processus pour aboutir au texte étiqueté se fait en deux grandes phases successives : la première est semi-automatisée et elle consiste en un étiquetage du texte²⁹ ; la seconde est une relecture des résultats fournis par l'étiquetage automatique, par des experts humains. Les résultats de l'étiquetage sont donc validés au final. Cette dernière phase a toute son importance car cela signifie que dans les objectifs à atteindre lors de la première étape, qui nous occupe dans ce travail, nous ne perdons pas de vue que les résultats vont être ensuite validés, voire corrigés. Cela signifie par exemple, que l'on n'est pas obligé d'avoir

28 Pour plus de détails sur les différentes méthodes d'étiquetage, on se reportera utilement à Brill (1999b), Daelemans (1999), El-Beze et Merialdo (1999), Paroubek et Rajman (2000), Voutilainen (1999).

29 Étape décomposée chez nous en un prétraitement du texte (chap. 4), la consultation de dictionnaires pour un étiquetage *a priori* (chap. 5) et l'application de grammaires de levée d'ambiguïtés (chap. 6).

tout étiqueté puisque des correcteurs humains finalisent le processus d'étiquetage. En d'autres termes, plusieurs étiquettes peuvent subsister en concurrence après la phase automatisée. Nous préférons cette solution à celle qui ne donnerait qu'une seule étiquette, au risque que celle-ci soit mauvaise (cf. § 8).

7 LA CAMPAGNE D'ÉVALUATION GRACE

Démarrée en 1994, sous la coordination de L'INALF et du LIMSI, l'action GRACE (Grammaires et Ressources pour les Analyseurs de Corpus et leur Évaluation) visait à évaluer des assignateurs de catégories morphosyntaxiques du français, en mettant en place un protocole qui permette la comparaison de systèmes créés selon des méthodes très différentes. En proposant un cadre de référence unique pour l'évaluation des analyseurs morphosyntaxiques, GRACE allait dans le sens des actions de constitution de larges ressources linguistiques ou encore de définitions de standards (TEI, EAGLES, MULTEXT).

La campagne s'est déroulée en trois phases successives, de janvier 1996 à 1998 (Adda *et al.* 1999) : une phase d'apprentissage, une phase d'essais et enfin une phase de tests qui s'est achevée par l'évaluation proprement dite des résultats.

Le corpus, composé d'extraits de la base textuelle FRANTEXT (textes littéraires) et du journal *Le Monde*, comprenait 650 000 formes. Vingt-et-une équipes ont participé au projet, parmi lesquelles treize l'ont mené jusqu'à son terme.

L'action GRACE visait à mesurer les performances effectives des systèmes et non leur fonctionnement interne (les systèmes fonctionnaient comme une boîte noire) sur un corpus de grande taille. L'évaluation a nécessité la mise en place d'un référentiel commun.

À notre sens, le principal apport de GRACE³⁰ a été de mettre clairement en évidence les difficultés de comparer des systèmes d'étiquetage différents et leurs résultats respectifs (Adda *et al.* 1998, 1999a et 1999b, Paroubek 1997, Paroubek et Rajman 2000)³¹.

Cette difficulté est également mise en exergue par Véronis et Khouri (1995) dans le cadre du projet multilingue MULTEXT. Les auteurs parlent de

³⁰ En effet, cette action n'a guère eu de suite. On aurait pu s'attendre, à l'issue d'une telle campagne, à des recommandations, des analyses d'erreurs fréquentes, des propositions de pistes de travail à développer, etc. Selon l'un des chercheurs impliqués dans l'action (communication personnelle), le manque de financement a stoppé la suite du projet.

³¹ Les mêmes difficultés sont rencontrées dans l'évaluation des analyseurs syntaxiques, dont les données en sortie sont aussi variées (Blache et Morin 2003, Carroll *et al.* 2002).

« la cuisante expérience de l'incomparabilité des jeux d'étiquettes ». Véronis et Khouri avancent :

de nombreux articles annoncent à l'heure actuelle des résultats de l'ordre de 95 ou 97 % de désambiguïisations correctes, mais la comparaison de ces chiffres n'a aucun sens, étant donné que l'on ne sait pas comparer les jeux d'étiquettes utilisés.

Et ils ajoutent en note :

Pour comprendre l'absurdité de la comparaison, il suffit de réaliser qu'à l'extrême un système n'ayant qu'une seule étiquette atteindra sans difficulté un taux de réussite de... 100%!

Nous reviendrons sur les résultats que nous obtenons dans l'évaluation préliminaire de notre étiquetage (cf. § 20).

8 FORMAT DE L'ÉTIQUETAGE

Nous l'avons dit, les textes annotés peuvent prendre plusieurs formes en ce qui concerne la présentation, la finesse des informations fournies dans les étiquettes ou encore la structuration interne de celles-ci. Mais les données présentées au terme de l'étiquetage peuvent également varier en ce sens que l'on peut avoir une seule étiquette par forme ou plusieurs étiquettes qui subsistent en parallèle.

8.1 Une seule étiquette par forme

Lorsque l'on a une seule étiquette, qui propose la bonne analyse, alors le taux de réussite est de 100 %.

Évidemment, pour reprendre l'exemple donné par Véronis et Khouri (1995), si le jeu d'étiquettes comporte une seule étiquette (on pourrait imaginer l'étiquette <MOT>), alors, l'étiquetage sera toujours réussi.

Sans aller jusqu'à cette caricature, si le jeu d'étiquettes est grossier au point de ne pas mettre en évidence les ambiguïtés – par exemple parce qu'elles sont « diluées », « dissoutes » dans des étiquettes fourre-tout elles-mêmes ambiguës – il est plus facile d'arriver à une seule étiquette qui sera la bonne que si le jeu d'étiquettes reflète bien les ambiguïtés lexicales présentes. Ainsi, on pourrait neutraliser dans une étiquette l'indicatif présent et le subjonctif présent, qui génère un grand nombre d'ambiguïtés pour les verbes en français. De même, on pourrait attribuer un type d'étiquettes particulier aux adjectifs épïcènes.

Dans le cas de notre corpus, on a pour certaines formes amorcées l'étiquette <AMO>. On pourrait considérer que la forme n'est pas ambiguë, puisqu'elle a bien une et une seule étiquette, et que cette étiquette est correcte (on est bien en présence d'une amorce de morphème). On voit néanmoins la lacune de cette catégorisation qui n'a de grammaticale que la forme.

La plupart des systèmes visent à attribuer une seule étiquette, même si celle-ci n'est pas nécessairement la bonne. Pourtant, comme le fait remarquer Laporte (1996 : 299) :

Dans le cas où on ne sait pas lever complètement les ambiguïtés, les systèmes qui donnent plusieurs solutions sont donc plus réalistes.

8.2 Plusieurs étiquettes par forme

Lorsque le système n'a pas réussi à lever toutes les ambiguïtés, plusieurs étiquettes coexistent en parallèle. Nous le verrons (cf. § 11), c'est le cas de notre système.

Ainsi, quand les règles que l'on a formulées échouent à fournir une et une seule analyse pour une forme *virtuellement* ambiguë, on préfère laisser les solutions alternatives au choix de l'annotateur humain qui validera les résultats.

Mais plusieurs étiquettes peuvent subsister (ou *devraient subsister*, dans le cas des systèmes qui fournissent une seule analyse) car elles correspondent en fait à des ambiguïtés *réelles*. C'est le cas des phrases classiques comme *la petite brise la glace* ou *le cuisinier sale la note*. Pour la première phrase, on a les deux analyses possibles après désambiguïsation³² :

catégories grammaticales

1^{re} analyse : {DET} {N} {V} {DET} {N}

2^e analyse : {DET} {A} {N} {PRO} {V}

catégories grammaticales et informations flexionnelles

1^{re} analyse : {DET:fs} {N:fs} {V:P3s} {DET:fs} {N:fs}

2^e analyse : {DET:fs} {A:fs} {N:fs} {PRO:fs} {V:P3s}

³² Nous ne parlons pas ici de toutes les analyses possibles pour cette phrase, à la manière du calcul présenté dans le § 4 de ce chapitre ou encore dans l'automate du texte ci-dessous, mais bien des deux seules bonnes analyses qui reflètent les ambiguïtés réelles de cette phrase.

catégories grammaticales, informations flexionnelles et lemmes

1^{re} analyse : {le,DET:fs} {petit,N:fs} {briser,V:P3s}
 {le,DET:fs} {glace,N:fs}

2^e analyse : {le,DET:fs} {petit,A:fs} {brise,N:fs}
 {le,PRO:fs} {glacer,V:P3s}

Les phrases effectivement ambiguës de ce type sont relativement rares. Dans la relecture partielle que nous avons faite sur notre corpus, nous n'en avons pas relevé. Clément (2001), dans le *Corpus de Paris 7*, n'en relève pas non plus³³.

Dans notre corpus, il est des cas où l'ambiguïté est impossible à lever, non seulement de façon automatisée, mais même par un humain. C'est le cas dans les énoncés inachevés qui laissent parfois le dernier morphème sans analyse catégorielle possible :

chaCL1 (...) et je me suis peut-être fait un jour euh /
 taper sur les doigts ou au moins attiré une réprimande
 d'un instituteur pour avoir dit je vais chercher mon à
 manger euh **le** // alors j'ai cessé de le dire (...)
 [chaCL1r]

On peut formuler l'hypothèse que le *le* est une anticipation de *le dire* qu'on retrouve plus loin dans l'énoncé. Cette hypothèse est probable, car on l'a dit, le français parlé procède souvent par anticipation, avec de nombreux allers et retours sur l'axe syntagmatique. Dans ce cas, il faut analyser la forme comme un pronom. Mais il peut aussi s'agir d'une bribe de syntagme nominal laissé inachevé (il n'est ni repris ni complété dans la suite de l'énoncé). Dans ce cas, la bonne analyse serait celle qui fait du *le* un déterminant. On voit que si l'on veut laisser les interprétations possibles pour les analyses ultérieures, il vaut mieux ne pas trancher dans ces cas, et laisser subsister les hypothèses concurrentes.

C'est évidemment la particularité de notre corpus qui produit de telles occurrences, mais elles participent des ambiguïtés réelles dans les données étiquetées, celles pour lesquelles plusieurs analyses doivent subsister en parallèle si l'on ne veut pas forcer l'interprétation.

³³ Cela dit, on peut formuler l'hypothèse suivante : on n'en relève pas parce qu'elles nous échappent. En effet, l'humain ne perçoit souvent qu'une des deux analyses et ne perçoit pas l'ambiguïté (Fuchs 1996). D'ailleurs, le lecteur a-t-il perçu l'ambiguïté présente dans « on n'en relève pas parce qu'elles nous échappent » ?

9 MOTS INCONNUS

Dans un système comme le nôtre, qui se base sur des ressources lexicales consignées dans des dictionnaires électroniques, sont considérés comme mots inconnus les mots pour lesquels on ne peut fournir une analyse, parce qu'ils ne sont pas répertoriés dans le dictionnaire. Ce problème est central pour les systèmes par dictionnaires.

Ces mots inconnus peuvent être des néologismes, des sigles, des noms propres, des variantes graphiques, des coquilles, des fautes d'orthographes, etc. (Cartoni 2006, Dister 2000a, Dister et Fairon 2004).

On a vu comment on pouvait partiellement résoudre ce problème. Dans le cas de notre corpus, on a construit un dictionnaire électronique qui recense les particularités lexicales du français en Belgique (chap. 5, § 5). Pour les noms propres et les sigles, on a une heuristique de la majuscule qui étiquette ces formes, mais sans indication quant aux informations flexionnelles de genre et de nombre. On a vu par ailleurs que pour une certaine catégorie d'amorces (chap. 3.4), on est incapable de fournir une hypothèse quant au lemme et à la catégorie grammaticale. Dans ce cas, celles-ci reçoivent l'étiquette <AMO> pour *amorce*.

Le formalisme que nous utilisons ne nous permet pas de poser des hypothèses de catégorie grammaticale quant à ces mots inconnus, comme d'autres peuvent le faire (Chanod et Tapanainen 1995b, Brill 1999a). Notre système n'est donc pas un système robuste, qui fournirait une analyse dans tous les cas, et ce malgré l'incomplétude des dictionnaires ou les « imperfections » du corpus.

10 TRAVAILLER SUR L'AUTOMATE DU TEXTE

Nous avons dit que notre étiquetage se faisait en deux temps : un étiquetage *a priori*, suite à la consultation du dictionnaire, puis une désambiguïsation qui vise à réduire le nombre des étiquettes fourni par cette première annotation. Celle-ci ne se présente pas sous la forme d'un texte linéaire ou structuré en colonnes, mais bien sous la forme d'un automate (cf. chap. 4, § 6.1 ; Dister 1999c).

Pour reprendre notre exemple de *la petite brise la glace*, voici l'automate du texte après consultation du dictionnaire :

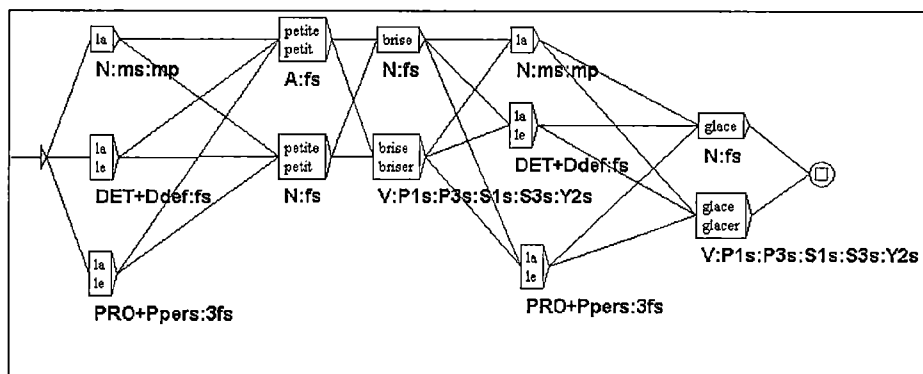


Figure 26. Automate du texte : la petite brise la glace

L'automate se lit de gauche à droite. On a pour chaque forme les analyses qui correspondent aux entrées de la forme dans le dictionnaire. Les formes du texte sont données dans les boîtes (ainsi que le lemme si celui-ci est différent), les informations de catégorie grammaticale et de flexion sont sous les boîtes. La forme *la* reçoit trois analyses (nom, déterminant et pronom) ; la forme *petite* deux analyses (nom et adjectif) ; *brise* reçoit deux analyses en termes de partie du discours (nom et verbe), mais cinq analyses de flexion pour le cas du verbe ; *glace* reçoit deux analyses (nom ou verbe), avec également cinq hypothèses dans le cas du verbe.

La représentation de la figure 26 est analogue à la représentation suivante où les boîtes contiennent les étiquettes complètes :

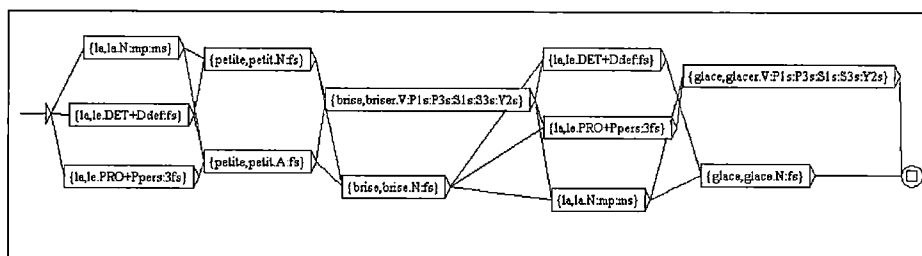


Figure 27. Automate du texte : la petite brise la glace

L'automate du texte ci-dessus ne représente pas visuellement toutes les ambiguïtés. En effet, regrouper les ambiguïtés flexionnelles, si elles se rapportent à une même catégorie grammaticale, sous une seule et même boîte peut occulter le phénomène en jeu. Par rapport à notre définition qui consiste à associer un lemme, une catégorie grammaticale et des informations flexionnelles uniques, on ne considère pas que *brise* est deux fois ambigu mais bien six fois :

- nom féminin (N:fs) ;
- verbe à la première personne du singulier de l'indicatif présent (V:P1s) ;
- verbe à la troisième personne du singulier de l'indicatif présent (V:P3s) ;
- verbe à la première personne du singulier du subjonctif présent (V:S1s) ;
- verbe à la troisième personne du singulier du subjonctif présent (V:S3s) ;
- verbe à la deuxième personne du singulier de l'impératif présent (V:Y2s).

La représentation graphique de l'automate ci-dessus illustre donc mieux ces ambiguïtés de flexion, où l'on voit aussi que *la* est quatre fois ambigu :

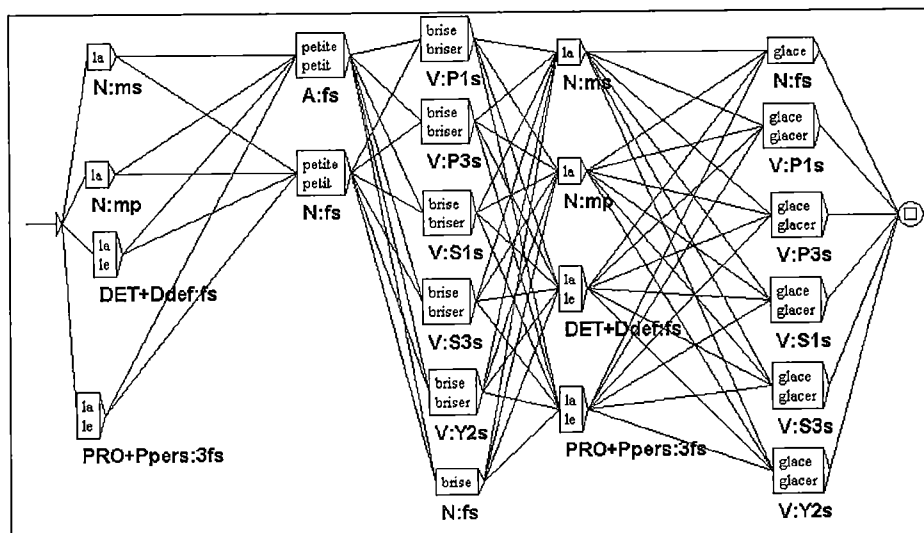


Figure 28. Automate explosé du texte : la petite brise la glace

Voici la seconde manière de représenter les ambiguïtés :

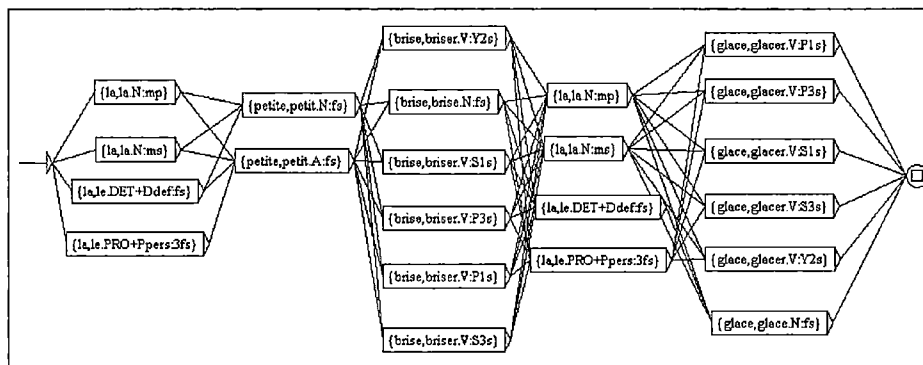


Figure 29. Automate du texte : la petite brise la glace

Cette représentation a le désavantage qu'elle devient vite illisible dès lors que le nombre d'ambiguïtés augmente. On lui préférera donc malgré tout la représentation de la figure 26, où l'automate regroupe dans une même boîte toutes les informations flexionnelles associées à un même lemme et à une même partie du discours.

L'intérêt de travailler sur l'automate du texte est que cela permet de mettre en parallèle les différentes analyses, mais aussi de tenir compte des ambiguïtés de découpage, dans le cas des mots composés. Nous l'avons vu au chapitre 5 (§ 4.5) avec l'exemple de *carte bleue*, où la suite libre est mise en parallèle dans l'automate du texte avec le mot composé.

C'est sur la base de cet automate du texte que va se faire la phase de désambiguïsation proprement dite, que nous allons maintenant décrire.

1.1 ELAG : ELIMINATION OF LEXICAL AMBIGUITIES BY GRAMMARS³⁴

Le système de levée d'ambiguïtés lexicales que nous utilisons est le système ELAG, acronyme de *Elimination of Lexical Ambiguities by Grammars*. L'acronyme est également un jeu de mots puisque, nous le verrons plus loin, le processus de levée d'ambiguïtés est un processus d'élagage de l'automate du texte.

ELAG a été développé initialement par Éric Laporte (Laporte et Monceaux 1999) pour le système INTEX (Silberztein 1993) ; il a ensuite été intégré au système UNITEX par Olivier Blanc (2006).

Le système s'appuie sur des dictionnaires électroniques à large couverture et sur des grammaires locales à base de règles construites par des experts humains.

Par ailleurs, nous l'avons déjà mentionné, cette approche vise à fournir le résultat de la désambiguïsation non pas sous la forme d'un texte linéaire totalement étiqueté, comme c'est classiquement le cas (Brill 1993, Schmid 1994), mais sous la forme d'un automate du texte partiellement désambiguïté (Koskenniemi 1990).

La philosophie d'ELAG pourrait être résumée comme suit : on préfère laisser des solutions en parallèle dans l'automate du texte plutôt qu'obtenir une seule analyse qui serait mauvaise. L'application d'une grammaire ELAG vise donc à augmenter la précision, sans réduire le rappel (cf. § 20).

Le mécanisme d'ELAG est puissant, et il permet à chaque utilisateur d'écrire ses propres grammaires. C'est ce que nous avons fait pour des textes écrits standard du français (Blanc et Dister 2004, Dister 2004) et, dans ce travail, pour des transcriptions du français parlé.

³⁴ Voir aussi Laporte et Monceaux (1999), Blanc (2006 : 44 et sv.) et Blanc dans (Paumier 2006), Dister (2004) pour une description plus détaillée d'ELAG, du formalisme et des exemples de grammaires pour l'écrit. Une partie des informations reprises dans cette section a été publiée dans Blanc et Dister (2004).

Certaines des grammaires que nous présentons dans ce chapitre ont été d'abord écrites par Laporte et Monceaux (1999) lorsque ELAG tournait sous le logiciel INTEX (Silberztein 1993). Nous les avons totalement réécrites, en partie avec Olivier Blanc, lors de l'implémentation d'ELAG sous UNITEX, dans le cadre du projet CONCORLEX (accord de coopération CGRI-FNRS-CNRS 2003-2004). Dans le cadre de ce travail, nous les avons modifiées, parfois fortement, en fonction de notre jeu d'étiquettes. Nous avons écrit également des grammaires strictement pour ce travail. Les noms du (des) auteur(s) figurent dans chaque graphe, même lorsque celui-ci est très éloigné de sa version première.

11.1 Grammaire locale

Une grammaire locale (Gross 1997a) est une représentation par automate de structures linguistiques complexes, difficilement formalisables dans des tables de lexique-grammaire ou dans des dictionnaires électroniques. Visuellement représentées sous formes de graphes (cf. chap. 4, § 6.1), les grammaires locales sont utilisées pour décrire des éléments qui relèvent d'un même domaine syntaxique (Fairon 2000) ou sémantique (grammaire des dates (Maurel 1990), grammaire des expressions de mesure (Constant 2000)).

Les descriptions linguistiques décrites sous la forme de grammaires locales sont utilisées pour une grande variété de traitements automatiques appliqués sur les corpus de texte. Elles sont utilisées notamment pour l'extraction d'informations (Poibeau 2003), le filtrage d'informations (Balvet 2001) ou encore dans des applications en traduction automatique (Fairon et Senellart 1999). Différentes méthodes de désambiguïsation lexicale ont été développées qui mettent en œuvre des contraintes grammaticales décrites à l'aide de ce type de graphes (Silberstein 1993, Dister 1999a, 1999b et 2004, Laporte, 1994a et 1994b, Laporte et Monceau 1999, Roche 1992). Et enfin, l'utilisation de graphes paramétrés en combinaison avec les tables du lexique-grammaire permet d'effectuer une analyse syntaxique pour les phrases simples (Paumier 2003).

Ces descriptions par grammaires locales s'accumulent dans de vastes bibliothèques de graphes (Constant 2003), qui peuvent notamment être utilisées pour rechercher des motifs linguistiques dans des textes (Nakamura 2003).

Les grammaires ELAG sont des grammaires locales dans lesquelles on va décrire des règles contextuelles afin de lever les ambiguïtés lexicales. Les règles concernent la plupart du temps des séquences réduites de mots, puisque le contexte local permet souvent de lever de nombreuses ambiguïtés. Néanmoins, à la différence de certains systèmes qui se limitent à des suites de deux ou trois mots (bigrammes ou trigrammes), l'empan d'une grammaire ELAG n'est pas limité.

11.2 Syntaxe d'une grammaire ELAG

Pour comprendre le fonctionnement d'une grammaire ELAG, il faut avoir à l'esprit que la grammaire s'applique sur un automate du texte, qui représente toutes les analyses possibles pour les mots du texte, analyses fournies par la consultation des dictionnaires.

Rappelons que le texte est découpé en unités plus petites dans une phase de prétraitement. Ces groupes ne sont pas des phrases, mais des unités de

traitement dont nous pensons qu'elles ont une pertinence pour l'étiquetage morphosyntaxique (cf. chap. 4). Dans la suite de ce travail, nous éviterons donc le mot *phrase* pour le remplacer par *unité de traitement (UT)*. Une grammaire ELAG ne s'applique, sauf mention explicite contraire, que dans les limites d'une *UT*.

Les grammaires manipulées par le systèmes ELAG, qui sont donc des grammaires de levée d'ambiguïtés, ont une syntaxe particulière. Une grammaire est composée de deux parties, appelées *si* et *alors*.

La chemin qui contient les boîtes $\langle ! \rangle$ constitue la partie *si* ; celui qui comprend les boîtes $\langle = \rangle$ est la partie *alors* de la grammaire. La partie *si* est classiquement représentée en haut du graphe et la partie *alors* dans le bas du graphe, mais il s'agit d'une convention qui facilite la lecture : inverser ces positions ne change évidemment rien au fonctionnement de la grammaire.

La sémantique d'une grammaire ELAG est la suivante : si on reconnaît dans l'automate du texte la séquence décrite dans la partie *si*, elle doit aussi correspondre à la séquence décrite dans la partie *alors* pour être conservée dans l'automate du texte. Si ce n'est pas le cas, elle est supprimée de l'automate. Pour le dire autrement : on n'a jamais l'analyse faite dans la partie *si*, sauf si elle est vérifiée par la partie *alors*. Pour le dire encore autrement : ELAG élimine l'analyse présente dans l'automate du texte si elle correspond dans le chemin $\langle ! \rangle$ mais pas dans le chemin $\langle = \rangle$.

La grammaire ELAG suivante nous permet d'illustrer le mécanisme :

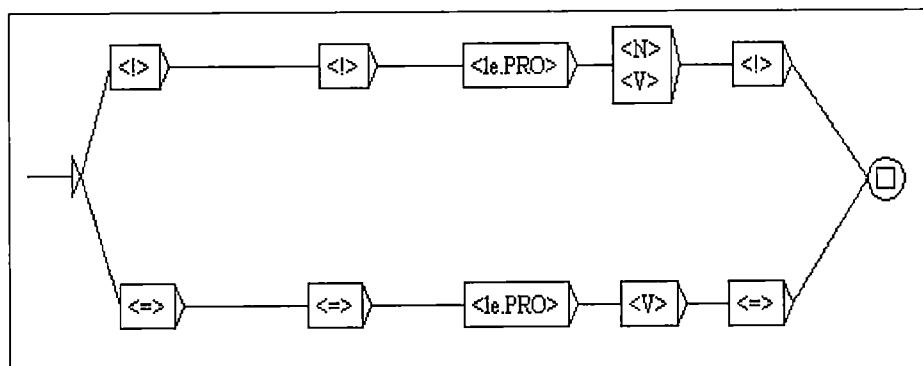


Figure 30. Grammaire de désambiguïsation : la suite $\langle \text{PRO} \rangle \langle \text{V} \rangle$

Si dans l'automate du texte on a le pronom *le* (ou l'une des formes fléchies *la*, *les*, *l'*), suivi d'un nom ou d'un verbe (partie *si*, représentée dans la grammaire par le chemin avec les symboles $\langle ! \rangle$), la partie *alors* (représentée par le chemin avec les symboles $\langle = \rangle$) impose que la forme qui suit *le* (ou *la*, *les*, *l'*) soit un verbe, et l'étiquette $\langle \text{N} \rangle$ est supprimée de l'automate du texte.

Si l'on reprend l'exemple que nous donnions ci-dessus pour la phrase *la petite brise la glace*, l'automate devient alors après l'application de cette grammaire :

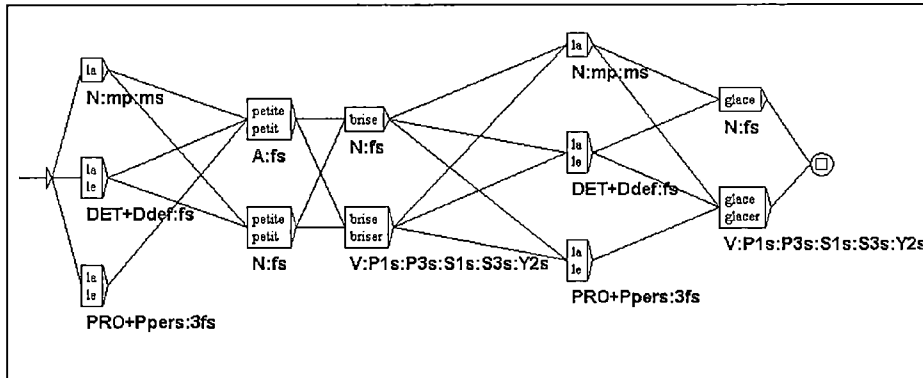


Figure 31. Automate de la petite brise la glace après une première désambiguïsation

À première vue, la grammaire a produit peu d'effets : toujours le même nombre de boîtes, toujours les mêmes hypothèses pour chaque forme. Pourtant, en comparant les deux automates, on constate que le chemin qui reliait *la* <PRO> à *petite* <N> a été supprimé, ainsi que celui qui reliait *la* <PRO> à *glace* <N>.

On a toujours un chemin entre le pronom *la* et l'adjectif *petite* : c'est que notre grammaire ne spécifie rien pour ce contexte. De même, le déterminant *la* relie toujours le verbe *glace*. Là aussi, notre grammaire ne traite pas ce contexte. Pour ce faire, il faut écrire une autre grammaire, comme la suivante :

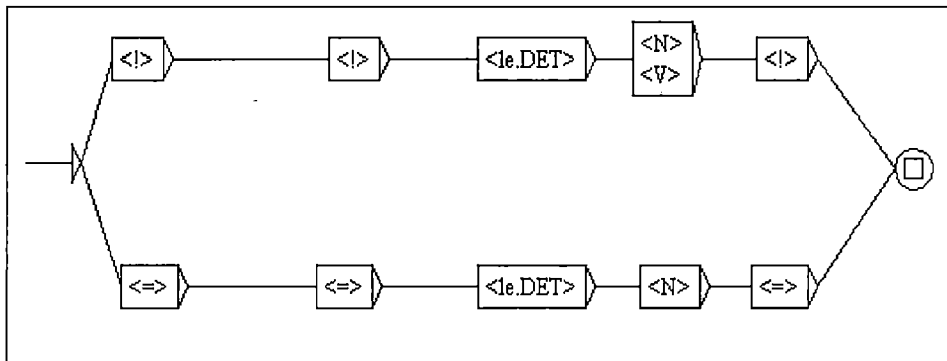


Figure 32. Grammaire de désambiguïsation : la suite <DET> <N>

Appliquée à l'automate du texte précédent, on obtient l'automate suivant :

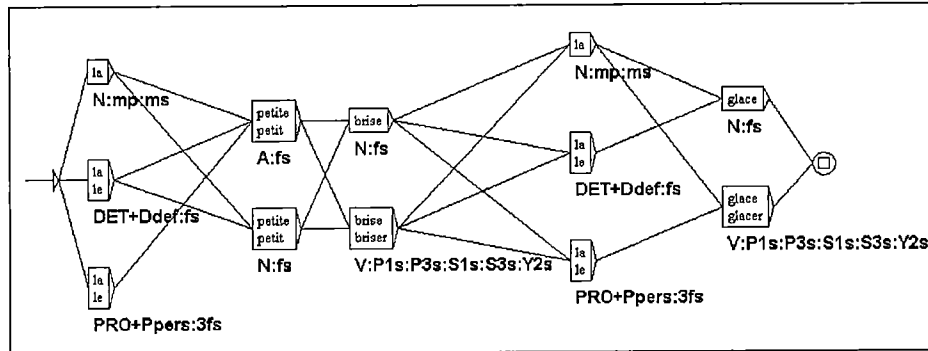


Figure 33. Automate de la petite brise la glace après une deuxième désambiguïsation

Maintenant, le chemin qui relie le déterminant *la* au verbe *glace* est bien supprimé. C'est le principe même d'ELAG que nous venons de mettre en évidence : appliquer des grammaires les unes après les autres afin d'élarguer l'automate du texte petit à petit.

Comme dans l'enquête criminelle d'Hercule Poirot citée en exergue, il s'agit d'éliminer les faux témoignages les uns après les autres, en espérant finalement garder le seul pertinent et trouver le coupable (en l'occurrence, la bonne analyse en contexte).

11.2.1 Le nombre de branches *alors*

On peut écrire des grammaires dans lesquelles on a plusieurs chemins *alors*. C'est le cas de la grammaire suivante :

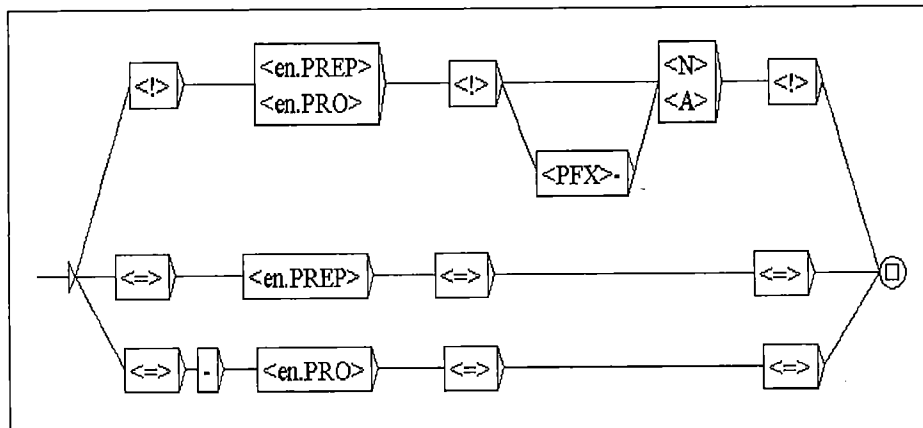


Figure 34. Grammaire en-N.grf

La règle est : « Si l'on a la forme *en* devant un nom ou un adjectif, alors c'est une préposition, sauf si elle est précédée d'un tiret, auquel cas c'est un pronom. »

Néanmoins, comme le précise la documentation du système (Blanc, dans Paumier 2006 : 133), limiter le nombre de branches *alors* permet d'optimiser les grammaires en diminuant le temps nécessaire à leur compilation.

11.2.2 Le point de synchronisation

Dans les deux parties *si* et *alors* d'une grammaire ELAG, on a respectivement les trois symboles $\langle ! \rangle$ et $\langle \Rightarrow \rangle$. Le deuxième de ces symboles fonctionne comme un point de synchronisation entre les parties *si* et *alors*.

La grammaire de désambiguïsation $\langle \text{PRO} \rangle \langle \text{V} \rangle$, que nous avons montrée ci-dessus (figure 30), est équivalente à celle-ci, où l'on ne répète pas dans la partie *alors* la séquence à gauche du point de synchronisation puisqu'elle reste la même :

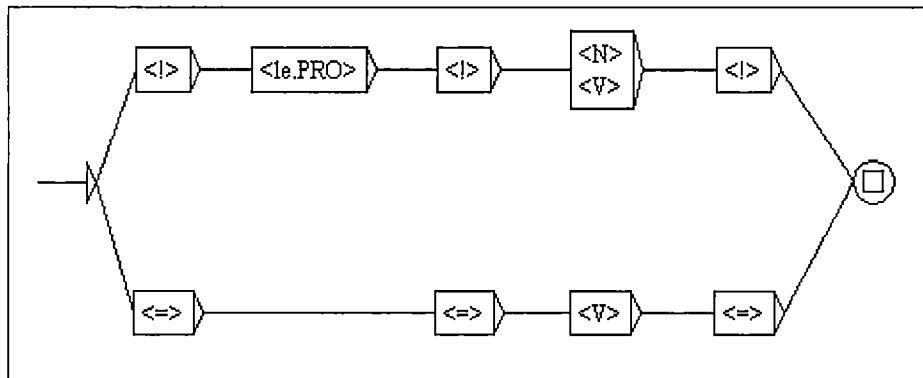


Figure 35. Grammaire de désambiguïsation : la suite $\langle \text{PRO} \rangle \langle \text{V} \rangle$, autre représentation

De même, la grammaire *en-N.grf* (figure 34) ne répète pas le motif donné dans la partie *si*, et noté à la droite du point de synchronisation.

12 LE JEU D'ÉTIQUETTES

Les données sur les mots utilisées dans les grammaires de levée d'ambiguïtés doivent correspondre aux informations présentes dans les dictionnaires électroniques. De ce point de vue, le système de levée d'ambiguïtés est étroitement lié aux ressources lexicales utilisées (tout comme le degré d'ambiguïté est lui-même lié à ces ressources (cf. chap. 5, § 3.1 et chap. 6, § 4).

12.1 Le jeu d'étiquettes : le fichier *tagset.def*

Pour fonctionner correctement, le programme ELAG a besoin d'une description formelle des étiquettes présentes dans les dictionnaires utilisés.

Cette description figure dans un fichier appelé *tagset.def*, qui liste les catégories grammaticales présentes dans le dictionnaire, en associant à chacune d'elles les éventuels codes flexionnels et syntaxiques.

Nous avons décrit au chapitre 5 nos choix linguistiques, liés notamment aux particularités de notre corpus. Ce sont ces choix qui sont formalisés dans le fichier *tagset.def*, dont voici un extrait. La description d'une catégorie commence par *POS* et se termine par la ligne qui contient un point. Une ligne de commentaires débute par le symbole *#*. Le fichier complet est fourni en annexe 7.

déterminants

POS DET

flex:

genre = m f

nombre = s p

cat:

souscat = Ddef Dind Dpos Ddem Dnum Dexi

possesseur = Dpos1s Dpos2s Dpos3s Dpos1p Dpos2p

Dpos3p

contraction = Prepdet

.

adjectifs

POS A

flex:

genre = m f

nombre = s p

.

verbes

POS V

flex:

temps = C F I J K P S T W Y G X B

pers = 1 2 3

genre = m f

nombre = s p

.

interjections

POS INTJ

12.2 La description du jeu d'étiquettes et la construction de l'automate du texte

La description du jeu d'étiquettes faite dans le fichier *tagset.def* a toute son importance. En effet, elle permet de s'éloigner d'une certaine manière des informations présentes dans le dictionnaire en ne définissant que celles qui seront maintenues dans l'automate du texte. Lors de la construction de l'automate du texte par UNITEX, une option permet de normaliser celui-ci en accord avec la description des codes faite dans le tagset.

Ainsi, on peut supprimer d'emblée un certain nombre d'ambiguïtés non désirables qui résultent de la consultation du dictionnaire. Prenons l'exemple suivant, qui concerne la sous-catégorisation des noms dans le dictionnaire. Pour *avocat*, on a les deux entrées suivantes :

avocat,.N+Conc+z1:ms

avocat,.N+Hum+z1:ms

La première entrée concerne le fruit (le trait sémantique « concret »), la seconde entrée l'homme (le trait « humain »). Si l'on décrit ces codes *+Conc* et *+Hum* dans le tagset, on aura les deux chemins en parallèle dans l'automate du texte, puisque les deux analyses sont considérées comme possibles :

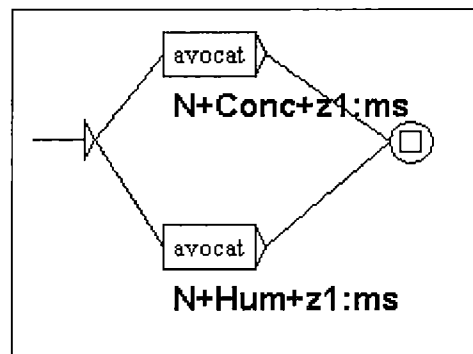


Figure 36. Ambiguïté de avocat dans l'automate

Si l'on supprime ces codes, en ne les intégrant pas dans le tagset comme codes possibles pouvant accompagner la catégorie grammaticale du nom, ces informations sont supprimées de l'automate du texte et les chemins proposés

en parallèle deviennent alors identiques, puisque ce sont ces seules informations qui dédoublaient les entrées. On n'a donc plus au final qu'un seul chemin³⁵ :

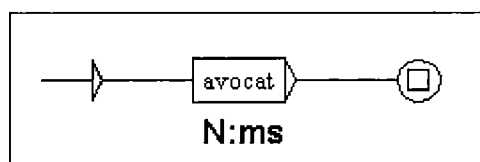


Figure 37. Automate non ambigu

On voit comment la description que l'on fait dans le tagset a déjà un pouvoir de désambiguïsation lors de la construction de l'automate du texte, avant même l'application de grammaires de levée d'ambiguïtés.

12.3 Les masques lexicaux

Les informations linguistiques que nous utilisons dans les grammaires de désambiguïsation ne correspondent pas nécessairement aux entrées lexicales complètes. En effet, il est possible de n'utiliser dans la description qu'une partie des informations présentes dans les entrées lexicales. On parle alors de *masque lexical*. Dans une grammaire ELAG, les masques lexicaux peuvent ainsi spécifier³⁶ :

- **une catégorie grammaticale** : <CAT>
<V> pour l'ensemble de la classe des verbes
- **une catégorie grammaticale et une sous-catégorie** : <CAT+sous-cat>
<DET+Ddef> pour les déterminants définis
- **une catégorie grammaticale et des informations flexionnelles** : <CAT: flexion>
<V:3p> pour les verbes conjugués à la 3^e personne du pluriel
- **un lemme, accompagné obligatoirement de sa catégorie grammaticale** : <lemme.POS>
<être.V> pour toutes les formes fléchies dont le lemme est le verbe *être*

³⁵ Nous n'avons pas intégré dans le tagset les données sur le numéro de couche du vocabulaire (cf. chap. 5), raison pour laquelle ces informations ne seront plus présentes dans l'automate.

³⁶ On garde la typographie de l'entrée : la catégorie grammaticale en majuscules, etc.

- **une catégorie grammaticale à l'exception de certains de ses lemmes** : `<!lemme1!lemme2.CAT>` :

Le masque `<!homme!femme.N+Hum>` reconnaît tous les noms de la sous-catégorie des noms humains, excepté les formes ayant *homme* ou *femme* pour lemme : sont ainsi reconnus *fille*, *scaphandriers*, *metteur en scène*, mais pas *femmes*, ni *réverbère*, ce dernier n'étant pas sous-catégorie « nom humain ».

À la différence de certains systèmes qui utilisent des étiquettes atomiques (van Halteren 1999, Marcus *et al.* 1993, Leech 1997b), nos étiquettes sont des étiquettes structurées. Elles respectent trois critères mis en avant par Leech (1997b : 25) pour le choix des étiquettes :

1. **Conciseness** : Brief labels are often more convenient to use than verbose, lengthy ones.
2. **Perspecuity**: Labels which can easily be interpreted are more userfriendly than labels which cannot. This means that tags should be, where possible, easily remembered : Preposition, by this criterion, is better than IN.
3. **Analysability** : Labels which are decomposable into their logical parts are better (particularly for machine processing, but also for human understanding) than those which are not.

12.4 Les symboles formels

Outre les codes grammaticaux présents dans les dictionnaires électroniques, ELAG utilise des symboles formels (symboles qui font référence à la forme des mots). Ces symboles sont notés en majuscules et entre crochets.

`<NB>` séquence d'un chiffre arabe. Si l'on veut mentionner un chiffre en particulier, celui-ci doit être noté tel quel, accompagné du code `<NB>`. Ainsi, `<3.NB>` reconnaît le chiffre 3. Pour reconnaître une suite de chiffres, il faut boucler sur `<NB>`³⁷ ;

`<PNC>` caractère séparateur, c'est-à-dire ponctuation, blanc ou apostrophe ;

³⁷ Notons que la boucle reconnaît aussi bien un des chiffres collés qu'une suite de chiffres séparés par des blancs.

Dans la mesure où les chiffres et les nombres de notre corpus sont transcrits en toutes lettres, ce symbole est absent de nos grammaires.

-
- <> séquence d'une ou de plusieurs lettres (donc aussi bien les mots répertoriés dans les dictionnaires que les mots inconnus), les signes de ponctuation et les chiffres ;
 - <MIN> séquence d'une ou de plusieurs lettres minuscules ;
 - <MAJ> séquence d'une ou de plusieurs lettres majuscules ;
 - <PRE> séquence d'une lettre majuscule suivie d'une ou de plusieurs lettre(s) minuscule(s) ;
 - <MIX> séquence d'une ou de plusieurs lettres minuscules et majuscules ;
 - <E> séquence vide ;
 - <^> début d'unité de traitement linguistique ;
 - <\$> fin d'unité de traitement linguistique.

Ces symboles formels, très utiles dans l'analyse de textes écrits standard, ne sont pratiquement pas utilisés dans nos grammaires, si ce n'est ceux du début et de fin d'unité de traitement linguistique.

13 MOT SIMPLE / MOT COMPOSÉ

Outre que l'on manque d'un recensement exhaustif des formes composées, la plupart des systèmes d'étiquetage préfèrent les ignorer parce que leur prise en compte multiplie considérablement le nombre d'ambiguïtés. Cela est vrai pour les noms, mais aussi pour tous les mots grammaticaux composés dont nous avons tenu compte (chap. 5, § 4.2).

On peut distinguer trois types de composés :

- ceux pour lesquels, en contexte, subsiste une vraie ambiguïté avec la suite de mots simples. C'est le cas dans *il recouvre la pomme de terre* ou *je l'aime tant que je suis en vie*. Ces cas sont très rares, le contexte permettant généralement de trancher. Néanmoins, lorsqu'ils se produisent, il faut garder les deux analyses ;
- ceux pour lesquels on a une ambiguïté artificielle avec la suite de mots simples, dans la mesure où l'un des constituants ne peut jamais se présenter comme un mot seul. Ainsi les composés *a priori*, *parce que*, *afin de*, etc. ;
- ceux pour lesquels la suite composée doit être privilégiée par rapport à la suite de mots simples. C'est le cas de *grand-mère*, *auto-école*, *cul-terreux*, etc.

Une des solutions pour pallier cette difficulté est de considérer que, dès lors que l'on a un mot composé, cette analyse est préférée à celle qui propose

l'analyse en une suite de mots simples. Cette heuristique est d'ailleurs souvent utilisée dans des systèmes de désambiguïsation (cf. Clément 2001 pour le *Corpus de Paris 7* par exemple).

Dans notre formalisme, il est impossible d'appliquer cette heuristique. Une autre solution serait d'utiliser des dictionnaires prioritaires, dans lesquels les mots composés non ambigus seraient consignés. Si une forme composée est reconnue par un dictionnaire prioritaire de mots composés, alors, ses composants ne sont pas analysés comme des mots simples par la consultation du dictionnaire des mots simples. C'est ce que propose le système INTEX, qui étiquette certains mots composés dans une phase de prétraitement (Silberztein 1993). À l'heure actuelle, cette procédure n'est pas possible avec UNITEX.

Une solution existe pour les mots composés d'un composé non autonome. Ces constituants non autonomes reçoivent le code XI.

Prenons l'exemple de *parce que* dont voici l'automate :

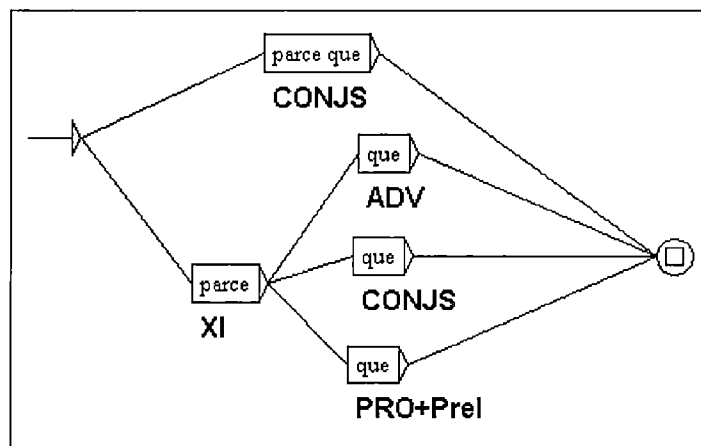


Figure 38. Automate de *parce que*

Parce que est artificiellement ambigu puisque la seule bonne analyse est celle qui analyse la séquence comme un composé

Nous avons donc construit une grammaire ELAG qui supprime du chemin de l'automate toutes les boîtes contenant des XI.

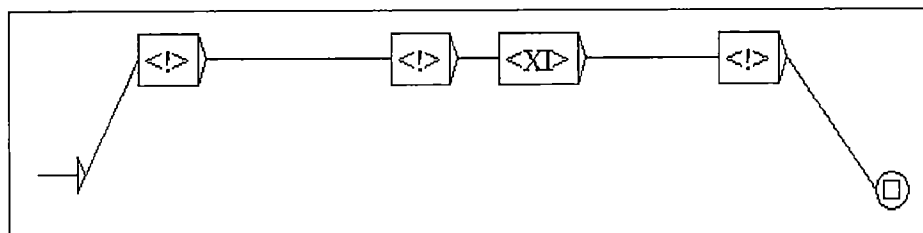


Figure 39. Grammaire qui supprime les boîtes contenant des symboles XI

Cette grammaire est différente des grammaires décrites ci-dessus, puisqu'on n'a que la branche *si* et pas la branche *alors*. La grammaire doit se lire : si on rencontre le symbole XI, alors, on supprime de l'automate du texte le chemin qui le contient.

Dans le cas de *parce que*, en supprimant le chemin qui contient un XI, on garde la bonne hypothèse, celle du composé.

Cette solution serait excellente si l'on n'avait pas, dans notre corpus, des occurrences où seule l'hypothèse XI est proposée dans le chemin de l'automate, sans chemin parallèle avec la forme composée. Cette configuration se produit dans deux cas de figure :

1. le locuteur s'est interrompu dans l'énoncé avant la production du dernier constituant ;
2. une marque est insérée dans le texte, qui ne permet pas de reconnaître la forme composée lors de l'application des dictionnaires. Dans le cas suivant, c'est la marque de début de chevauchement qui est insérée entre les constituants *parce* et *que* ³⁸:

blaAD1 écoute c'est incroyable -| ben nous notre
maison je sais pas combien on la revendrait maintenant
hein / (silence) **parce | - que**
blaNB1 oui -| tu as déjà quand même un avoir (...)
[blaNB1r]

En fait, le problème dans ce cas est que supprimer de l'automate du texte les symboles XI a pour conséquence que l'on supprime le seul chemin dans l'automate du texte, puisqu'on n'a pas le chemin qui propose l'analyse de la conjonction composée.

La solution pour pallier ce problème, qui se produit pour d'autres formes à de nombreuses reprises dans le corpus, est de supprimer les entrées avec les codes XI du dictionnaire. On obtient alors un automate dans lequel la

³⁸ Nous avons dit que cette marque avait été traitée lors du prétraitement et étiquetée de telle sorte qu'elle ne gêne pas l'application des grammaires. Cela est vrai, mais ne concerne pas la reconnaissance des mots composés lors de l'application des dictionnaires.

forme *parce* ne reçoit aucune analyse, car elle est considérée comme un mot inconnu :

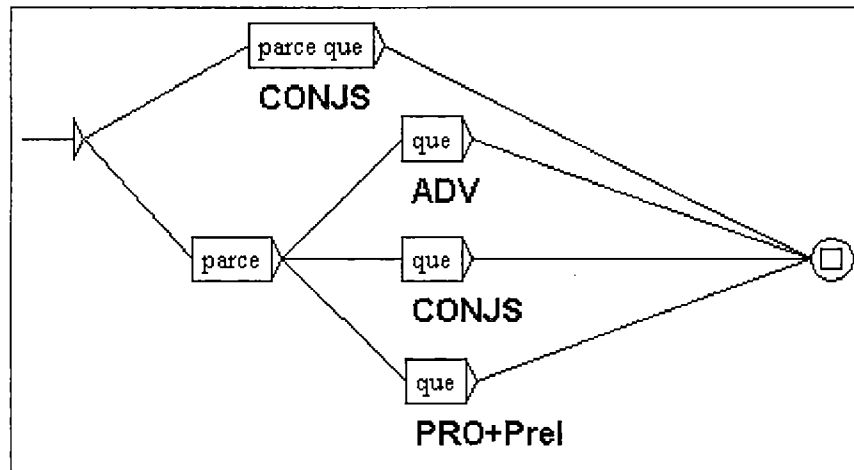


Figure 40. Automate de parce que, avec parce mot inconnu

Si cette solution est satisfaisante car elle ne supprime pas le seul chemin de l'automate dans le cas des énoncés non achevés, elle ne résout pas le problème de l'ambiguïté dans tous les autres cas.

Nous avons donc adopté la procédure suivante, pour tous ces constituants qui ne reçoivent pas de code : on applique la fonction « Clean Texte FST » proposée par UNITEK lors de la construction de l'automate du texte. Celle-ci supprime de l'automate du texte les chemins qui contiennent des mots inconnus s'ils sont en parallèle avec des chemins qui n'ont pas de mots inconnus. Ainsi, dans le cas de *parce que*, cette heuristique permet de supprimer automatiquement les analyses du *que* en pronom, conjonction ou adverbe au profit de la seule bonne analyse de *parce que* comme conjonction de subordination :

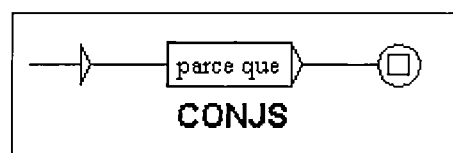


Figure 41. Automate de parce que non ambigu

Si nous insistons un peu sur la procédure adoptée ici, c'est que sa mise au point a nécessité des interventions non seulement sur le dictionnaire, mais

aussi sur la procédure de construction de l'automate du texte, alors que la méthode qui consiste à supprimer les codes XI de l'automate du texte est une solution efficace sur des corpus écrits standards. C'est encore une fois la particularité de nos données qui a voulu que nous procédions de la sorte, pour un processus pourtant éprouvé maintes fois auparavant, sur d'autres données « standard ».

Si l'on a également développé ce point, c'est que l'on a été confrontée aux règles qui suppriment tous les chemins de l'automate et aboutissent aux rejets de la phrase par la grammaire. Nous reviendrons sur cet aspect au § 17 de ce chapitre.

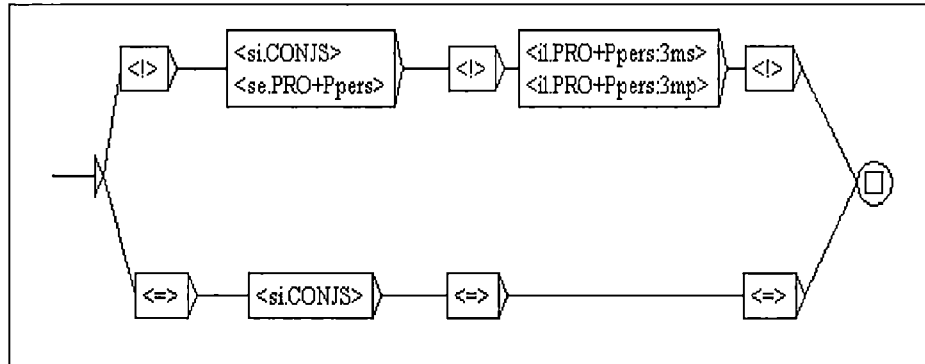
14 INTÉRÊT DE SPÉCIFIER DES SOUS-CATÉGORIES

Nous avons dit que moins le jeu d'étiquettes était fin, plus il était facile statistiquement d'attribuer la bonne étiquette à une forme dans le texte. Néanmoins, le revers est que moins les catégories sont fines, moins il est facile d'écrire des règles contextuelles qui permettent de lever des ambiguïtés. En effet, des catégories fines donnent la possibilité de mieux décrire les contraintes distributionnelles en jeu. Ainsi, en tenant compte de la distinction entre pronom sujet et pronom complément, on peut écrire des grammaires qui se basent sur le trait *PpvIL* (sujet) et qui ne vont concerner le pronom que lorsqu'il est sujet. Nous avons donné les raisons qui nous poussent à ne pas faire cette distinction au chapitre 5 (§ 4.2.2).

Nous allons illustrer l'intérêt de spécifier des sous-catégories avec deux exemples concrets.

14.1 L'ambiguïté du *s* élide

La forme élidée *s'* est associée à deux lemmes différents dans le dictionnaire : *se* (pronom personnel) et *si* (conjonction de subordination). Une règle relativement simple consiste à dire que devant les pronoms *il* et *ils*, le *s'* est la conjonction. Dans tous les autres contextes, il s'agit du pronom. Voici la grammaire *s_elide.grf* qui formalise cette règle :

Figure 42. Grammaire *s_elide.grf*

Chaque fois que l'on rencontrera un pronom *se* (élide ou non) devant *il* ou *ils*, le chemin sera supprimé, pour ne garder que celui avec la conjonction.

Cette règle a une règle corollaire : après la conjonction *s'*, on ne peut rencontrer que les pronoms *il* ou *ils*. Cette règle, qui ne souffre pas d'exception, est impossible à formaliser dans ELAG. En effet, on ne peut référer à une forme dans une grammaire que via sa catégorie grammaticale, son lemme et sa flexion. Ainsi, *s'* doit être formalisé *<se.PRO+Ppers:3s:3p>* ou *<si.CONJS>*. Mais, écrites comme cela, ces étiquettes renvoient également à *se* et à *si*, les formes non élidées. Or, la règle que nous énoncions plus haut ne vaut que pour les formes élidées : en effet, on rencontre aussi la conjonction sous sa forme pleine dans d'autres contextes que celui des pronoms *il* et *ils*. Ainsi, si l'on veut pouvoir écrire une grammaire qui traite les formes élidées et uniquement celles-là, il faut ajouter un code dans les entrées du dictionnaire, qui permet de distinguer forme élidée et forme pleine.

C'est ce que nous avons fait, et les entrées sont les suivantes :

s',se.PRO+Ppers+Eli:3s:3p
s',si.CONJS+Eli

On peut donc maintenant écrire des grammaires qui ne prennent en compte que les formes élidées, en spécifiant le code *Eli*. La grammaire *s_elideBis.grf* impose que dans le contexte droit d'un *s* élide qui est une conjonction de subordination, on ait les formes *il* ou *ils*.

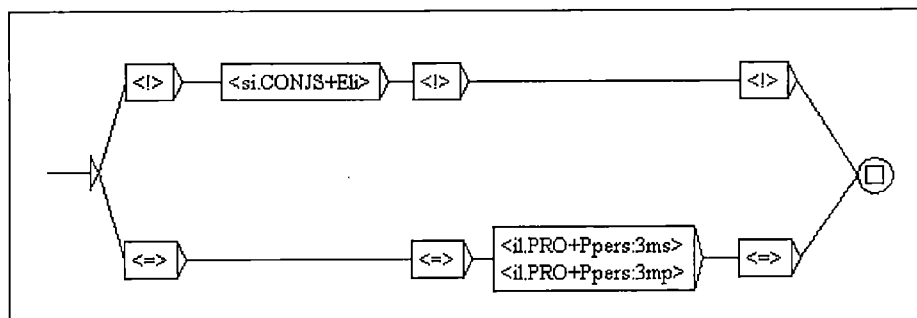


Figure 43. Grammaire s_elideBis.grf

Ce code de l'élision est nécessaire pour la construction de cette grammaire. Dans les autres grammaires pour lesquelles on n'observe pas un comportement différent entre la forme élidée et la forme non élidée (en fait tous les autres cas), on n'indique plus ce code (cf. le concept de masque lexical développé au § 12.3).

La nécessité de spécifier des traits supplémentaires est liée dans ce premier cas au formalisme d'ELAG.

14.2 L'ambiguïté des déterminants numéraux et des noms homographes

Les déterminants numéraux sont ambigus avec les noms homographes. Cette ambiguïté est gênante et systématique pour tous ces déterminants. Nous avons donc construit une grammaire qui lève l'ambiguïté entre les déterminants numéraux et les noms ambigus avec ceux-ci. Il s'agit d'une grammaire fortement lexicalisée, puisque l'on se base sur certaines formes à droite du déterminant : *heures*, *euros*, *ans*, etc. On pose l'hypothèse que devant ces formes, on a bien un déterminant et non un nom³⁹. Dans les dictionnaires du LADL, on a le code sémantique *Nnum* pour tous ces noms (noms numéraux). Le conserver permet de pouvoir considérer ces formes, sans devoir les énumérer toutes dans le graphe, ce qui serait fastidieux. Voici notre graphe, construit sur base des données de notre corpus, et qui demanderait bien entendu à être complété (avec les noms de monnaie ou de mesure par exemple) :

³⁹ On aurait pu mettre simplement la catégorie grammaticale <N> mais le risque d'éliminer de bonnes analyses est trop élevé.

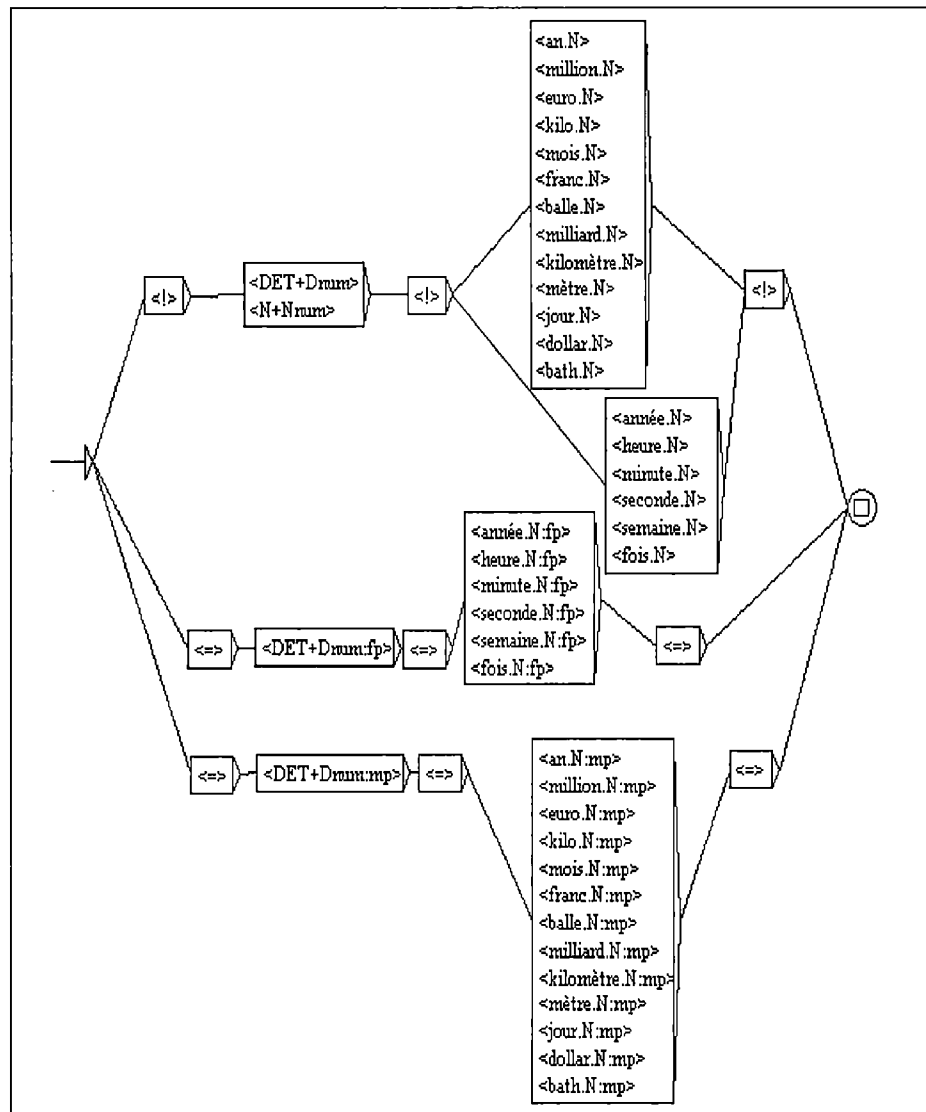


Figure 44. Grammaire DnumNom.grf

L'intérêt de la sous-catégorie n'est pas ici lié au formalisme d'ELAG, comme c'était le cas pour traiter les formes élidées. On voit néanmoins ici tout l'avantage d'une description exhaustive de certains traits du lexique, qui évitent des énumérations fastidieuses dans les grammaires.

15 INDÉPENDANCE DES GRAMMAIRES

Les grammaires ELAG ont ceci de particulier qu'elles peuvent s'appliquer indépendamment les unes des autres. Ce n'est pas le cas de tous les systèmes : certaines des grammaires de désambiguïsation que nous avons

construites avec INTEX (Dister 1999b) étaient dépendantes les unes des autres : l'application d'une grammaire A, sans application de la grammaire B complémentaire, risquait de fournir un étiquetage erroné.

Comme le mentionnent Laporte et Monceaux (1999 : 355) :

(...) the only possible strategy for building a disambiguation grammar is to accumulate separate rules and to use them together, in order to benefit from the addition of their effects.

Ainsi, les règles sont écrites séparément, et les effets des grammaires sont cumulatifs. Quand on ajoute une nouvelle grammaire, on n'a pas à se préoccuper d'une mauvaise interaction avec les grammaires déjà présentes⁴⁰.

Avec le système ELAG, on peut donc envisager que plusieurs personnes écrivent des grammaires séparément. La seule contrainte est l'utilisation des mêmes ressources lexicales et du même jeu d'étiquettes.

L'application des grammaires peut se faire selon deux modes différents : on peut appliquer toutes les grammaires en même temps (on compile les règles), ou les appliquer les unes après les autres, dans un ordre indifférent. Dans ce cas d'application séquentielle, on sauve à chaque étape l'automate du texte partiellement désambiguïsé. Nous reviendrons sur le mode d'application que nous avons choisi (§ 20.2).

16 LES UNITÉS DE TRAITEMENT

Au chapitre 4, nous avons détaillé la phase de prétraitement appliquée à notre corpus, motivée notamment par l'étude des disfluences mais aussi par l'analyse d'autres marques présentes dans le corpus (le point d'interrogation, certaines amorces, etc.). Le prétraitement visait notamment à découper le texte en unités plus petites, que nous appelons *unités de traitement* (cf. chap. 4). Une UT débute et s'achève par les symboles $\{S\}$. L'intérêt de ces unités de traitement est que les grammaires de désambiguïsation ne s'appliquent pas au-delà de ces limites.

Nous avons dit que nous considérons comme pertinents les constituants minimaux lors du découpage en UT (cf. chap. 4)⁴¹. Le début d'un constituant nominal peut être marqué par une préposition⁴². Plusieurs de nos grammaires

⁴⁰ On suppose ici que les règles formalisées dans les grammaires sont correctes, évidemment.

⁴¹ Nous ne voulons pas dire que le texte est découpé en constituants minimaux, mais que le fait qu'il le soit parfois (plutôt qu'en propositions par exemple) n'est pas gênant.

⁴² Font exception les prépositions qui peuvent subir un effacement comme *avec* ou *contre*, qui ne sont pas nécessairement à l'initiale d'un syntagme.

utilisent ainsi la préposition pour contraindre l'analyse du contexte droit. Pour simplifier, on dira que la suite préposition + déterminant ne peut être suivie que par un nom, un adjectif (ou un participe passé) ou encore un adverbe. Mais il arrive que le syntagme ainsi initié ne soit pas complété, comme dans les deux exemples suivants :

blaBD1 mais c'est sans / son truc est sans rapport avec
le // son élection est sans rapport avec le fait qu'il a ce
 rôle |- -là [blaGD1]

accGP1 enfin dans le // même dans ces milieux-là
 hein il y a des différences [accGP1r]

La phase de prétraitement insère les balises suivantes :

{S} {blaBD1,.NLoc} {S} mais c'est sans {/,PAB} son
 truc est sans rapport avec le {S} {/,PAL} {S} son
 élection est sans rapport avec le fait qu'il a ce rôle {|-
 ,IGN+ChO} -là {S}

{S} {accGP1,.NLoc} {S} enfin dans le {S} {/,PAL} {S}
 même dans ces milieux-là hein il y a des différences {S}

Ici, on n'a donc ni un nom, ni un adjectif, ni un adverbe qui suit le déterminant. L'analyse doit donc s'arrêter. Afin de permettre à celle-ci de s'arrêter sans échouer, nous insérons dans la grammaire la possibilité que l'on soit en fin d'UT grâce au symbole {S}. Dans le cas des deux énoncés repris ici, cela évite d'analyser *son* comme un nom et *même* comme un déterminant, analyses qui auraient été conservées si l'on avait reconnu des syntagmes dans les séquences *avec le son* et *dans le même*.

De la même manière, *à* et *sur* se présentent à l'initiale d'un syntagme. Ce n'est pas le cas dans les deux exemples suivants, et le fait de noter dans la grammaire la possibilité d'une interruption après ceux-ci permet de ne pas rejeter ces UT.

ileFN1 dans un un créneau -| bien précis parce que
 c'est vrai que / si on s'enferme dans un créneau bien
 précis en disant moi je ne traite que des problèmes
 d'enseignement par exemple bon euh / c'est vrai que on
 va simplement ouvrir la porte à |- certains <ileGF0> oui
 c'est vrai -| certains abus ou |- on va rester à

norHJ1 vous connaissez qu'il était il avait ses Conseils-
 aux-Wallons puis |- orthophonie n'est-ce pas oui hein
 Remacle <norFA0> oui j'ai / j'ai lu ces deux ces deux
 livres -| connaissait bien aussi a fait de bonnes études
sur // mais // alors euh // vous connaissez l'ouvrage de
 Lafontaine donc |- euh n'est-ce pas je vous dit n'est-ce
 pas oui [norHJ1r]

On a aussi de nombreux exemples de ce type pour les syntagmes verbaux initiés par des pronoms sujet. Nous développerons cela ci-dessous avec le pronom *je* (§ 17.3).

17 UNITÉS DE TRAITEMENT REJETÉES

Il arrive que certaines UT soient rejetées après l'application des grammaires ELAG, alors que les règles formalisées dans les grammaires sont des règles linguistiquement correctes⁴³. Cela se produit quand il n'y a plus de chemin permettant d'aller de l'état de départ à l'état d'arrivée de l'automate de l'UT. Pour le dire autrement, on a trop élagué l'automate. Apparaît au début de celui-ci une boîte qui contient l'information REJECTED.

L'élagage intempestif est dû à trois causes principales.

17.1 Coquilles et fautes d'orthographe

Des coquilles ou des fautes d'orthographe gênent la bonne application des grammaires. La grammaire suivante est une grammaire qui normalement produit un taux d'erreur nul, aussi bien sur des corpus écrits standard que sur des transcriptions de l'oral. La règle dit : « Si on a les adverbes *ci* ou *là* précédés d'un tiret, alors on a dans le contexte gauche un nom. »

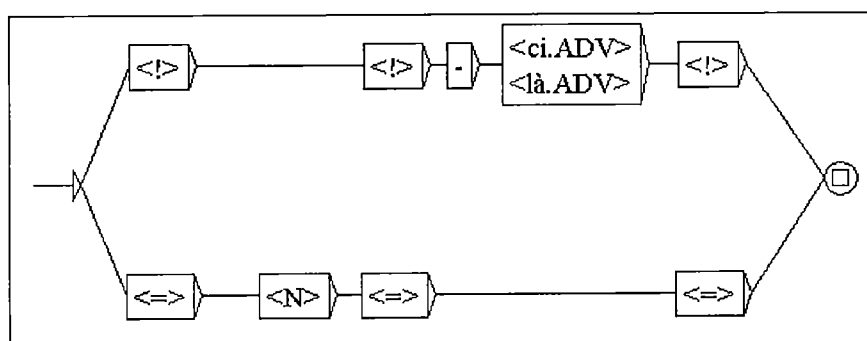


Figure 45. Grammaire Tiret-ci-la.grf

Cette règle est correcte, et les séquences comme *celui-là*, *jusqu'ici*, etc. n'y dérogent pas : ces formes sont consignées comme des mots composés

⁴³ Évidemment, en testant une nouvelle grammaire, il arrive que certaines UT soient rejetées parce que les règles sont fausses, incomplètes, mal définies, etc. De nombreux tests permettent d'éviter ces problèmes et de construire des grammaires pertinentes. Nous parlons ici de problèmes liés aux données et non aux grammaires dans leur phase de test.

dans les ressources lexicales. L'automate de *celui-là*, sans désambiguïsation, est le suivant :

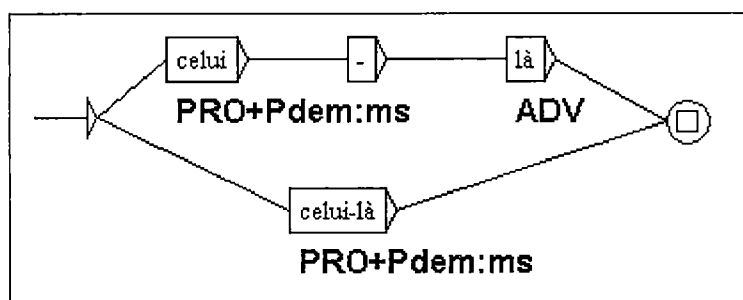


Figure 46. Automate de celui-là

Après l'application de notre grammaire, le chemin du dessus est supprimé puisqu'on n'a pas un nom dans le contexte gauche de *-là*.

Non seulement cette grammaire désambigüise les séquences *image-là*, *classe-là*, *taille-là*, etc., où le nom est ambigu avec un verbe, mais elle lève l'ambigüité entre un mot composé recensé dans les dictionnaires et une suite de mots simples, puisqu'elle supprime le chemin de la suite libre.

Dans notre corpus, cette grammaire rejette pourtant une UT, à cause d'une erreur orthographique :

famED1 quand tu as acheté -| les deux **autres-là** tu
en as acheté quatre ou six en |- même temps [famES1r]

En effet, la seule analyse proposée pour *autres* est adjectif :

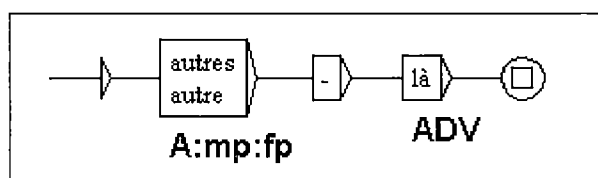


Figure 47. Automate de autres-là

Dans la mesure où l'on impose que la forme à gauche du tiret soit un nom, il y a incompatibilité entre la règle de la grammaire et la description dans l'automate du texte. Le chemin de l'automate est coupé et l'UT est rejetée.

De la même manière, nous avons une grammaire qui supprime les chemins contenant des apostrophes, puisque celles-ci sont maintenant

intégrées directement dans les formes codées dans le dictionnaire. Nous avons dans le corpus *quelqu'un* écrit avec un espace après l'apostrophe. Cela a pour effet que la séquence n'est pas reconnue comme une forme composée, bien qu'elle possède une entrée dans le dictionnaire. Le seul chemin est donc celui qui analyse une suite de deux formes simples séparées par une apostrophe. Lorsque l'on supprime celle-ci, on supprime le seul chemin possible dans l'automate et l'UT est rejetée.

17.2 Incompatibilité liée aux données des dictionnaires

Une deuxième raison qui explique la coupure de tout chemin dans l'automate est liée à nos ressources linguistiques.

Prenons l'exemple de la grammaire suivante, dont la règle est : « Si on a un verbe après une préposition, celui-ci doit être à l'infinitif, au gérondif ou au participe passé. Certaines prépositions peuvent être suivies d'un verbe conjugué, il s'agit des prépositions avec effacement possible. »⁴⁴

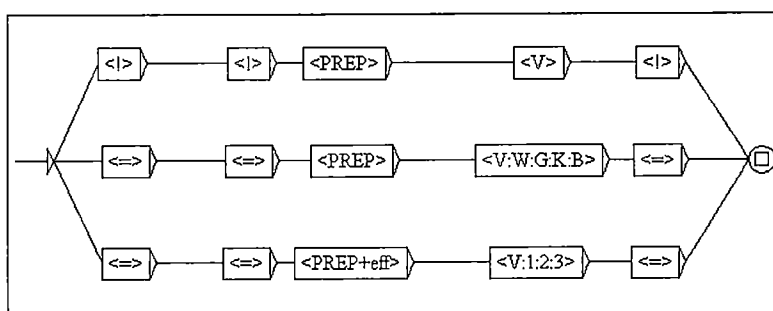


Figure 48. Grammaire PrepVAnne.grf

Appliquée à notre corpus, cette grammaire rejette l'UT suivante : *à brûle-pourpoint*, dont voici l'automate après consultation des dictionnaires :

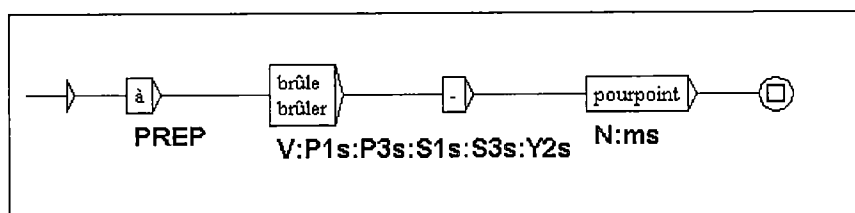


Figure 49. Automate de à brûle-pourpoint

⁴⁴ Cf. chap. 5, § 4.2.6.

Il apparaît clairement que ce n'est pas notre grammaire qui doit être remise en cause mais le fait que *à brûle-pourpoint* n'est pas recensé comme une forme composée dans nos ressources lexicales. Ici, la suite de mots simples contrevient à la règle énoncée dans notre grammaire et la séquence est rejetée.

Les deux exemples ci-dessus pourraient donner l'impression que les problèmes sont liés au figement. Or, on rencontre ces problèmes indépendamment de la non-prise en compte des composés, comme dans l'exemple suivant :

blaJV1 non non je marchais ah non non j'ai pas couru /
et j'ai fait encore un détour pour aller trouver une clope
/ chez au truc **à pitas** quoi / c'était fermé donc j'ai fait
même un détour / jusqu'à l'Espérance puis je suis
revenu [blaJV1]

Voici l'automate pour la séquence qui nous intéresse :

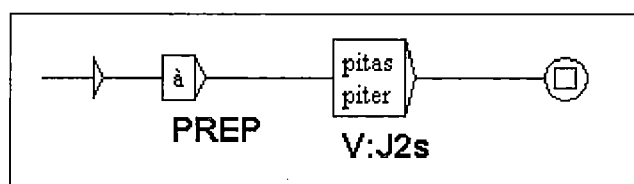


Figure 50. Automate de la séquence à pitas

La seule analyse proposée après consultation du dictionnaire pour *pitas* est le verbe *piter* à la 2^e personne du singulier du passé simple. Il y donc incompatibilité avec la règle de la grammaire et l'UT est rejetée.

On voit avec ces exemples qu'ELAG est un système « sensible ». Non seulement il ne cherche pas à proposer une solution unique à tout prix (nous l'avons dit, c'est un avantage), mais il affiche son incapacité à le faire devant les contradictions générées par les trois types de données dont il dispose : le texte (sous la forme de l'automate), les ressources lexicales et les grammaires de désambiguïsation. Néanmoins, la manière dont ELAG affiche son impossibilité à traiter la séquence présente elle aussi un avantage indéniable : au lieu d'attribuer une mauvaise étiquette, la grammaire n'est tout simplement pas appliquée et l'automate du texte est laissé tel quel.

17.3 Les inachèvement syntaxiques

La troisième cause de rejet n'est pas liée à des erreurs graphiques ou à des lacunes dans les ressources lexicales mais bien à la particularité même

de nos données. En effet, les inachèvements syntaxiques, sur lesquels nous avons déjà beaucoup insisté dans ce travail, sont la cause majeure des rejets d'UT.

Dans le cadre de l'analyse des disfluences faite au chapitre 3, nous avons tenté de trouver des marques formelles qui nous permettraient d'identifier les inachèvements syntaxiques, définis comme des inachèvements dans la construction d'un constituant minimal (et non d'une unité qui s'approcherait de celle de la « phrase »). On a vu que les analyses que nous avons faites, basées sur les disfluences étudiées, ne nous ont pas permis d'isoler des marques ou des séquences de marques qui seraient de bons indicateurs d'une rupture. Au cours de ce travail, nous avons montré de nombreux exemples d'inachèvements. Dans les exemples donnés au point précédent, c'est la marque $\{S\}$, notée à l'emplacement d'une rupture, qui permet d'éviter de mauvaises analyses.

Nous allons illustrer les problèmes posés par les inachèvements avec la grammaire du pronom sujet *je*. Celui-ci initie un constituant verbal. La grammaire pour le français écrit est la suivante :

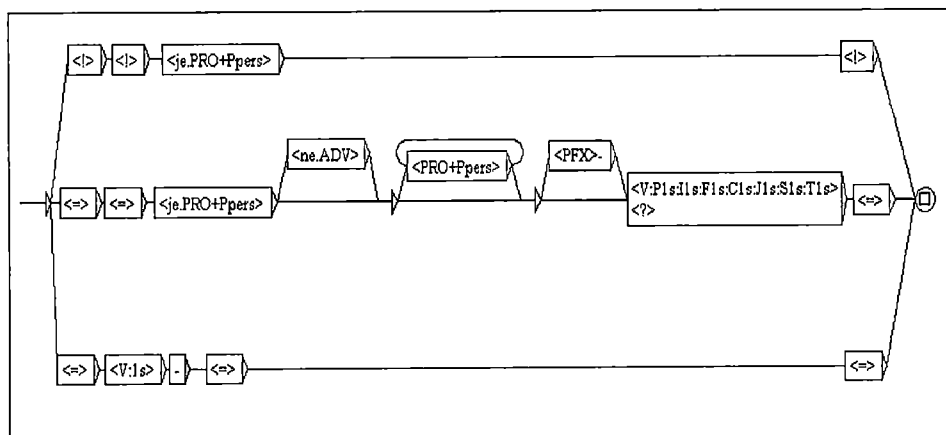


Figure 51. Grammaire je.grf de l'écrit

Cette grammaire est à lire comme suit : « Dans le contexte droit d'un pronom *je*, on ne peut avoir qu'un verbe conjugué à la première personne du singulier, éventuellement précédé d'un *ne* de négation et/ou d'un (ou de plusieurs) pronom(s) personnel(s) ; dans le cas où le *je* est précédé d'un tiret, on ne dit rien sur le contexte droit. »

Cette grammaire reconnaît les séquences suivantes :

je porte
je la porte
je lui en porte

je le lui porte
j'en porte

et lève les ambiguïtés sur les formes ainsi identifiées : non seulement on reconnaît la catégorie grammaticale de *porte*, mais également une partie de sa flexion (première personne du singulier)⁴⁵ ; les pronoms personnels ambigus – dans ces exemples, avec le pronom (*le*, *la*), mais aussi avec la préposition (*en*) ou le participe passé (*lui*) – sont également désambiguïsés.

Par contre, lorsque le *je* est précédé d'un tiret comme dans :

puis-je vous aider ?
suis-je en classe ?

on ne modifie pas l'automate du texte, alors que l'on a bien dans le contexte droit des formes qui pourraient théoriquement être analysées respectivement comme un pronom personnel et comme un verbe (et qui le sont d'ailleurs dans le premier des exemples).

Dans un texte de français standard écrit, cette grammaire est une grammaire qui présente un taux d'erreur nul : non seulement, elle analyse correctement toutes les séquences qu'elle prend en considération, mais en plus, elle ne rejette aucune phrase.

Ce n'est pas le cas dans notre corpus, puisque cette grammaire rejette des unités de traitement. C'est le cas de la séquence suivante :

norHJ1 hein -| alors vous voyez // bon / cela dit donc
je moi je crois vous savez **je / m bon** peut-être qu'on
allait (...) [norHJ1r]

ilpMJ1 (...) parce que **je moi je** suis mal à l'aise quand
je ne sens pas le public [ilpMJ1r]

On est là devant un format de répétition que nous n'avons pas pris en compte (cf. chap. 3.2). Les exemples de ce type ne posent guère de problème puisque l'on peut insérer dans la grammaire le pronom *moi* et revenir ensuite au *je* avant de débiter véritablement l'analyse de la séquence (cf. figure 52).

Dans notre grammaire, nous ajoutons aussi la possibilité qu'une interjection suive le *je*. Cette présence possible d'une interjection règle le problème rencontré pour la seconde occurrence de *je* dans le premier exemple ci-dessus, ou encore celles des énoncés suivants :

⁴⁵ On ne peut évidemment rien dire du temps. Ainsi pour reprendre l'exemple ici donné de *porte*, on reconnaît qu'il s'agit d'un verbe à la 1^{re} personne (*porte* peut aussi être un nom, ou un verbe à la 3^e personne du singulier), mais on ne dit pas s'il s'agit de l'indicatif présent ou du subjonctif présent.

ilrLD3 ah ouais / ça ne me fait ni chaud ni froid quoi je ne suis pas plus content / il y en a ils voient la mer ils sont tout contents moi je bon / ça m'est égal quoi (...) [ilrLD3]

chaBP1 (...) et c'est parti hein / ça je pff / effarant quoi [chaBP1r]

chaBP1 (...) ça reste hein il y a comme ça des choses qui restent / savez en en Allemagne en captivité lorsque l'on vit / euh ensemble jour et nuit on ne peut pas se cacher / l'un à l'autre n'est-ce pas hein on est réellement tel qu'on est ici vous vous mettez sous un bel aspect moi aussi hein je / hein mais mais autrement quoi / et alors on s'aperçoit qu'il y a des tas de / de valeurs qui s'écroulent (...) [chaBP1r]

Voici notre grammaire du *je* :

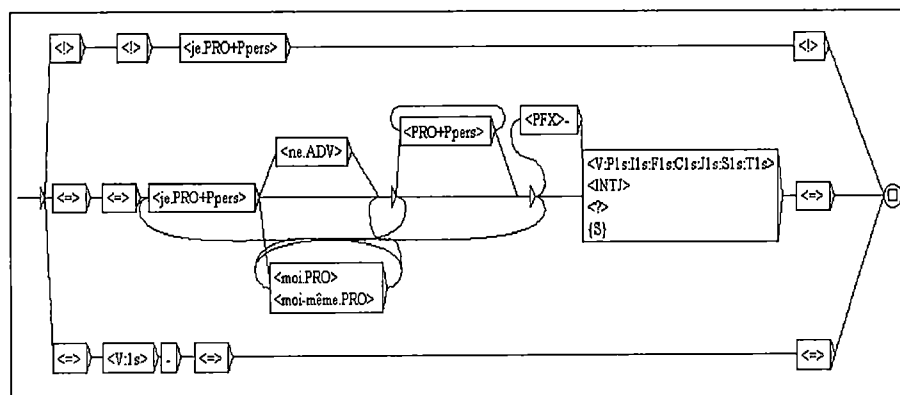


Figure 52. Grammaire je.grf de l'oral

Pour les cas mentionnés ci-dessus, cette grammaire fonctionne parfaitement. Par contre, il existe une série d'UT rejetées par cette grammaire et pour lesquelles nous n'avons aucune solution. Il s'agit des inachèvements syntaxiques qui n'ont pas été repérés dans le texte. Les exemples suivants en sont illustratifs :

(1) **blaAT1** (...) enfin j/ / jamais je jamais enfin maintenant j'ai l'impression que j'étais inconsciente (...) [blaAT1l]

(2) **blaJV1** bè si ben là je c'est parce que elle avait un problème de boiler et j'ai été lui lui remettre un nouveau quoi (...) [blaJV1l]

(3) **blaNB1** (...) je suppose qu'ils ont été au courant / je n'en sais rien / et c'est comme ça que / simplement le on / le le terrain mais c'est grand je c'est / je suppose que ça descend jusqu'aux étangs (...) [blaNB1]

(4) **blaJV1** non non non du tout non -| mais j/ je vraiment c'est pas un reproche et je l'ai dit à et je l'ai dit à ta soeur d'ailleurs (...) [blaJV1]

(5) **blaND1** mais je franchement je ne / sais pas / comment (silence) enfin comment il faut f/ [blaAT1]

(6) **blaGD1** (...) ben oui pour avoir récupéré sa mise et cetera mais je / au niveau de la bourse / tu aurais acheté des trucs il y a // quatorze mois d'ici [blaGD1]

(7) **blaJV1** ouais non -| non non il dit quoi je / déjà au quinze août il était venu me trouver j'étais il il était mort et moi aussi / il m' a demandé euh mon défi pour l'année prochaine (...) [blaJV1]

(8) **famVV1** mais vendredi je / euh Olivier arrive vers vingt-et-une heures |- à Bruxelles (...) [famVV1r]

(9) **famVI1** j'avais pas de carte Mastercard / parce que j'avais une carte Diners {mais je} / Diners ils n'acceptent pas |- au Laos (...) [famVV1r]

(10) **chaHI1** ben (silence) non euh / ça / comment est-ce que je vais expliquer ça / dans le temps / les filles je / couraient quand même un peu après les garçons / ou les garçons couraient après les filles (...) [chaHI1r]

(11) **blaJD1** (...) et ils avaient déjà été à l'hôpital on leur avait laissé voir cinq minutes / ah ça (xx) / je ne / on ne ferait pas ça ici (...) [blaJD1]

(12) **blaAD1** ben / parce que / si moi je suis pas là le ma/ ben tu vois je com/ |- comme je pars quand même souvent [blaNB1]

Dans les exemples 6 à 11, on a une pause brève après le *je*. On pourrait modifier notre phase de prétraitement et considérer que la pause brève est également un endroit de l'énoncé où l'on insérerait une marque {S}. Cela étant, cette modification n'apporterait pas de réponse au problème posé par les énoncés 1 à 5 pour lesquels on n'a aucune marque après le *je*. L'exemple 5 illustre un format de répétition que nous n'avons pas pris en compte : considérer n'importe quelle insertion possible entre répétable et répété, nous l'avons dit, occasionne un bruit ingérable.

Pour l'exemple 9, le chercheur a noté qu'il n'était pas certain de sa transcription. À la lire dans un autre contexte que celui de cette section, on pourrait penser qu'il se trompe assurément. Or, les autres constructions de ce type illustrées ici montrent que cette transcription est tout à fait plausible.

Le cas de figure illustré en 11 est sans doute l'une des occurrences pour lesquelles une amélioration de la phase de prétraitement pourrait apporter une solution. En effet, nous avons balisé un certain nombre d'autocorrections, et notamment celles qui concernent les pronoms sujets (chap. 3.3, § 2.2). Mais on s'est limitée aux pronoms contigus. En étendant le *reparandum* à l'adverbe de négation qui suit le pronom sujet, on réglerait le problème posé par cet exemple.

En ce qui concerne l'exemple 12, ce sont également les choix posés lors de la phase de prétraitement qui sont en cause. En effet, ils concernent les amorces que nous avons baptisées *amorces bégayantes*. Cette catégorie est constituée par les amorces complétées directement et pour lesquelles il y a identité graphique entre l'amorce et le *repair*. Celles-ci reçoivent l'étiquette *IGN+Amo*. L'étiquetage de l'énoncé après le prétraitement est le suivant :

{S} {blaAD1,.NLoc} {S} ben {/,PAB} parce que {/,PAB}
 si moi je suis pas là le {S} {ma/,AMO} {S} ben tu vois
je {<a> com/ ,.IGN+Amo} {|-,.IGN+ChO}
comme je pars quand même souvent

Si l'amorce est bien complétée, le problème est que le *repair*, *comme*, n'appartient pas aux catégories décrites dans le contexte droit du *je*, et que l'UT est donc rejetée. Ce choix de traitement de l'amorce bégayante, mis en défaut dans cet énoncé, s'avère pourtant pertinent dans la quasi-totalité des occurrences, et ce contre-exemple ne suffit donc pas à le remettre en cause. Nous allons maintenant passer en revue certaines des solutions apportées par le prétraitement dans notre corpus.

18 LES ÉTIQUETTES <IGN>

Le prétraitement que nous avons appliqué à notre corpus visait à le scinder en unités plus petites mais aussi à ajouter des étiquettes <IGN> (pour « ignorer »). Le but est que les séquences ainsi balisées soient ignorées par les grammaires de levée d'ambiguïtés. Concrètement, cela signifie que si, dans le chemin de l'automate, on rencontre une boîte dont la catégorie est <IGN>, la grammaire fait comme si cette boîte n'existait pas.

Prenons l'exemple de l'énoncé ci-dessous :

accFJ1 (...) quand euh on s'aperçoit que j'ai un accent
 // **je n/ ne** fais pas en sorte de le perdre / mais alors à
 ce moment-là j'utilise un autre registre de vocabulaire
 [accFJ1r]

Comme dans le dernier énoncé de la section précédente, on est devant une amorce bégayante. Celle-ci reçoit donc l'étiquette *<IGN>* au prétraitement. La portion de l'automate qui nous intéresse pour notre propos est la suite :

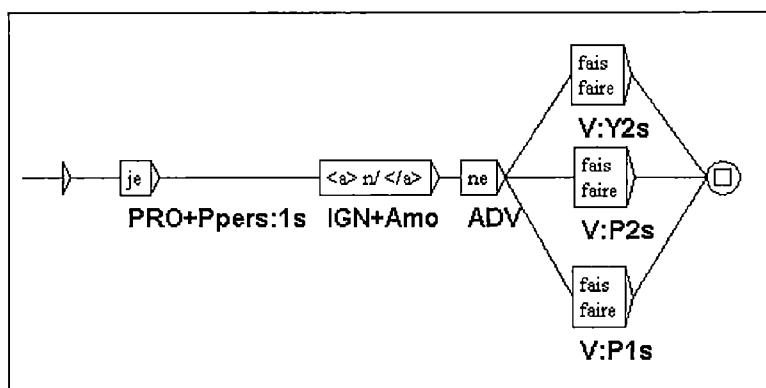


Figure 53. Automate de je n/ ne fais

Si l'on applique la grammaire *je.grf* ci-dessus, on voit que dans le contexte droit du *je*, on doit avoir un *ne*, un pronom, un verbe à la première personne du singulier, un mot inconnu, une interjection ou le symbole de fin d'UT. Or ici, on a une forme catégorisée *<IGN+Amo>*. Ainsi, en fonction du mécanisme d'ELAG, notre grammaire devrait rejeter l'UT dans laquelle figure cette portion de texte. Or, il n'en est rien. En effet, le formalisme d'ELAG a été modifié⁴⁶ afin d'ignorer dans l'automate les étiquettes *<IGN>*. Dans le cas présent, tout se passe comme si, à la droite du *je*, on avait l'adverbe *ne*. On voit l'intérêt d'un tel mécanisme. En effet, à la différence de la phase de prédiction faite par Valli et Véronis (1999), notre texte garde toutes les données de départ.

Pour l'application correcte de la grammaire *je.grf*, on peut encore citer les exemples suivants, dans lesquels les marques de début et de fin de chevauchement sont insérées à l'intérieur du constituant verbal et ont été étiquetées *<IGN>*:

norIN1 (...) je ne suis pas frappée par une entité précise en telle ou telle région **je |- trouve** qu'il y a des
norFA0 l'accent parisien -| aussi

ilcMF0 c'est cela / |- et

ilcCF1 mais **je -| ne connais** pas beaucoup d'hommes provenant de Bastogne donc euh |- (rire)

⁴⁶ Cette modification, tout comme les autres changements dont nous avons besoin, a été faite par Olivier Blanc que nous remercions encore.

Lors du prétraitement, on extrait les chevauchements internes en laissant la trace de l'emplacement grâce au signe @ et à un numéro d'ordre (cf. chap. 4, § 5.3). Ces marques sont également étiquetées <IGN>. Voici la séquence initiale et la séquence après la phase de prétraitement :

norKJ1 (...) comment on prononce tel nom de village /
|- euh là-bas bè je le <norFA0> (souffle en riant) m
-l saurai (...) [norKJ1r]

norKJ1 (...)comment on prononce tel nom de village
{/,PAB} {-,.IGN+ChO} {euh,.IGN+Euh} là-bas bè je le
{@1416,.IGN+Loc} {-|,.IGN+ChF} saurai (...) [norKJ1r]

Dans l'extrait ci-dessus, deux étiquettes <IGN> se suivent, à l'intérieur d'une séquence. Cela ne pose aucun problème pour que celle-ci soit reconnue.

Dans les énoncés suivants, c'est le balisage de l'autocorrection, elle-même étiquetée <IGN>, qui permet de ne pas analyser le pronom *je* comme le début d'un constituant verbal :

ilrTC1 c'est embêtant enfin parce que <c> je </c> on
perd un an [ilrTC1r]

ilpHJ1 mais <c> je consi/ </c> j'estime je suppose
que les autres sont habitués de m'entendre (...)

ilrLP1 oui oui non je sais <c> je s/ </c> il y a de
toute façon un contact qui doit passer (...) [ilrPL1r]

La répétition a été repérée dans les énoncés suivants, que celle-ci soit séparée ou non par une amorce de morphème :

ileDC1 (...) pour ma part euh <r> je ne </r> je ne
veux pas euh j'appartiens pas au terrorisme linguistique
/ consistant / à traquer euh la faute chez les gens (...) [ileDC1r]

iljVM1 et bien moi je crois que <r> je pou/ </r> je
l'encouragerais (...) [iljVM1r]

Dans l'énoncé suivant, on a trois *je*, que la grammaire de l'écrit n'aurait pas pu analyser et où les séquences auraient été rejetées :

iljBB1 (...) moi je pff mes enfants ne parlent pas le
wallon / euh moi je le / parle très peu / mal /
comprends bien mais je <r> le / </r> le parle mal /
en pluS c'est quoi le wallon (...) [iljBB1r]

Ces trois cas de figure sont correctement gérés par notre grammaire je de l'oral :

- le 1^{er} *je* est suivi d'une interjection, ce que permet notre grammaire ;
- la séquence *je PRO V* est interrompue par une pause brève, étiquetée <IGN>;
- le 3^e *je* est suivi par un pronom répété et une pause, également étiquetés <IGN>.

Enfin, des cas comme le suivant, où le *je* est suivi d'une pause longue, sont également bien gérés, puisque le symbole de fin d'UT est inséré :

accGP1 moi je {S} {//,PAL} {S} peut-être que ça peut être dévalorisant (...) [accGP1r]

On le voit avec les exemples ci-dessus, la phase de prétraitement facilite considérablement la tâche d'étiquetage. Si des problèmes se présentent, ils ont pour conséquence de rejeter des UT : c'est-à-dire que la grammaire ne s'appliquera pas sur elles. Le résultat en est donc que l'UT ne sera pas désambiguïsée par la grammaire, ce qui, de notre point de vue, est beaucoup moins gênant que si un étiquetage erroné était fourni.

1 9 SUGGESTION POUR UNE AMÉLIORATION DU FORMALISME

À notre sens, on pourrait encore améliorer le formalisme d'ELAG en permettant de considérer des catégories selon des ambiguïtés spécifiques. À l'heure actuelle, on ne peut pas traiter dans une boîte un type d'ambiguïté particulier en termes de catégorie, mais on peut le faire au niveau des formes fléchies ou des lemmes. Prenons l'exemple suivant : on peut dire que la forme ambiguë *marche* (nom ou verbe) est un verbe après *ça*. La grammaire prend la forme suivante :

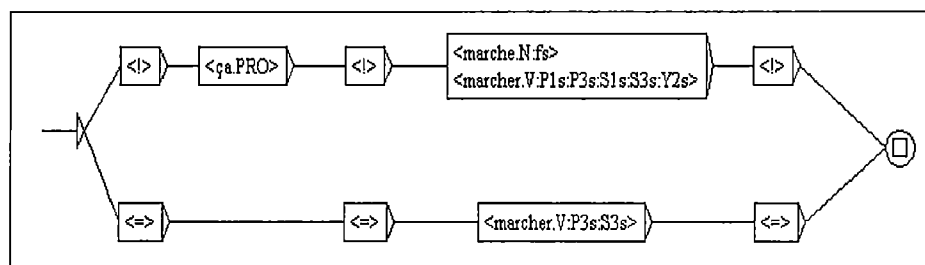


Figure 54. Grammaire ça marche

Mais si au lieu de noter la forme précise *marche* on veut généraliser à toute la catégorie, on a le graphe suivant :

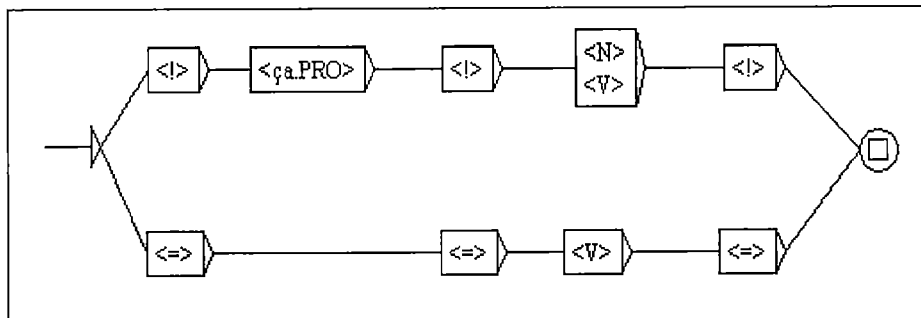


Figure 55. Grammaire *ça marche* généralisée aux catégories *N* et *V*

Or, cette grammaire ne veut pas dire que si on a une forme ambiguë entre un nom et un verbe après *ça*, on choisit le nom. La grammaire dit : « Pour les séquences *ça* *<V>* et *ça* *<N>*, on ne garde que les chemins *ça* *<V>*. » Les chemins qui relient *ça* et *<N>* sont supprimés.

Dans le cas de *tu dois vérifier si ça marche*, l'une ou l'autre grammaire est équivalente. La deuxième grammaire fonctionne aussi très bien dans beaucoup d'occurrences du corpus comme *ça fait*, *ça passe*, *ça reste*, etc.

Par contre, pour la séquence suivante, la deuxième grammaire a d'autres implications :

blanB1 (...) c'est synonyme de vacances hein moi ca
euh / feuilletton de l'été (...) [blanB1]

En effet la séquence *ça euh / feuilletton* est reconnue puisque le *euh* et la pause brève sont ignorés par les grammaires. Or, on dit que *on* ne peut avoir un chemin qui relie *ça* à *<N>*. le chemin est donc supprimé, alors qu'il est le seul. Cela a pour conséquence que la phrase est rejetée par la grammaire. Si l'on pouvait formaliser qu'on ne souhaite traiter que les cas d'ambiguïté entre nom et verbe, on ne reconnaîtrait pas cette séquence et la phrase ne serait pas rejetée.

Si l'exemple ci-dessus peut paraître plus anecdotique, le formalisme d'ELAG ainsi modifié permettrait de lever de nombreuses ambiguïtés, notamment entre les homographies de participes passés et de nom, dans le contexte droit des verbes *avoir* et *être* conjugués (*est arrivée*, *avait conduit*, *ont fait*, *étaient mort*, *est pensionné*, etc.).

Par ailleurs, on pourrait imaginer, en fin de processus de déambiguïsation avec des grammaires linguistiques (c'est-à-dire construites sur la base de descriptions linguistiques contextuelles fines), de lever les

ambiguïtés restantes avec des heuristiques du type de celles proposées par Jacques Vergne lors de l'évaluation GRACE⁴⁷. En cas d'ambiguïtés spécifiques entre deux catégories grammaticales, l'auteur préconise l'une des catégories dans la mesure où cette hypothèse est beaucoup plus souvent actualisée en corpus. Les heuristiques sont classées **par ordre de préférence** :

- en cas d'ambiguïté Nom/Verbe, choisir Nom ;
- en cas d'ambiguïté Adjectif/Verbe, choisir Adjectif ;
- en cas d'ambiguïté Adverbe/Nom, choisir Adverbe ;
- en cas d'ambiguïté Nom/Adjectif, choisir Nom ;
- en cas d'ambiguïté Déterminant/Pronom, choisir Déterminant ;
- autrement choisir au hasard une étiquette parmi celles proposées.

Évidemment, avec de telles heuristiques, on s'éloigne des règles contextuelles purement linguistiques pour entrer dans la méthode de l'étiquetage probabiliste.

20 ÉVALUATION : RÉSULTATS PRÉLIMINAIRES

Nous l'avons dit, la plupart des auteurs relèvent la non-pertinence de la comparaison des résultats de l'étiquetage faits avec des jeux d'étiquettes différents. Néanmoins, on ne peut se passer de donner une évaluation, même préliminaire, afin de situer la pertinence de notre approche.

En général, les systèmes affichent des taux de réussite aux alentours de 95 %, voire plus. Dans le calcul de l'ambiguïté que nous avons donné ci-dessus (§ 4), avec la phrase d'Alain Guillet *la porte du car se ferme automatiquement*, nous donnions le degré d'ambiguïté de chaque forme, afin de calculer le nombre d'analyses possibles pour la phrase. Comme le dit Gross (2001 : 30) :

Ce nombre [calcul de l'ambiguïté] s'obtient en effectuant le produit des degrés d'ambiguïté de chaque mot. Pour des phrases longues, la croissance du degré est exponentielle, on parle d'explosion combinatoire. Toutefois, cette manière de compter représente mal les phénomènes en jeu. En effet, les ambiguïtés sont souvent locales, c'est-à-dire limitées à une partie de la phrase, voire à des mots simples dans le cas des

⁴⁷ <http://www.limsi.fr/TLP/grace/www/res.labels.html>. Nous n'avons pas trouvé dans les textes de Vergne que nous avons consultés la formulation de cette heuristique.

homographies, elles n'interfèrent alors aucunement entre elles.

Ainsi, les résultats de la levée d'ambiguïté se calculent rarement de cette façon. Il existe plusieurs manières de présenter les résultats, qui correspondent à des mesures quantitatives différentes : le *rappel* (*recall*) et la *précision* (*precision*).

Le *rappel* correspond à la proportion des étiquettes correctes proposées par le système par rapport au nombre d'étiquettes correctes qu'il aurait fallu donner de manière idéale. La différence entre le nombre de bonnes étiquettes total et le nombre d'étiquettes correctes proposées est appelé *silence* ($\text{silence} = 1 - \text{rappel}$).

La *précision* mesure le nombre d'étiquettes correctes produites par le système sur le nombre d'analyses produites par le système (que celles-ci soient correctes ou incorrectes). Pour le dire autrement, la précision mesure le nombre de bonnes étiquettes sur le nombre d'étiquettes fournies. La précision permet de calculer le *bruit* produit par le système ($\text{bruit} = 1 - \text{précision}$).

Un rappel de 100 % signifie que toutes les étiquettes pertinentes sont fournies (mais il peut y en avoir d'autres). Une précision de 100 % signifie que toutes les étiquettes fournies sont correctes.

Dans les évaluations de l'action GRACE, une autre mesure de performance a été développée : la *décision*. « La décision mesure la capacité à fournir, dans le formalisme de référence, une étiquette totalement désambiguïsée (des réponses partiellement ambiguës sont admises). »⁴⁸

Nous allons appliquer ces trois mesures d'évaluation sur nos données.

20.1 Nos grammaires

Nous avons construit, à l'heure actuelle, 83 grammaires de levée d'ambiguïtés. Ce chiffre ne veut pas dire grand-chose. En effet, une seule grammaire peut regrouper plusieurs règles qui pourraient être représentées dans des grammaires séparées, et inversement, on pourrait imaginer de regrouper deux grammaires en une seule, soit parce qu'elles traitent des phénomènes proches, soit par simple commodité. On pourrait donc annoncer un chiffre de 300 grammaires dont le nombre de règles serait équivalent à celui présent dans nos 83 grammaires, et donc, sans que davantage d'ambiguïtés ne soient levées.

⁴⁸ <http://www.limsi.fr/RS99FF/CHM99FF/TLP99FF/tlp10/>.

Nos grammaires traitent des contextes réduits, dépassant rarement les 5 mots. Nous l'avons dit, rien n'empêche d'étendre ce contexte si cela s'avère nécessaire. Nous nous sommes concentrée principalement sur les pronoms préverbaux, les accords dans les groupes nominaux et les suites initiées par des prépositions. Nous avons vu que lever l'ambiguïté sur ces mots grammaticaux permet de lever celle sur les mots lexicaux dans le contexte proche (§ 5). Nous avons aussi levé l'ambiguïté de certains mots lexicaux fréquents dans le corpus.

Les grammaires ont été testées sur différentes parties de notre corpus, afin de faire varier le « corpus d'entraînement » et de vérifier la validité des règles. L'objectif est toujours d'augmenter la précision, sans diminuer le rappel : on lève donc davantage d'ambiguïtés, en essayant de ne jamais supprimer la bonne analyse.

Pour présenter les résultats du système dans son état actuel (avec le prétraitement tel que nous l'avons détaillé au chapitre 4 et en utilisant les ressources lexicales décrites au chapitre 5), nous l'avons testé sur un texte, chaBR1r. Nous avons choisi ce texte pour deux raisons : 1) il n'a pas fait partie des données sur lesquelles nous avons testé nos grammaires ; c'est une entrevue semi-dirigée – comme la majorité des textes qui composent notre corpus –, mais dans laquelle le locuteur se livre à de longs monologues.

Le texte comprend 9598 mots graphiques et il est divisé, lors de l'étape de prétraitement, en 401 unités de traitement.

20.2 Les unités de traitement rejetées

Parmi celles-ci, onze UT sont rejetées (cf. § 17). Les causes de rejet sont de cinq ordres :

- erreur de marquage dans le prétraitement : 2 ;
- faute d'orthographe : 3 ;
- mot composé non répertorié : 1 ;
- énoncé inachevé : 1 ;
- « problème » d'accord en genre ou en nombre dans l'énoncé : 4.

Nous n'avons pas identifié le type d'erreur lié au prétraitement dans les autres phases de test que nous avons faites. Il se produit à cause de balises qui se chevauchent. On a en effet une répétition balisée dans une multitranscription : la balise ouvrante de la répétition apparaît avant la balise fermante de la multitranscription, ce qui gêne le processus d'étiquetage des répétitions et des séquences de multitranscription dans le prétraitement fait avec UNITEX :

chaBR1 (...) et nous étions dans les premiers jours
 <m1> dans </m1> <m2> <r> d'ent/ </m2> </r>
 d'entrée à l'école technique (...) [chaBR1r]

Pour résoudre ce problème, nous proposons de ne pas baliser les répétitions dans les multitranscriptions qui ne sont pas prises en considération par l'étiquetage. Ici, il s'agit en fait d'une seule erreur de marquage, mais qui a pour conséquence de générer deux UT qui sont toutes deux rejetées.

Trois rejets concernent des fautes d'orthographe. Ici, on constate que c'est trois fois le même type d'erreur qui se produit, d'ailleurs un peu surprenant : le pronom sujet pluriel est orthographié *il*, alors que le verbe qui suit est bien au pluriel. Notre grammaire du pronom au singulier spécifie que le verbe qui suit doit être au pluriel, ce qui a pour effet de rejeter les trois UT concernées par cette erreur orthographique.

Nous avons un seul rejet à cause d'un inachèvement. Il s'agit de l'UT suivante, rejetée à cause de la suite *des arrêtez*. Nous avons en effet une grammaire qui spécifie le contexte droit de *des*, et un verbe conjugué ne peut s'y trouver, ce qui est le cas ici :

chaBR1 (...) (silence) pf j'ai connu j'ai connu tellement
des arrêtez un petit peu que je pense à ça
 (interruption de l'enregistrement) (...) [chaBR1r]

Ce seul rejet suite aux inachèvements est sans doute dû au fait que ceux-ci sont peu nombreux dans l'entrevue que nous avons choisie⁴⁹.

Nous avons un rejet lié aux ressources lexicales. Il s'agit de la séquence *au revoir*, qui n'est pas reconnue comme un mot composé. Or, une grammaire interdit un verbe à l'infinitif dans le contexte droit de *au*, ce qui a pour effet de rejeter l'UT.

Nous avons enfin quatre rejets liés à un problème d'accord entre le déterminant et le nom qui le suit. Les quatre séquences concernées sont les suivantes :

chaXX2 on fait une enquête // -| un étude pour les
 une étude un mémoire

chaBR1 (...) tout le monde était hors des voitures il y
 avait un émeute sur le sur la route là (...)

⁴⁹ Nous ne l'avons pas choisie pour cette raison, mais nous nous en sommes rendu compte lors de la validation des résultats de la désambiguïsation. On aurait sans doute eu des rejets plus nombreux dans d'autres textes comme *iljDV1r* ou *ilePA2r*, par exemple, où nous avons pu constater que les locutrices faisaient de nombreuses ruptures syntaxiques. C'était le cas de *blaNb1l*, qui nous a servi de corpus d'entraînement.

chaBR1 (...) c'était un type / qui conduisait du charbon avec sa cheval et sa charrette comme dit ma ma chef là (...)

chaBR1 (...) celui qui n'avait pas les possibilités de ramener un défunt / de de d'un hôpitaux ou l'autre (...)

Les trois premiers cas concernent une inadéquation du genre du déterminant et du genre du substantif, tandis que dans le quatrième cas, c'est une inadéquation de nombre, avec un déterminant au singulier et un nom au pluriel. Ces quatre rejets sont produits par la même grammaire, qui est très efficace pour toutes les autres occurrences du corpus.

Le rejet d'unités de traitement a des implications sur le mode d'application des grammaires que nous avons choisi. En effet, nous avons dit que les grammaires ELAG pouvaient s'appliquer sur l'automate du texte selon deux modes différents : 1) toutes ensemble, en étant compilées dans une seule règle, ou 2) de manière séquentielle, l'une après l'autre. Dans ce cas, après l'application d'une grammaire, on sauve l'automate du texte partiellement désambiguïsé ; c'est sur cet automate qu'est appliquée la grammaire suivante.

L'intérêt d'une application en cascade est évident pour notre corpus. En effet, quand une UT est rejetée, aucune des règles de la grammaire ne s'applique sur elle, et l'automate reste le même que l'automate de départ (mise à part l'indication REJECTED dans la première boîte de l'automate).

Si l'on applique toutes les grammaires en même temps – tout se passe comme si elles ne formaient plus qu'une seule grande règle – et qu'une UT est rejetée, aucune des différentes règles regroupées ne s'applique. On a donc des séquences identiques qui sont désambiguïsées dans certaines UT et pas dans d'autres, parce que celles-ci sont rejetées à cause d'une autre grammaire.

Dans le cas de l'UT rejetée à cause de la séquence *des arrêtez*, l'automate du texte ne subit aucune désambiguïsation, alors que des règles présentes dans d'autres grammaires pourraient s'appliquer. Par exemple, le verbe *pense* est 5 fois ambigu, et seule l'ambiguïté entre subjonctif présent et indicatif présent resterait si la grammaire du *je* pouvait s'appliquer, ce qui n'est pas le cas quand toutes les grammaires sont compilées en même temps.

Afin que les autres règles décrites dans les autres grammaires puissent produire leurs effets de désambiguïsation, on préférera donc appliquer les grammaires les unes à la suite des autres, en sauvant après chaque application l'automate partiellement désambiguïsé.

20.3 Résultats de l'étiquetage⁵⁰

20.3.1 Six catégories d'étiquettes

Le tableau ci-dessous présente les résultats globaux, distillés en six catégories :

	résultats	pourcentage
1	un lemme, une catégorie, une info flexionnelle	72,44 %
2	un lemme, une catégorie, plusieurs info flexionnelles	4,50 %
3	un lemme, plusieurs catégories	12 %
4	plusieurs lemmes, une catégorie	2,31 %
5	le reste	8,48 %
6	pas d'analyse	0,27 %

Tableau 42. Résultats de la désambiguïsation selon le type d'étiquettes

Parmi les six catégories de ce tableau, une seule correspond à une étiquette totalement désambiguïsée (catégorie 1), avec un lemme, une catégorie et une information flexionnelle. Cette catégorie représente 72,44 % des formes du corpus. Pour reprendre la mesure d'évaluation de la décision, développée dans GRACE, on peut dire que notre étiquetage présente une décision dans 72,44 % des cas⁵¹.

Cela signifie que 37,56 % des formes ne sont pas totalement désambiguïsées, puisqu'elles reçoivent plusieurs étiquettes (catégories 2 à 5) ou pas d'étiquettes du tout (cat. 6). Ces derniers cas sont relativement rares : ils ne représentent que 0,27 % des occurrences. Les formes concernées sont des formes non répertoriées dans les dictionnaires, qui appartiennent au dialecte (le wallon) mais sont utilisées telles qu'elles dans un énoncé en français.

⁵⁰ Valli et Véronis (1999) affirment que la plupart des évaluations « adoptent une position généreuse » dans la présentation des résultats. Ce n'est pas la position que nous adoptons ici.

⁵¹ Cette mesure de décision se base sur la définition que nous avons donnée d'une étiquette non ambiguë (un lemme, une catégorie grammaticale et une information de flexion) : on pourrait augmenter ce score, en considérant que les occurrences des catégories 2 et 4 sont satisfaisantes, comme c'est le cas dans de nombreux systèmes qui ne fournissent ni lemme ni informations flexionnelles dans l'étiquette.

Parmi les cas pour lesquels subsistent plusieurs étiquettes, on a plusieurs cas de figure, qui correspondent aux cas énoncés au § 4.

Les cas de type 3 sont les plus nombreux (12 %), pour lesquels on a un lemme de même forme qui correspond à plusieurs analyses catégorielles : c'est le cas des pronoms ambigus avec les déterminants (*le, la, les, leur*, etc.), mais aussi des ambiguïtés entre nom et adjectif (*mortuaire, civil, célibataire, parents*, etc.) ou encore des formes *de, tout, que*, etc.

Viennent ensuite les cas où l'ambiguïté subsiste entre des analyses en lemme et en catégorie différents, avec 8,48 % des formes : c'est le cas des ambiguïtés entre nom et verbe (*porte, diner, danses*, etc.).

On a aussi 4,5 % des occurrences où l'on a levé l'ambiguïté quant au lemme et à la catégorie grammaticale, mais où subsistent plusieurs analyses de flexion (cat. 2) : ce cas de figure est fréquent pour la conjugaison des verbes (ambiguïté indicatif et subjonctif présent, ambiguïté première personne et troisième personne du singulier, ambiguïté présent de l'indicatif et impératif des premières et deuxième personnes du pluriel, etc.) ou le genre et le nombre des adjectifs épiciques (*vieux, facile*, etc.), mais aussi de certains noms (*l'interview, les enfants*, etc.).

20.3.2 Le rappel

Dans notre étiquetage, le *rappel* avoisine les 100 %⁵². En effet, hormis les 0,27 % de formes qui ne reçoivent pas d'analyse, on a quelques cas où la bonne analyse n'est pas proposée par le système :

- soit que cette analyse n'est pas présente dans les ressources lexicales : c'est le cas des composés *au revoir, arrières-grands-parents, n'importe, pour cent* et *tout à fait*, que nous estimons devoir être analysés comme des composés et qui ne sont pas répertoriés comme tels dans les dictionnaires (8 cas) ;
- soit que le codage dans le dictionnaire est mauvais : c'est le cas de *c'*, qui ne reçoit que le code singulier, alors que l'analyse en un pluriel devrait être proposée, dans *c'étaient* par exemple. Ceci est une erreur de codage de notre part, facile à corriger (11 cas) ;
- soit que la grammaire génère une mauvaise analyse : dans notre corpus, cela se produit à cinq reprises. Voici les occurrences concernées⁵³ :

⁵² Cette affirmation ne vaut évidemment que si l'on est d'accord avec les choix linguistiques que nous avons posés.

⁵³ Ces cinq extraits sont attribués à chaBR1 et ne constituent pas un tour de parole complet.

- (1) c'est terminé hein m'fève
 (2) oh oui hein m'feye
 (3) pour une telle personne qu'habite dans une telle rue
 (4) il y en a eu les quatre-vingts pour cent qui sont morts et qui n'ont pas atteint les cinquante-cinq ans / qu'ont eu leur pension à quarante-cinq cinquante-cinq ans
 (5) et on met le le chose en marche / c'est un un / des pressions à air

Dans les deux premiers cas, le *m'* est analysé comme le pronom. C'est d'ailleurs la seule analyse possible selon les données du dictionnaire. Or ici, il s'agit du déterminant élide, devant une forme nominale dialectale. Ces formes devraient être balisées dans le texte afin de ne pas être analysées. Les erreurs produites dans les exemples 3 et 4 sont liées à un non-respect des conventions de transcription. En effet, le pronom *qui* élide dans la prononciation doit être transcrit sous sa forme pleine. Or, ce n'est pas le cas ici. Les données du dictionnaire ne mentionnent pas *qu'* avec le lemme *qui* mais uniquement *que*. On fournit donc ici une analyse erronée en termes de lemme. Le dernier cas cité pourrait figurer parmi les exemples de codage manquant dans les ressources lexicales. En effet, le locuteur utilise ici *chose* comme il utiliserait *machin* ou *truc*, au masculin. Or, l'information flexionnelle présente pour *chose* dans le dictionnaire indique le féminin. À la différence des cas que nous citons ci-dessus, l'UT dans laquelle apparaît *le chose* n'est pas rejetée car on ne supprime pas le seul chemin de l'automate ; une désambiguïsation partielle est proposée. Voici l'automate de départ de *le chose*, et l'automate partiellement désambiguïsé :

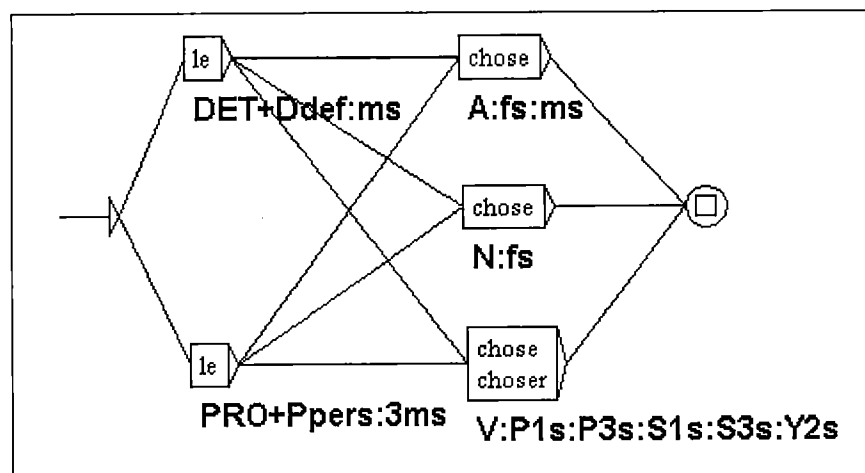


Figure 56. Automate de le chose

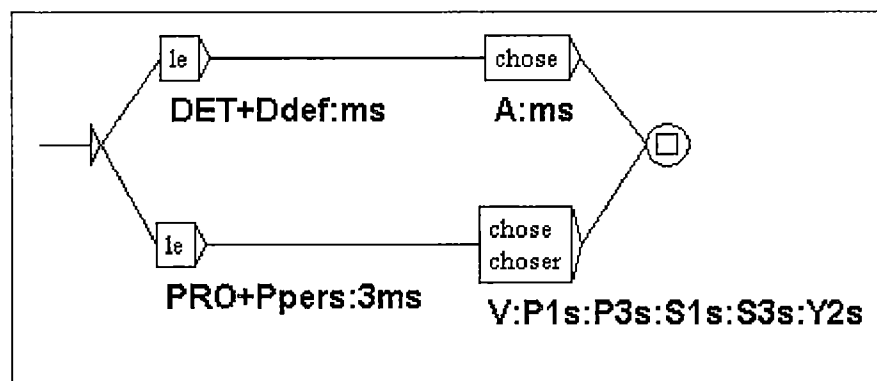


Figure 57. Automate partiellement désambiguïsé de le chose

Si l'UT n'est pas rejetée, à la différence des cas que nous avons analysés ci-dessus, c'est que l'on a aussi deux autres analyses pour *chose* : A et V. Dans l'automate partiellement désambiguïsé, on voit que le déterminant au masculin contraint l'adjectif d'être au masculin (l'hypothèse *fs* est supprimée), et que l'analyse de la forme en un nom a été supprimée, puisque celui-ci était au féminin. Si l'on souhaite donc obtenir les occurrences du lemme *chose* en tant que substantif, notre système produit pour cette occurrence du silence.

20.3.3 La précision

La précision calcule le nombre de bonnes étiquettes fournies par le système par rapport au nombre d'étiquettes total (bonnes ou pas) fournies par le système. Si l'on considère que l'on a une étiquette chaque fois que l'on a un lemme ou une catégorie ou une information flexionnelle non unique, on a dans l'automate de notre texte 17 206 étiquettes, parmi lesquelles 65,16 % présentent l'analyse souhaitée.

Si l'on regroupe maintenant dans une seule étiquette les informations flexionnelles se rapportant à un lemme unique ET à une partie du discours unique, on a alors 14 518 étiquettes dans l'automate, et le taux de précision augmente à 77,23 %.

La précision est fortement diminuée dans notre système à cause de la prise en compte des formes composées, où la couverture est très bonne en ce qui concerne les mots grammaticaux. Nous l'avons déjà mentionné, il n'y a pas actuellement de solution dans UNITEX ni dans ELAG qui permettrait de préférer une forme composée non ambiguë à une suite de mots simples. Ainsi, alors que l'on a *grands-parents* comme composé dans le dictionnaire, à l'heure actuelle, cette analyse du composé subsiste dans un chemin en

parallèle avec l'analyse en mots simples. Cela augmente considérablement le nombre d'étiquettes et diminue de ce fait la précision.

Dans notre système, nous ne regroupons pas des formes particulièrement ambiguës, comme c'est le cas de Winbrill (Lecomte 1998) pour le *que* par exemple, qui reçoit l'étiquette SUB\$. Nous avons fait un essai, en neutralisant l'ambiguïté des formes *de* et *que* dans une étiquette unique : on obtient alors, toujours sur le même texte, un taux de précision qui passe à 79,82 %.

Quoi qu'il en soit, la présentation des résultats en termes de précision ne rend pas tout à fait justice au travail partiel de désambiguïsation qui est fait sur l'automate du texte : en effet, de nombreux chemins sont supprimés, sans qu'on obtienne au final une seule étiquette non ambiguë (par exemple, entre déterminant et verbe). On peut espérer qu'en ajoutant de nouvelles grammaires, on « démêle les témoignages contradictoires » afin de ne conserver que la bonne analyse.

21 PERSPECTIVES

Nous avons défini les unités de traitement comme des unités de travail, sans statut théorique. On constate que nos choix de découpage sont pertinents, car dans les résultats que nous obtenons, c'est rarement eux qui génèrent des erreurs dans l'annotation morphosyntaxique. Le seul problème, auquel nous n'avons pas pu apporter de solution, est que l'on voudrait trouver une fin d'UT à l'endroit des inachèvements syntaxiques.

Néanmoins, les UT ont ceci d'insatisfaisant qu'elles sont de tailles très variables et ne représentent aucune unité de structure syntaxique. Certaines sont très brèves et d'autres très longues, comme c'est le cas dans chaBR1r, où la majorité des pauses sont des pauses brèves, même dans les longs monologues. On a donc peu de découpage. Pour comparaison, le texte blaNB1l, qui contient 6614 mots graphiques, est divisé en 697 UT, principalement parce qu'il contient plus de pauses longues.

Nous aimerions donc procéder à un découpage du texte qui se baserait sur des indices plus sûrs. Nous pensons par exemple aux unités minimales du discours, telles que celles dégagées par Degand et Simon (2005). C'est d'ailleurs selon ce type d'unités que sont maintenant alignées les transcriptions de VALIBEL. Récupérer les données sous cette forme, plutôt que sous le format d'un enchaînement en tours de parole, ne pose donc aucun problème d'ordre technique.

En ce qui concerne les disfluences, nous ne les avons guère abordées dans cette évaluation. C'est qu'elles ne posaient dans notre texte de test aucun problème en ce qui concerne l'annotation morphosyntaxique – ni

fausse analyse, ni non-reconnaissance de mots composés –, si ce n'est le cas de mauvais balisage effectué au prétraitement dans une multitranscription.

Nous avons, dans le texte que nous avons pris en considération, six séquences où une répétition a été reconnue alors qu'elle relève de la syntaxe : dans cinq cas, il s'agit de *vous vous souvenez* et, à une reprise, de *vous vous rendez compte*. Baliser ces séquences n'a pas posé de problème pour l'annotation morphosyntaxique : la bonne étiquette a été donnée à ces formes et pour les formes dans le contexte. Reste maintenant à savoir le format final que nous allons donner au texte étiqueté. En effet, deux solutions sont possibles : 1) soit on annote les disfluences de la même manière que le dernier répété qui, lui, fait partie du texte annoté, 2) soit on conserve les disfluences à l'intérieur des balises actuelles. Cette dernière solution aurait l'avantage de faciliter les recherches sur le corpus annoté, puisque la recherche d'un motif syntaxique se ferait de la même manière que lors de l'application des grammaires de levée d'ambiguïtés : en ignorant les séquences disfluentes. Dans ce cas, il faudrait alors régler le problème des séquences balisées qui doivent apparaître dans l'énoncé, comme les pronoms dans les exemples que nous venons de citer.

Le taux de précision que nous avançons peut paraître faible en regard des résultats annoncés par d'autres systèmes. Nous avons vu qu'il est lié à la finesse de nos données linguistiques, notamment à la prise en compte des mots composés, et à notre volonté de ne pas éliminer de bonnes analyses lors de l'étiquetage. Au regard de cet objectif, on peut considérer que les résultats que nous obtenons ici sont plus qu'encourageants.

En effet, les grammaires que nous avons construites sont loin de recenser toutes les structures syntaxiques en jeu dans le corpus. Elles demanderaient inévitablement à être complétées, voire affinées, sans doute en augmentant le nombre de grammaires lexicalisées. De plus, certaines ambiguïtés sont impossibles à lever de manière automatisée, car le contexte seul ne le permet pas. Ainsi, si l'on veut une annotation non ambiguë (lorsque celle-ci est possible), il sera toujours nécessaire qu'un expert humain valide les résultats en fin de processus.

CONCLUSION

Ils se marièrent, vécurent heureux et eurent beaucoup d'enfants.

Parvenus au terme de ce parcours qui envisage les différentes étapes qui mènent de la transcription de corpus oraux à leur étiquetage morphosyntaxe, il nous paraît important de revenir brièvement sur les implications et les enjeux de la série de choix que nous avons posés. En effet, notre recherche se situe dans un entre-deux qui, a priori, peut apparaître comme gênant, voire insatisfaisant d'un point de vue épistémologique :

D'un côté, notre travail a été orienté par un objectif pratique précis : étiqueter morphosyntaxiquement des corpus oraux retranscrits, ce travail ayant vocation à préparer des exploitations ultérieures de ces corpus. Il ne fait pas de doute que ces exploitations prendront la forme de recherches susceptibles de contribuer à l'enrichissement des théories linguistiques existantes. Dans cette optique, on pourrait concevoir qu'idéalement, la série d'interventions que nous avons décrites pour aboutir à l'étiquetage morphosyntaxique soient « neutres » d'un point de vue théorique. Or, et c'est précisément toute la difficulté, ce travail à vocation pratique repose en réalité sur une série de choix préalables, souvent implicites, mais dont la portée est elle aussi théorique. D'où la nécessité de manifester, tout au long de notre travail et pour chacune des étapes envisagées, les choix que nous avons opérés, et d'indiquer, dans toute la mesure du possible, leurs enjeux et leurs conséquences.

D'autre part, une difficulté supplémentaire s'est immédiatement présentée, qui touche aux contingences pratiques et empiriques imposées par le matériau de base : les choix de transcription posés en amont de notre intervention, et la décision de ne pas retourner aux données sonores pour trancher d'éventuels (et inévitables) problèmes d'interprétation. En somme,

les choix que nous avons dû faire ont tenté de concilier ces contingences matérielles et les enjeux théoriques.

En effet, nous avons été largement tributaire des conventions de transcription établies au centre de recherche VALIBEL : absence de ponctuation des textes, présentation des données en tours de parole, notation subjective des pauses, absence de méta-information dans le texte valant comme une préanalyse sur laquelle se baser, par exemple, pour le découpage du texte ou encore pour le repérage des disfluences et des inachèvements syntaxiques, etc. Le fait que les textes ne fournissent pas d'informations quant à l'allongement phonétique a parfois rendu difficile la comparaison de nos données avec les résultats de recherches antérieures.

Si la plupart des conventions sont explicites – et justifiées théoriquement –, nous avons également été confrontée à des cas pour lesquels la variation est gênante, principalement du point de vue du traitement informatisé : on pense à la notation des éléments paraverbaux, qui s'insèrent soit à la fin d'un constituant, soit au milieu de celui-ci, à la notation d'une même forme tantôt considérée comme un nom propre, tantôt non, ou encore aux indications de phonétique qui, la plupart du temps, sont des ajouts au texte mais qui, parfois, apparaissent en lieu et place du texte transcrit. Cette variation, qui n'a pas de répercussion sur les analyses faites jusqu'à présent dans le cadre des recherches menées au Centre VALIBEL (notamment parce qu'elle n'a guère d'enjeu théorique), est relativement gênante dans le cadre de notre travail, où l'on aimerait travailler sur des objets « uniformes » et stables.

Il nous a donc fallu tenir compte de tous ces éléments pour envisager au mieux la tâche d'annotation morphosyntaxique.

Parmi les ressources disponibles dans la banque de données VALIBEL, nous avons choisi, pour constituer notre corpus, 60 textes transcrits (450 000 mots), issus de 12 sous-corpus différents. Ces textes ont été sélectionnés en fonction de critères sociologiques caractérisant les locuteurs, et en tenant compte également de la qualité sonore des enregistrements. Nous avons, autant que faire se peut, essayé d'équilibrer les différents paramètres, mais on a vu l'impossibilité devant laquelle nous étions. Notre corpus présente donc des disproportions, par exemple en termes de niveau d'études des locuteurs ou encore selon leur répartition géographique. Dans le corpus ainsi constitué, les situations de parole sont peu variées – principalement des entretiens semi-dirigés –, et elles relèvent toutes de l'oral non planifié. Néanmoins, et à notre connaissance cela est unique dans la constitution d'un corpus comme le nôtre, cette absence apparente de variété est compensée par une stricte documentation des données (les données sociologiques sur les locuteurs, mais aussi les liens qui les unissent, le lieu de l'interaction, etc.). De plus, certains des résultats que nous avons pu mettre en évidence mettent à mal les limites strictes que l'on voudrait établir entre les genres : des

entrevues semi-dirigées contiennent de longs monologues qui s'apparentent plus au récit de vie qu'à une situation contraignante pour les locuteurs.

Avant de mettre en place la phase de prétraitement des données, nous avons analysé de manière approfondie, dans notre corpus, certaines marques typiques des énoncés en cours d'élaboration, marques que nous avons appelées *disfluences*. Celles-ci sont reconnues d'emblée comme la pierre d'achoppement des traitements sur l'oral (non seulement pour l'étiquetage, mais aussi pour la reconnaissance de la parole ou le dialogue homme-machine). Les marques que nous avons prises en considération sont le *eah*, les répétitions, les autocorrections immédiates et les amorces de morphèmes.

Notre volonté était d'éclairer des phénomènes qu'il est possible repérer formellement. En effet, le but de l'analyse que nous avons faite est de pouvoir intégrer ces marques dans le processus de levée d'ambiguïtés. Dans le cadre de ce travail, étudier un phénomène qu'il nous sera impossible d'intégrer dans un repérage ou une analyse semi-automatisés n'aurait guère eu de sens, même s'il n'eût évidemment pas manqué d'intérêt.

Les *eah* et les amorces sont des formes graphiques non ambiguës, aisément repérables. Il n'en est pas de même pour les répétitions, que nous avons circonscrites à une identité formelle et à une stricte contiguïté – mise à part une liste réduite d'insertions possibles. Cette définition de la répétition est réductrice par rapport à la réalité du phénomène, et écarte de l'analyse des occurrences très proches mais qui ne répondent pas à nos critères d'identité formelle et de contiguïté. Inversement, nous regroupons sous la même étiquette des répétitions qui relèvent effectivement de la disfluence et des répétitions tout à fait standard, liées soit aux exigences de la grammaire, soit à un effet de style.

Pour ce qui est des réalisations de répétitions très proches que nous avons écartées, une étude approfondie permettrait sans doute de mieux les caractériser, afin de tenter de les repérer formellement. C'est d'ailleurs ce que nous avons fait en partie, en considérant les autocorrections immédiates, qui sont une variante de la répétition telle que nous l'avons définie.

Par ailleurs, englober des répétitions fonctionnellement différentes sous le seul label *répétition* a l'avantage de ne pas trancher en fonction d'un critère qui serait celui de l'intention du locuteur. Nous avons donc soumis toutes les répétitions qui suivent le même schéma formel à une analyse identique.

Pour les marques que nous avons étudiées, nous avons mis en évidence, comme d'autres auteurs avant nous, certaines régularités indéniables. Cela étant, on a échoué à trouver des indices qui seraient la marque d'une rupture syntaxique. En effet, si les disfluences constituent un empiètement sur l'axe syntagmatique, celui-ci se poursuit ensuite normalement, avec une reprise du déroulement linéaire de l'énoncé. Mais dans certains cas, on n'assiste pas à

une reprise et le constituant entamé est inachevé. On a tenté de trouver des traces de ces inachèvement dans les marques de disfluence ou dans le cumul de ces marques, mais on n'a pu mettre en évidence aucun résultat probant. De plus, les inachèvements sont bien plus nombreux sans aucune trace de disfluence à l'emplacement de la rupture.

L'étude des disfluences nous a permis de poser des choix quant à leur intégration dans le processus d'étiquetage. Nous avons décidé de les gérer lors du prétraitement des données, plutôt que de les intégrer dans les grammaires de levée d'ambiguïtés. Les disfluences sont donc balisées et étiquetées ; quand on a une suite de disfluences, c'est la séquence maximale qui est reconnue. Traiter les disfluences dans une phase de prétraitement a néanmoins des répercussions sur la phase d'étiquetage proprement dite puisque nous avons ajouté les étiquettes insérées lors de cette étape au jeu d'étiquettes du système de désambiguïsation, dont l'étiquette <IGN> qui a la particularité d'être ignorée dans le processus de levée d'ambiguïtés.

Tenant compte des paramètres liés aux locuteurs, nous avons pu mettre en avant la difficulté de relier ceux-ci à la fréquence des marques de disfluences. Celles-ci sembleraient plutôt relever du style du locuteur que devoir être attribuées à des facteurs tels que le degré de scolarité ou encore l'aspect « professionnel de la parole ». Des études plus approfondies, qui analyseraient le même locuteur dans d'autres situations de communication, permettraient de confirmer ou d'infirmer cette hypothèse. Par ailleurs, certains locuteurs, présents dans plusieurs conversations différentes, présentaient des taux de marques disfluentes très variables, sans que l'on puisse clairement en établir la cause.

Le prétraitement, s'il concerne ces disfluences, concerne aussi de nombreuses autres marques présentes dans les textes. En effet, si elles sont bien dans un format compatible avec le logiciel UNITEX que nous utilisons, les transcriptions de l'oral comportent des marques qu'il faut prendre en considération dans la perspective de l'annotation morphosyntaxique : balises de chevauchement de parole, codes des locuteurs, marques des pauses, indications paraverbales, etc. Il a également fallu harmoniser les transcriptions, faites à diverses époques selon des conventions parfois un peu différentes.

Nous avons donc modifié le texte afin de lui donner une « forme » acceptable par notre système d'annotation. Ces modifications ont été pensées dans un constant aller-retour entre données de départ et résultats souhaités.

Les différentes marques dont il est question ont ainsi reçu une étiquette spécifique, intégrée également parmi les étiquettes prises en compte par le système de levée d'ambiguïtés.

Lors du prétraitement, nous scindons le texte en unités plus petites, qui ne correspondent pas à une unité théorique mais simplement à une unité de traitement adéquate.

Si l'étape du prétraitement vise à préparer les données en vue de leur annotation, il ne s'agit pas à proprement parler d'une normalisation. Par exemple, on ne vise pas à ponctuer le texte, à compléter les amorces, à résoudre les inachèvements ou à effacer les segments corrigés par le locuteur. Notre prétraitement consiste en des ajouts au texte de départ (des étiquettes ou des marques de segmentation), sans perte, à aucun moment du traitement, des informations présentes dans le texte brut de départ.

Nous avons détaillé cette étape de prétraitement car elle est la plupart du temps présentée comme une phase purement pratique du processus d'étiquetage, consistant simplement à résoudre des problèmes techniques. Or, il s'agit d'une étape importante, non seulement parce que, dans le cadre d'un corpus comme le nôtre, elle est relativement complexe et couteuse en temps, mais aussi parce que les choix qui y sont opérés ont des répercussions sur les analyses qui seront faites ultérieurement sur les données.

Dans l'évaluation que nous avons faite des résultats partiels obtenus après l'application de nos grammaires de désambiguïsation, s'il reste un grand nombre de choses à améliorer, on peut affirmer que l'étape de prétraitement que nous avons choisie, tant en ce qui concerne les disfluences que les autres marques, s'avère pertinente.

Un autre sujet de réflexion, dans l'élaboration de ce travail, a concerné le jeu d'étiquettes à prendre en considération pour l'annotation. La particularité de nos données a conditionné le choix du jeu d'étiquettes, lui-même lié aux ressources lexicales que nous utilisons. Nous avons montré comment, en fonction des étiquettes d'une part et, d'autre part, selon les informations associées au lexique, le concept d'ambiguïté était variable. Nous avons détaillé les choix que nous avons fait pour le codage des mots grammaticaux. Ici, nous avons dû articuler deux programmes de recherche différents l'un à l'autre : celui du LADL et celui du centre de recherche VALIBEL. En effet, les recherches menées dans le cadre du lexique-grammaire ont permis une catégorisation en partie du discours qui se distingue des typologies classiques. Néanmoins, nous avons préféré des étiquettes plus consensuelles, qui à bien des égards sont très proches de celles de la grammaire scolaire, afin que les données annotées puissent être utilisées par des chercheurs non au fait de la théorie du lexique-grammaire. Ce choix est critiquable à plus d'un égard, mais à l'heure actuelle, en l'absence de consensus, il nous paraît le plus approprié.

Étant donné la double spécificité de notre corpus de français *parlé* en *Belgique*, nous avons également enrichi les ressources lexicales par une double intervention : d'une part par la prise en compte de certaines formes

de l'oral, telles que les interjections et les onomatopées ; d'autre part, en créant un dictionnaire électronique du français en Belgique, le DELFRABEL, qui comprend près de 5500 entrées. L'ajout de ces ressources assure, sur notre corpus, une couverture de pratiquement 100 %.

En ce qui concerne les ressources lexicales, le travail encore à effectuer, d'envergure, consiste à compléter les ressources pour les formes composées, principalement les noms.

Pour ce qui est de l'étiquetage proprement dit, nous avons choisi de développer le système ELAG. Au vu des résultats que nous obtenons, ce système semble bien adapté à des données telles que les nôtres. Son intégration dans le système UNITEX offre des avantages indéniables, au nombre de trois : d'abord, la majeure partie de la phase de prétraitement est effectuée sous UNITEX ; ensuite, les ressources lexicales distribuées avec UNITEX sont très complètes ; enfin, nous avons pu ajouter nos propres ressources lexicales, ce qui était indispensable pour pouvoir analyser correctement nos données.

De plus, le système ELAG nous permet de définir notre jeu d'étiquettes, en déterminant les catégories, sous-catégories et traits spécifiques à retenir pour l'annotation. Cette manière de réduire les ambiguïtés non désirées se fait grâce à la définition du jeu d'étiquettes dans un fichier spécial (*tagset.def*).

Par ailleurs, l'avantage principal d'ELAG, conçu par d'autres comme un inconvénient majeur, est que le résultat de la levée d'ambiguïtés est présenté sur un automate du texte partiellement désambiguïté et non dans un texte linéaire où chaque mot recevrait une étiquette unique. Si nous considérons que cette solution est un plus, c'est que notre objectif n'est pas d'annoter automatiquement tout le corpus (au risque que de nombreuses étiquettes soient fausses), mais bien de faire une désambiguïsation partielle (mais correcte) qui sera validée et complétée par un expert humain. De plus, nous l'avons vu dans notre corpus, certaines ambiguïtés sont impossibles à lever et vouloir attribuer une étiquette unique est un non-sens.

À l'heure actuelle, les grammaires que nous avons construites s'appliquent de manière très satisfaisante. Nous avons construit d'autres grammaires qui sont en cours de validation ; de nombreuses grammaires sont encore à écrire, qui traiteront de phénomènes que nous n'avons pas encore pris en considération.

S'il reste inévitablement du travail à faire (ajouter des grammaires, valider l'annotation), le processus qui part d'une transcription VALIBEL pour aboutir à une annotation morphosyntaxique partielle de celui-ci est explicite et opérationnel.

Nous avons dit dans notre introduction que l'annotation serait, à n'en pas douter, une plus-value pour nos données. Les perspectives de recherche

ouvertes sont nombreuses. L'une qui vient immédiatement à l'esprit est de permettre des travaux comme ceux menés par Douglas Biber sur les types de texte. Pour cela, il faudrait varier les données, par la prise en compte d'autres types d'interactions langagières.

Nous avons mis en évidence, tout au long de ce travail, un fait indéniable : la constitution et l'enrichissement de corpus est un travail couteux, et il l'est encore plus lorsqu'il s'agit de données orales.

Il convient donc de donner aux données la plus grande diffusion possible, afin que la communauté scientifique puisse les exploiter à la hauteur du travail qui a été investi.

BIBLIOGRAPHIE

* Les articles dont la pagination n'est pas donnée ont été consultés dans une version électronique téléchargeable non paginée.

- ABEILLÉ Anne, CLEMENT Lionel (1999). « A tagged reference Corpus for French », *Actes de LINC-EACL* (June12, Bergen).
- ABNEY Steve (1991). « Parsing by chunks », R. Berwick, S. Abney et C. Tenny (Éds.), *Principle-based parsing : Computation and Psycholinguistics*, Boston, Kluwer Academic Publishers, pp. 257-278.
- ABNEY Steve (1996). « Part-of-speech tagging and partial parsing », S. Young et G. Bloothoof (Éds.), *Corpus-based Methods in Language and Speech*, Dordrecht, Kluwer Academic Publishers, pp. 118-136.
- ADDA Gilles, LECOMTE Josette, MARIANI Joseph, PAROUBEK Patrick, RAJMAN Martin (1997). « Les procédures de mesure automatique de l'action GRACE pour l'évaluation des assignateurs de parties du discours pour le français », *Premières Journées scientifiques et techniques du Réseau francophone de l'Ingénierie de la Langue de l'Aupelf-Uref*, Avignon.
- ADDA Gilles, LECOMTE Josette, MARIANI Joseph, PAROUBEK Patrick, RAJMAN Martin (1998). « The GRACE French Part-of-Speech Tagging Evaluation Task », *Proceedings of the First International Conference on Language Resources and Evaluation (LREC)* (May 28-30, Granada).
- ADDA Gilles, MARIANI Joseph, PAROUBEK Patrick, RAJMAN Martin, LECOMTE Josette (1999a). « L'action GRACE d'évaluation de l'assignation des parties du discours pour le français » *Langues* 2 (1), pp. 119-129.
- ADDA Gilles, MARIANI Joseph, PAROUBEK Patrick, RAJMAN Martin, LECOMTE Josette (1999b). « Métrique et premiers résultats de l'évaluation GRACE des étiqueteurs morphosyntaxiques pour le français », *Actes de TALN 1999* (12-17 juillet 1999, Cargèse).
- ADDA-DECKER Martine, HABERT Benoît, BARRAS Claude, ADDA Gilles, BOULA DE MAREÛIL Philippe, PAROUBEK Patrick (2003). « A disfluency study for cleaning spontaneous speech automatic transcripts and improving speech language models », *ISCA Tutorial and Research Workshop on Disfluency in Spontaneous Speech (DiSS'03)*, Gothenburg, University of Gothenburg, pp. 67-70.
- ADDA-DECKER Martine, HABERT Benoît, BARRAS Claude, ADDA Gilles, BOULA DE MAREÛIL Philippe, PAROUBEK Patrick (2004). « Une étude

des disfluences pour la transcription automatique de la parole spontanée et l'amélioration des modèles de langage », *Actes des 25^{es} Journées d'Étude sur la Parole (JEP'2004)* (19-22 avril, Fès).

ANTOINE Jean-Yves, LETELLIER-ZARSHENAS Sabine, NICOLAS Pascale, SCHADLE Igor, CAELEN Jean (2002). « Corpus OTG et ECOLE_MASSY : vers la constitution d'une collection de corpus francophones de dialogue oral diffusés librement », *Actes de TALN 2002* (24-27 juin, Nancy), pp. 319-324.

ANTOINE Jean-Yves, GOULIAN Jérôme, VILLANEAU Jeanne (2003). « Quand le TAL robuste s'attaque au langage parlé : analyse incrémentale pour la compréhension de la parole spontanée », *Actes de TALN 2003* (11-14 juin, Batz-sur-Mer), pp. 25-34.

ANTONI Marie-Hélène, DEMONET Marie-Luce (2000). « Adaptation d'un lemmatiseur au corpus rabelaisien : naissance d'Humanistica », *Actes des JADT 2000, 5^{es} Journées Internationales d'Analyse des Données Textuelles* (9-11 mars, Lausanne), pp. 493-500.

ARRIVÉ Michel (1968). « Les formes des déterminants et des substantifs », *La Grammaire du français parlé, Le Français dans le Monde* 47, juin 1968, pp. 34-37.

ARRIVÉ Michel, GADET Françoise, GALMICHE Michel (1986). *La Grammaire d'aujourd'hui : guide alphabétique de linguistique française*, Paris, Flammarion.

ATKINSON Maxwell J., HERITAGE John (1999). « Jefferson's transcript notation », A. Jaworski et N Coupland (Éds.), *The Discourse Reader*, London, Routledge, pp. 158-166.

ATWELL Eric (1983). « Constituent-Likelihood Grammar », *Newsletter of the International Computer Archive of Modern English (ICAME NEWS)* 7, Norwegian Computing Centre for the Humanities, Bergen University, pp. 34-67.

BAAYEN Harald, SPROAT Richard (1996). « Estimating lexical priors for low-frequency morphologically ambiguous forms », *Computational Linguistics* 22 (2), pp. 155-166.

BAL Willy, DOPPAGNE Albert, GOOSSE André, HANSE Joseph, LENOBLE-PINSON Michèle, POHL Jacques, WARNANT Léon (1994). *Inventaire des particularités lexicales du français en Belgique*, Louvain-la-Neuve, Duculot.

BALLESTER Almudena, SANTAMARÍA Carmen, MARCOS-MARIN Francisco A. (1993). « Transcription conventions used for the Corpus of Spoken Contemporary Spanish », *Literary and Linguistic Computing* 8 (4), pp. 283-92.

-
- BALLY Charles (1935). *Le Langage et la Vie*, Zurich, Max Niehans (2^e éd.).
- BALVET Antonio (2001). « Grammaires locales et lexique-grammaire pour le filtrage d'information. Vers une (ré)utilisabilité des ressources linguistiques pour la recherche d'information », *Actes des quatrièmes rencontres Terminologie et Intelligence Artificielle* (3-4 mai 2001, Nancy), pp. 201-211.
- BARBERIS Jeanne-Marie (1992). « Onomatopée, interjection : un défi pour la grammaire », *L'Information grammaticale* 53, pp. 52-57.
- BEALE Andrew David (1985). « Grammatical analysis by computer of the Lancaster-Oslo/Bergen (LOB) corpus of British English texts », *Proceedings of the 23rd annual Meeting on Association for Computational Linguistics*, pp. 293-298.
- BEAR John, DOWDING John, SHRIBERG Elizabeth (1992). « Integrating multiple knowledge sources for detection and correction of repairs in human-computer dialogue », *Actes de ACL 1992*, pp. 56-63.
- BÉGUELIN Marie-José (1999). « De la phrase aux énoncés : bilan et perspective », *TRANEL* 31, pp. 143-153.
- BÉGUELIN Marie-José (Dir.) (2000). *De la phrase aux énoncés : grammaire scolaire et descriptions linguistiques*, Bruxelles, De Boeck & Larcier.
- BÉGUELIN Marie-José (2002). « Clause, période ou autre ? La phrase graphique et la question des niveaux d'analyse », M. Charolles, P. Le Goffic et M.-A. Morel (Éds.), *Y a-t-il une syntaxe au-delà de la phrase ?*, *Verbum* 24 (1-2), pp. 85-107.
- BEL-ENGUUX Gemma, JIMÉNEZ-LÓPEZ Maria Dolores (2006). « Ambiguous turn-taking games in conversations », P. Mertens, C. Fairon, A. Dister et P. Watrin (Éds.), *Verbum ex machina, Actes de TALN 2006, Cahiers du Cental* 2 (1), Presses universitaires de Louvain, Louvain-la-Neuve, pp. 400-406.
- BÉNARD Frédérique (2005). *Normalisation de corpus oraux : des métadonnées à l'annotation des transcriptions*, Université Paris-3, Sorbonne Nouvelle, Mémoire de maitrise.
- BENZITOUN Christophe (2004). « L'annotation syntaxique de corpus oraux constitue-t-elle un problème spécifique ? », *Actes de RÉCITAL* (21 avril 2004, Fès).
- BENZITOUN Christophe, CAMPIONE Estelle, DEULOFEU José, HENRY Sandrine, SABIO Frédéric, TESTON Sandra, VALLI André, VÉRONIS Jean (2004). « L'analyse syntaxique de l'oral : problèmes et méthode », *Journée d'étude de l'ATALA sur l'annotation syntaxique de corpus* (15 mai 2004, Paris).

-
- BENZITOUN Christophe, VERONIS Jean (2005). « Problèmes d'annotation d'un corpus oral dans le cadre de la campagne EASY », *Actes de TALN 2005* (6-10 juin, Dourdan), t. 2, pp. 13-16.
- BERRENDONNER Alain (1993). « La phrase et les articulations du discours », *Des pratiques de l'écrit, Le Français dans le monde*, n° sp., février-mars 1993, pp. 20-26.
- BERRENDONNER Alain (1999). « Histoire d'une transposition didactique ; les "Types de phrase" », *TRANEL* 31, pp. 37-54.
- BERRENDONNER Alain, REICHLER-BÉGUELIN Marie-José (1989). « Décalages : les niveaux de l'analyse linguistique », *Langue française* 81, pp. 99-125.
- BERTHOUD Anne-Claude, MONDADA Lorenza (Éds.) (2000). *Modèles du discours en confrontation*, Berne, Peter Lang.
- BIBER Douglas (1988). *Variation across Speech and Writing*, Cambridge, Cambridge University Press.
- BIBER Douglas (1993a). « Representativeness in Corpus Design », *Literary and Linguistic Computing* 8 (4), pp. 243-257.
- BIBER Douglas (1993b). « Using Register-Diversified Corpora for General Language Studies », *Computational Linguistics* 19 (2), pp. 219-241.
- BIBER Douglas, JOHANSSON Stig, LEECH Geoffrey, CONRAD Susan, FINEGAN Edward (1999). *Longman Grammar of Spoken and Written English*, London, Pearson ESL.
- BILGER Mireille (1982). « Contribution à l'analyse en grilles », *Recherches sur le français parlé* 4, Université de Provence, pp. 195-215.
- BILGER Mireille (2000). « Petite typologie des conventions de transcription de l'oral. Quelques aspects pratiques et théoriques », M. Bilger (Éd.), *Linguistique sur corpus. Études et réflexions, Cahiers de l'Université de Perpignan* 31, pp. 77-92.
- BILGER Mireille, CAMPIONE Estelle (2002). « Propositions pour un étiquetage en "séquences fonctionnelles" », *Recherches sur le français parlé* 17, Université de Provence, pp. 117-136.
- BLACHE Philippe, MORIN Jean-Yves (2003). « Une grille d'évaluation pour les analyseurs syntaxiques », *Atelier sur l'évaluation des analyseurs syntaxiques, Actes de TALN 2003* (11-14 juin, Batz-sur-Mer), pp. 77-86.
- BLACHE Philippe, GUÉNOT Marie-Laure, PORTES Cristel (2005). « Une plateforme d'analyse lexicale pour le développement du Français Fondamental », *Actes de Français Fondamental : 50 ans de travaux et d'enjeux* (8-10 décembre 2005, Lyon).

-
- BLANCHE-BENVENISTE Claire (1985). « La dénomination dans le français parlé : une interprétation pour les “répétitions” et les “hésitations” », *Recherches sur le français parlé* 6, Université de Provence, pp. 109-130.
- BLANCHE-BENVENISTE Claire (1987). « Syntaxe, choix de lexique et lieux de bafouillage », *DRLAV* 36-37, pp. 123-157.
- BLANCHE-BENVENISTE Claire (1991). « À propos des énoncés sans verbe : les énoncés réponses », *Recherches sur le français parlé* 11, Université de Provence, pp. 57-85.
- BLANCHE-BENVENISTE Claire (1993). « Faire des phrases », *Le Français aujourd'hui* 101, pp. 7-15.
- BLANCHE-BENVENISTE Claire (1995). « De la rareté de certains phénomènes syntaxiques en français parlé », *Journal of French Language Studies* 5, pp. 17-29.
- BLANCHE-BENVENISTE Claire (1997). « Transcriptions et technologies », *Recherches sur le français parlé* 14, Université de Provence, pp. 87-99.
- BLANCHE-BENVENISTE Claire (2000a). « Convergences de matériel grammatical permettant d'établir des typologies textuelles », M. Bilger (Éd.), *Linguistique sur corpus. Études et réflexions*, Cahiers de l'Université de Perpignan 31, pp.103-116.
- BLANCHE-BENVENISTE Claire (2000b). « Transcription de l'oral et morphologie », M. Gille et R. Kiesler (Éds.), *Romania Una et diversa, Philologische Studien für Theodor Berchem*, Tübingen, Gunter Narr, pp. 61-74.
- BLANCHE-BENVENISTE Claire (2000c). *Approches de la langue parlée en français*, Paris, Ophrys.
- BLANCHE-BENVENISTE Claire (2002a). « Phrase et construction verbale », M. Charolles, P. Le Goffic et M.-A. Morel (Éds.), *Y a-t-il une syntaxe au-delà de la phrase ?*, *Verbum* 24 (1-2), pp. 7-22.
- BLANCHE-BENVENISTE Claire (2002b). « Quel est le rôle du français parlé dans les évolutions syntaxiques ? », *L'Information grammaticale* 94, pp. 11-17.
- BLANCHE-BENVENISTE Claire (2002c). « Réflexions sur les transcriptions de corpus de français parlé », *Revue Parole* 22-23-24, pp.91-117.
- BLANCHE-BENVENISTE Claire (2003a). « La langue parlée », M. Yaguello (Dir.), *Le Grand Livre de la langue française*, Paris, Seuil, pp. 317-343.
- BLANCHE-BENVENISTE Claire (2003b). « La naissance des syntagmes dans les hésitations et répétitions du parler », J.-L. Araoui (Éd.), *Le sens et la mesure. Hommages à Benoît de Cornulier*, Paris, Honoré Champion, pp. 40-55.

-
- BLANCHE-BENVENISTE Claire (2005). « L'étude grammaticale des corpus de langue parlée en français », G. Williams (Éd.), *La Linguistique de corpus*, Rennes, Presses universitaires de Rennes, pp. 47-66.
- BLANCHE-BENVENISTE Claire, BOREL Bernard, DEULOFEU José, DURAND Jacky, GIACOMI Alain, LOUFRANI Claude, MEZIANE Boudjema, PAZERY Nelly (1979). « Des grilles pour le français parlé », *Recherches sur le français parlé 2*, Université de Provence, pp. 163-205.
- BLANCHE-BENVENISTE Claire, JEANJEAN Colette (1987). *Le Français parlé. Transcription et édition*, Paris, Didier Érudition.
- BLANCHE-BENVENISTE Claire, BILGER Mireille, ROUGET Christine, VAN DEN EYNDE Karel (1990). *Le Français parlé. Études grammaticales*, Paris, CNRS Éditions.
- BLANCHE-BENVENISTE Claire, BILGER Mireille (1999). « Français parlé - oral spontané, Quelques réflexions », *Revue Française de Linguistique Appliquée* 4 (2), pp. 21-30.
- BLANCHE-BENVENISTE Claire, ROUGET Christine, SABIO Frédéric (Éds.) (2002). *Choix de textes de français parlé. 36 extraits*. Paris, Honoré Champion.
- BLANKENSHIP Jane, KAY Christian (1964). « Hesitation phenomena in English Speech : a study in distribution », *Word* 20, pp. 360-372.
- BLASCO-COUTURIER Mylène, SAVELLI Marie-Josée (1990). « Transcrire, traduire, orthographier le français parlé. De l'impossible copie à la falsification des données orales », *Recherches sur le français parlé* 10, Université de Provence, pp. 19-39.
- BLASCO-DULBECCO Mylène, CAPPEAU Paul (2005). « Ce que les corpus oraux nous apprennent sur les adjectifs », G. Williams (Éd.), *La Linguistique de corpus*, Rennes, Presses universitaires de Rennes, pp. 69-80.
- BOERSMA Paul, WEENINK David (2006). *Praat : doing phonetics by computer* (Version 4.5.08) <http://www.praat.org/>.
- BOLASCO Sergio (1993). « Choix de lemmatisation en vue des reconstructions syntagmatiques du texte par l'analyse des correspondances », *Actes des Secondes Journées Internationales d'Analyse Statistique de Données Textuelles (JADT 1993)*, Paris, Telecom Paris.
- BOLASCO Sergio (2000). « TALTAC : un environnement pour l'exploitation de ressources statistiques et linguistiques dans l'analyse textuelle. Un emploi d'application au discours politique », *Actes des JADT 2000, 5^{es} Journées Internationales d'Analyse des Données Textuelles* (9-11 mars, Lausanne).

- BONHOMME Patrice (2000). « Codage et normalisation de ressources textuelles », J.-M. Pierrel (Éd.), *Ingénierie des langues*, Paris, Hermès, pp. 173-191.
- BOOMER Donald S., DITTMAN Allen T. (1962). « Hesitation pauses and juncture pauses in speech », *Language and Speech* 5, pp. 215-220.
- BOSREDON Bernard, TAMBA Irène (1991). « Verre à pied, moule à gaufres : prépositions et noms composés de sous-classe », *Langue française* 91, pp. 40-55.
- BRANIGAN Holly, LICKLEY Robin, MCKELVIE David (1999). « Non-linguistic influences on rates of disfluency in spontaneous speech », *Proceedings of the ICPhS*, San Francisco, pp. 387-390.
- BRES Jacques, GARDES-MADRAY Françoise (1991). « Ratages et temps de l'à-dire », H. Parret (Éd.), *Le Sens et ses hétérogénéités*, Paris, Éditions du CNRS.
- BRILL Eric (1993). *A Corpus-Based Approach To Language Learning*, Department of Computer and Information Science, University of Pennsylvania, Thèse non publiée.
- BRILL Eric (1995). « Transformation-based error-driven learning and natural language processing : a case study in part-of-speech tagging », *Computational linguistics* 21 (4), pp. 543-565.
- BRILL Eric (1999a). « Tagging unknown words », H. Van Halteren (Éd.), *Syntactic Wordclass Tagging*, Dordrecht, Kluwer Academic Publishers, pp. 207-211.
- BRILL Eric (1999b). « Corpus-based rules », H. Van Halteren (Éd.), *Syntactic Wordclass Tagging*, Dordrecht, Kluwer Academic Publishers, pp. 247-262.
- BRUNET Étienne (1983). « Le vocabulaire de Proust », *Travaux de linguistique quantitative* 18, Genève-Paris, Slatkine-Champion.
- BRUNET Étienne (2000). « Qui lemmatise, dilemme attise », *Lexicometrica* 2, <http://www.cavi.univ-paris3.fr/lexicometrica/article/numero2/brunet2000.html>.
- BURNAGE Gavin (1990). *CELEX. A Guide for Users*, Nijmegen, Center for lexical Information, University of Nijmegen.
- BURNARD Lou (1995). « The Text Encoding Initiative : an overview », G. Leech, G. Myers, J. Thomas (Éds.), *Spoken English on Computer. Transcription, Mark-up and Application*, New York, Longman, pp. 69-81.
- BUVET Pierre-André (1994). « Détermination : les noms », *Lingvisticae Investigationes* XVIII (1), pp. 121-150.



Cahiers GUTenberg 24 1996. *TEI : Text Encoding Initiative*.

CAMPIONE Estelle (2004). « Étude des interactions entre pauses silencieuses et pauses remplies en français parlé », *Recherches sur le français parlé* 18, Université de Provence, pp. 185-200.

CAMPIONE Estelle, VÉRONIS Jean (2001). « Étiquetage prosodique semi-automatique des corpus oraux », *Actes de TALN 2001* (2-5 juillet 2001, Tours).

CAMPIONE Estelle, VÉRONIS Jean (2004). « Pauses et hésitations en français spontané », *Actes des 25^{es} Journées d'Étude sur la Parole (JEP'2004)* (19-22 avril, Fès), pp. 109-112.

CAMPIONE Estelle, VÉRONIS Jean, DEULOFEU José (2005). « The French corpus », E. Cresti et M. Moneglia (Éds.), *C-ORAL-ROM, Integrated Reference Corpora for Spoken Romance Languages* (Amsterdam, John Benjamins, pp. 111-133.

CAMUGLI GALLARDO Catherine (2003). « *Qu'est-ce que tu chantes là ?* Syntaxe et lexique dans les expressions métaphoriques figées », *Cahiers de lexicologie*, pp. 175-192.

CANDEA Maria (2000a). *Contribution à l'étude des pauses silencieuses et des phénomènes dits "d'hésitation" en français oral spontané*, Université de Paris-3 Sorbonne-nouvelle, Thèse non publiée.

CANDEA Maria (2000b). « Les *euh* et les allongements dits "d'hésitation" : deux phénomènes soumis à certaines contraintes en français oral non lu », *Actes des 23^{es} Journées d'Étude sur la Parole (JEP'2000)* (19-23 juin, Aussois).

CANDEA Maria (2002). « Le *e d'appui* parisien : statut actuel et progression », *Actes des 24^{es} Journées d'Étude sur la Parole* (24-27 juin 2002, Nancy).

CAPPEAU Paul (1997). « Données erronées : quelles erreurs commettent les transcripteurs ? », *Recherches sur le français parlé* 14, Université de Provence, pp. 117-126.

CAPPEAU Paul (1998). « Quelques mots sur quelques bribes liés au genre », M. Bilger, K. van den Eynde et Fr. Gadet (Éds.), *Analyse linguistique et approches de l'oral. Recueil d'études offert en hommage à Claire Blanche-Benveniste*, *Orbis Supplementa* 10, Leuven, Peeters, pp. 301-311.

CARROLL John, FRANK Annette, LIN Dekang, PRESCHER Detlef, USZKOREIT Hans (Éds.) (2002). *Proceedings of the Workshop Beyond PARSEVAL. Towards Improved Evaluation Measures for Parsing Systems, the 3rd International Conference on Language Resources and Evaluation* (June 2, Las Palmas, Gran Canaria).

-
- CARTONI Bruno (2006). « Constance et variabilité de l'incomplétude lexicale », P. Mertens, C. Fairon, A. Dister et P. Watrin (Éds.), *Verbum ex machina, Actes de TALN 2006, Cahiers du Cental 2* (2), Presses universitaires de Louvain, Louvain-la-Neuve, pp. 661-669.
- CATACH Nina (1994). *La Ponctuation*, Paris, Presses universitaires de France (coll. « Que sais-je ? »).
- CHAFE Wallace (1995). « Adequacy, user-friendliness, and practicality in transcribing », G. Leech, G. Myers, J. Thomas (Éds.), *Spoken English on Computer. Transcription, Mark-up and Application*, New York, Longman, pp. 54-61.
- CHANOD Jean-Pierre, TAPANAINEM Pasi (1995a). « Tagging French. Comparing a statistical and a constraint-based method », *Proceedings of EACL'95*, pp. 149-156.
- CHANOD Jean-Pierre, TAPANAINEM Pasi (1995b). « Creating a tagset, lexicon and guesser for a French tagger », *From Texts To Tags : Issues In Multilingual Language Analysis, EACL SIGDAT workshop*, Dublin, University College Dublin, pp. 58-64.
- CHEEPEN Christine (1995). « Discourse considerations in transcription and analysis », G. Leech, G. Myers, J. Thomas (Éds.), *Spoken English on Computer. Transcription, Mark-up and Application*, New York, Longman, pp. 135-143.
- CHEVALIER Gisèle, KASPARIAN Sylvia, SILBERZTEIN Max (2004). « Éléments de solution pour le traitement automatique d'un français oral régional », *TAL* 45 (2), pp. 41-62.
- CHEVALIER Jean-Claude (1966). « Éléments pour une description du groupe nominal. Les prédéterminants du substantif », *Le Français moderne* 34, pp. 241-253.
- CHEVALIER Jean-Claude (1976). « Le jeu des exemples dans la théorie grammaticale, étude historique », J.-Cl. Chevalier (Éd.), *Grammaire transformationnelle : syntaxe et lexique*, Lille, Presses universitaires de Lille, pp. 235-261.
- CLARK Herbert, CLARK Eve (1977). *Psychology and Language*, New York, Harcourt-Brace-Jovanovitch.
- CLÉMENT Lionel (2001). *Construction et exploitation d'un corpus syntaxiquement annoté pour le français*, Université Paris-7, Thèse non publiée.
- CLERC-RENAUD Jeanne, VASILESCU Ioana, CANDEA Maria, ADDA-DECKER Martine (2004). « Étude acoustique et perceptive des hésitations autonomes multilingues », *Actes des 25^{es} Journées d'Étude sur la Parole (JEP'2004)* (19-22 avril, Fès).

-
- CLINQUART Anne-Marie (2000). « La répétition, une figure de reformulation à revisiter », P. Anderson, M. Mandini, A. Chauvin-Vileno (Éds.), *Répétition, altération, reformulation*, Besançon, Presses de l'Université de Franche-Comté, pp. 323-349.
- CLOEREN Jan (1999). « Tagsets », H. Van Halteren (Éd.), *Syntactic Wordclass Tagging*, Dordrecht, Kluwer Academic Publishers, pp. 37-54.
- CONSTANT Matthieu (2000). « Description d'expressions numériques en français », A. Dister (Éd.), *RISSH : Revue, Informatique et Statistiques dans les sciences humaines* 36, Liège, CIPL, Université de Liège.
- CONSTANT Matthieu (2003). *Grammaires locales pour l'analyse automatique de textes : Méthodes de construction et outils de gestion*, Université de Marne-la-Vallée, Thèse non publiée.
- COOK Guy (1995). « Theoretical issues : transcribing the untranscribable », G. Leech, G. Myers, J. Thomas (Éds.), *Spoken English on Computer. Transcription, Mark-up and Application*, New York, Longman, pp. 35-53.
- COOK Mark (1969). « Anxiety Speech Disturbance And Speech Rate », *British Journal of Social and Clinical Psychology* 8, pp. 13-21.
- COOK Mark (1971). « The Incidence of Filled Pauses in Relation to Part of Speech », *Language and Speech* 14, pp. 135-150.
- COSNIER Jacques (1987). « L'éthologie du dialogue », J. Cosnier, C. Kerbrat-Orecchioni (Éds.), *Décrire la conversation*, Lyon, Presses Universitaires de Lyon (coll. « Linguistique et Sémiologie »), pp. 291-315.
- COSNIER Jacques, KERBRAT-ORECCHIONI Catherine (Dir.) (1987). *Décrire la conversation*, Lyon, Presses Universitaires de Lyon (coll. « Linguistique et Sémiologie »).
- COSTE Daniel (1986). « Auto-interruptions et reprises », *Paroles inachevées. Études linguistiques pragmatiques et processuelles de conversations, brouillons et textes*, *DRLAV* 34-35, pp. 127-139.
- COURTOIS Blandine (1990). « Un système de dictionnaires électroniques pour les mots simples du français », *Langue française* 87, Paris, Larousse, pp. 11-22.
- COURTOIS Blandine (1994). « Buts et méthodes de l'élaboration des dictionnaires électroniques du LADL », *Cahiers du CIEL* 1994-1995, pp. 87-107.
- COURTOIS Blandine (1996). « Les formes ambiguës de la langue française », *Linguisticae Investigationes* 20 (1), Amsterdam-Philadelphia, John Benjamins, pp. 167-202.

-
- COURTOIS Blandine, SILBERZTEIN Max (Éds.) (1990). *Dictionnaires électroniques du français, Langue française* 87.
- CROWDY Steve (1993). « Spoken corpus design and transcription », *Literary and Linguistic Computing* 8 (4), pp. 259-265.
- CROWDY Steve (1995). « The BNC spoken corpus », G. Leech, G. Myers, J. Thomas (Éds.), *Spoken English on Computer. Transcription, Mark-up and Application*, New York, Longman, pp. 224-234.
- DAELEMANS Walter (1999). « Machine Learning approaches », H. Van Halteren (Éd.), *Syntactic Wordclass Tagging*, Dordrecht, Kluwer Academic Publishers, pp. 285-304.
- DANJOU-FLAUX Nelly, FICHEZ-VALLEZ Élisabeth (1985). « Linguistique taxinomique et grammaire générative : le traitement de l'exception », D. Corbin, A.-M. Dessaux-Berthonneau (Dir.), *L'Exception, Langue française* 66, pp. 99-121.
- DANLOS Laurence (Éd.) (1988). *Les Expressions figées, Langages* 90.
- DEBAISIEUX Jeanne-Marie (2005). « Les corpus oraux : situation, exploitation linguistique. Bilan et perspectives », *SCOLIA* 19, pp. 9-40.
- DEGAND Liesbeth, SIMON Anne Catherine (2005). « Minimal Discourse Units : can we define them, and why should we ? », *Actes de SEM'05*. <http://www.univ-tlse2.fr:8880/erss/index.jsp?perso=bras&subURL=sem05/>.
- DELATTE Louis, DUCHESNE-DEGEY M., GOVAERTS Susanne, DENOZ Joseph (1977). « Le traitement automatique de la langue française au Laboratoire d'Analyse Statistique des Langues Anciennes de l'Université de Liège », *Cahiers de lexicologie* 31, pp. 3-32.
- DELIC (2004). « Présentation du Corpus de référence du français parlé », *Recherches sur le français parlé* 18, pp. 11-42.
- DEROUBAIX Jean-Claude, GOBIN Corinne (2000). « Le roi règne, il ne gouverne pas. Il s'adresse au peuple belge », *Actes des JADT 2000, 5^{es} Journées Internationales d'Analyse des Données Textuelles* (9-11 mars, Lausanne), pp. 103-110.
- DISTER Anne (1995). *L'Apocope, un phénomène vivant du français contemporain*, Département d'Études romanes, Université de Liège, Mémoire inédit de licence.
- DISTER Anne (1999a). « Construire des grammaires de levée d'ambiguïtés pour INTEX », C. Fairon (Éd.), *Analyse lexicale et syntaxique : le système INTEX, Lingvisticæ Investigationes* 22 (1-2), Amsterdam-Philadelphia, John Benjamins, pp. 233-247.

-
- DISTER Anne (1999b). « La levée d'ambiguïtés morphologiques par grammaires locales. Le système Intex », *Actes des XIIIes Journées de Linguistique (25 et 26 mars 1999)*, Québec, Presses de l'Université Laval, p. 57-66.
- DISTER Anne (1999c). « De l'étiquetage traditionnel au transducteur du texte : la levée d'ambiguïtés par grammaires locales », *Revue, Informatique et Statistiques dans les sciences humaines* 35, Liège, CIPL, p. 9-24.
- DISTER Anne (2000a). « La presse québécoise vue à travers l'analyse de ses particularités lexicales », *Actes des 14^{es} Journées de Linguistique (23 et 24 mars 2000)*, Québec, Presses de l'Université Laval, pp. 57-66.
- DISTER Anne (2000b). « Réflexions sur l'homographie et la désambiguïsation des formes les plus fréquentes », *Actes des JADT 2000, Journées d'Analyse des Données Textuelles*, Lausanne, p. 131-138.
- DISTER Anne (2001). « Levée d'ambiguïtés sur les mots lexicaux et grammaticaux » (fascicule spécial *Description et levée des ambiguïtés*, É. Laporte (Éd.), *Linguisticæ Investigationes* 24 (1), Amsterdam/Philadelphia, John Benjamins, pp. 105-126.
- DISTER Anne (2004). *Les grammaires ELAG pour le français*. <http://www-igm.univ-mlv.fr/~unitex/>.
- DISTER Anne, FAIRON Cédric (2004). « Extension des ressources lexicales grâce à un corpus dynamique », *Lexicometrica*. <http://www.cavi.univ-paris3.fr/lexicometrica/>.
- DISTER Anne, MOREAU Marie-Louise (2006). « Dis-moi comment tu féminises, je te dirai pour qui tu votes. Les dénominations des candidates dans les élections européennes de 1989 et de 2004 en Belgique et en France », *Langage et Société* 115, pp. 5-45.
- DISTER Anne, FRANCARD Michel, GERON Geneviève, HAMBYE Philippe, SIMON Anne Catherine, WILMET Régine (2006). *Conventions de transcription régissant les corpus de la banque de données VALIBEL*. <http://valibel.fltr.ucl.ac.be/>, rubrique corpus oraux, conventions de transcription.
- DISTER Anne, SIMON Anne Catherine, à paraître (2007). « La transcription synchronisée des corpus oraux. Un aller-retour entre théorie, méthodologie et traitement informatisé », *Arena Romanistica*, Bergen.
- DISTER Anne, FAIRON Cédric, WATRIN Patrick, à paraître (2007a). « Recherche d'expressions figées du français en Belgique. Méthode et observations », *Cahiers de l'Institut de Linguistique de Louvain*.
- DISTER Anne, FRANCARD Michel, HAMBYE Philippe, SIMON Anne Catherine, à paraître (2007b). « Du son, du texte, des méta-données.

-
- L'évolution de la banque de données textuelles orales VALIBEL (1989-2006) », *Interactions & Langages*.
- DOPPAGNE Albert (1966). *Trois Aspects du français contemporain*, Paris, Larousse.
- DOPPAGNE Albert (1979). *Belgicisms de bon aloi*. Bruxelles, Office du bon langage (Fondation Plisnier).
- DOSTIE (Gaétane) 1995. « Réflexions sur le phénomène des mots-phrases », J. Labelle et Ch. Leclère (Éds.), *Lexiques-grammaires comparés en français. Actes du colloque international de Montréal, Lingvisticæ Investigationes Supplementa 17*, Amsterdam-Philadelphia, John Benjamins, pp. 175-188.
- DRILLON Jacques (1991). *Traité de la ponctuation française*, Paris, Gallimard.
- DU BOIS John W. (1991). « Transcription design principles for spoken discourse research », *Pragmatics* 1 (1), pp. 71-106.
- DU BOIS John W., SCHUETZE-COBURN Stephan (1993a). « Representing hierarchy : Constituent Structure for Discourse Databases », J.A. Edwards and M.D. Lampert (Éds.), *Talking Data. Transcription and Coding in Discourse Research*, Hillsdale, Lawrence Erlbaum Associates, pp. 221-260.
- DU BOIS John W., SCHUETZE-COBURN Stephan, CUMMING Susanna, PAOLINO Danae (1993b). « Outline of Discourse Transcription », J.A. Edwards and M.D. Lampert (Éds.), *Talking Data. Transcription and Coding in Discourse Research*, Hillsdale, Lawrence Erlbaum Associates, pp. 45-89.
- DUEZ Danielle (1991). *La Pause dans la parole de l'homme politique*, Paris, Éditions du CNRS.
- DUEZ Danielle (1995). « Perception of hesitations in spontaneous French speech », *Proceedings of the 23rd International Congress of Phonetic Sciences*, vol. 2, Stockholm, pp. 498-501.
- DUEZ Danielle (1997). « La signification des pauses dans la production et la perception de la parole », *Revue Parole* 3-4, pp. 275-299.
- DUEZ Danielle (2001). « Signification des hésitations dans la production et la perception de la parole spontanée », *Revue Parole* 17-18-19, pp. 113-137.
- DURAND Jacques (1970). « Rhétorique et image publicitaire », *Communications* 15, pp. 70-95.
- EDWARDS Jane A. (1993). « Principles and Contrasting Systems of Discourse Transcription », J.A. Edwards and M.D. Lampert (Éds.),

-
- Talking Data. Transcription and Coding in Discourse Research*, Hillsdale, Lawrence Erlbaum Associates, pp. 3-31.
- EDWARDS Jane A. (1995). « Principles and alternative systems in the transcription, coding and mark-up of spoken discourse », G. Leech, G. Myers, J. Thomas (Éds.), *Spoken English on Computer. Transcription, Mark-up and Application*, New York, Longman, pp. 19-34.
- EHlich Konrad (1993). « HIAT : A transcription System for Discourse Data », J.A. Edwards and M.D. Lampert (Éds.), *Talking Data. Transcription and Coding in Discourse Research*, Hillsdale, Lawrence Erlbaum Associates, pp. 123-166.
- EHMER Oliver (2006). « The Tranformer – a tool for working with aligned language data », communication présentée à la 3^e rencontre fribourgeoise de la linguistique sur corpus appliquée aux langues romanes, Freiburg-im-Breisgau, 14-17 septembre 2006.
- EL-BÈZE Marc, MERIALDO Bernard (1999). « Hidden Markov Models », H. Van Halteren (Éd.), *Syntactic Wordclass Tagging*, Dordrecht, Kluwer Academic Publishers, pp. 263-284.
- EL-BÈZE Marc, SPRIET Thierry (1995). « Étiquetage probabiliste et contraintes syntaxiques », *Actes de TALN 1995* (15-16 juin, Marseille).
- ERBACH Gregor (1991). « A Flexible Parser for a Linguistic Development Environment », *Text Understanding in LILOG*, Berlin, Springer (coll. « Lectures Notes on Artificial Intelligence »), pp. 74-87.
- ESQUENET-BERNAUDIN Monique (1985). « Lisibilité de l'oral : une gageure », M.-A. Morel (Dir.), *L'Oral du débat, Langue française* 65, pp. 17-27.
- ESTÈVE Yannick, DELÉGLISE Paul, JACOB Bruno (2004). « Système de transcription automatique de la parole et logiciels libres », *TAL* 45(2), pp. 15-39.
- FAIRON Cédric (1998). « La presse belge au crible de l'analyse automatique. Application du système INTEX », J. Klein, B. Lamiroy et J.-M. Pierret (Éds.), *Théorie Linguistiques et applications informatiques. Actes du 16^e colloque européen sur la grammaire et le lexique comparés (24-27 septembre 1997)*, Cahiers de l'Institut de linguistique de Louvain, Louvain-la-Neuve.
- FAIRON Cédric (2000). *Structures non connexes. Grammaire des incises en français : description linguistique et outils informatiques*, Université de Paris-7, thèse non publiée.
- FAIRON Cédric, COURTOIS Blandine (1998). « Les corpus dynamiques et GlossaNet. Extension de la couverture lexicale des dictionnaires

- électroniques anglais », *Actes des JADT 2000, 5^{es} Journées Internationales d'Analyse des Données Textuelles* (9-11 mars, Lausanne).
- FAIRON Cédric, SENELLART Jean (1999). « Réflexion sur la localisation, l'étiquetage, la reconnaissance et la traduction d'expressions linguistiques complexes », *Actes de TALN 1999* (12-17 juillet 1999, Cargèse).
- FAYOL Michel (1989). « Une approche psycholinguistique de la ponctuation. Étude en production et compréhension », *Langue française* 81, pp. 21-39.
- FERRÉ Gaëlle (2003). « Les pauses démarcatives déplacées en anglais spontané : marquages prosodiques et kinésiques », *LIDIL* 26, pp. 155-169.
- FERRÉ Gaëlle (2004). « Les pauses intra-constituants en anglais spontané », *Actes des 25^{es} Journées d'Étude sur la Parole (JEP'2004)* (19-22 avril, Fès), pp. 217-220.
- FIALA Pierre, BOUTET Josiane (1986). « Les télescopages syntaxiques », *Paroles inachevées. Études linguistiques pragmatiques et processuelles de conversations, brouillons et textes, DRLAV* 34-35, pp. 111-126.
- FLETCHER Paul, GARMAN Michael (1995). « Transcription, segmentation and analysis : corpora from the language-impaired », G. Leech, G. Myers, J. Thomas (Éds.), *Spoken English on Computer. Transcription, Mark-up and Application*, New York, Longman, pp. 116-127.
- FONTANEY Louise (1987). « Intonation et régulation de l'intonation », J. Cosnier, C. Kerbrat-Orecchioni (Éds.), *Décrire la conversation*, Lyon, Presses Universitaires de Lyon (coll. « Linguistique et Sémiologie »), pp. 225-267.
- FOUCOU Pierre-Yves (1998). « Formes répétées, formes figées », *BULAG* 23, pp. 67-88.
- FRANCARD Michel (1990). « Le français parlé des corpus oraux », *Travaux de Linguistique* 21, pp. 53-63.
- FRANCARD Michel (1997). « Le français en Wallonie », D. Blampain, A. Goosse, J.-M. Klinkenberg et M. Wilmet (Dir.), *Le Français en Belgique*, Louvain-la-Neuve, Duculot, pp. 229-237.
- FRANCARD Michel, PERONNET Louise (1989). « La transcription de corpus oraux dans une perspective comparative. La démarche du projet PLURAL », *Recherche en linguistique appliquée à l'informatique (RELAI)*, Québec, CIRB, pp. 295-307.

-
- FRANCARD Michel, LAMBERT Joëlle, MASUY Françoise (1993). *L'Insécurité linguistique en Communauté française de Belgique*, Bruxelles-Louvain-la-Neuve, Service de la langue française-Duculot (coll. « Français & Société », n°6).
- FRANCARD Michel, GERON Geneviève, WILMET Régine (2002). « Diffusion et vitalité des particularités lexicales du français en Belgique : une enquête sociolinguistique », R. Wakely (Dir.), *Les Belges : enregistreurs de tous les usages*, Edinburgh, French Section and Centre de recherches francophones belges of the School of European Languages and Cultures, University of Edinburgh, pp. 11-32.
- FRANCARD Michel, GERON Geneviève, WILMET Régine (2003). « Les 'belgicismes' sont-ils 'belges' ? La nomenclature du Dictionnaire du français en Belgique », P. Nobel (Éd.), *Variations linguistiques. Koinè, dialectes, français régionaux*, Besançon, Presses Universitaires de Franche-Comté, pp. 137-150.
- FRANCIS W. Nelson, KUCERA Henry (1979 [1964]). *Brown Corpus Manual*. <http://icame.uib.no/brown/bcm.html>.
- FRANCOPOULO Gil (2003). « TagChunker : mécanisme de construction et évaluation », *Actes de TALN 2003* (11-14 juin, Batz-sur-Mer), pp. 95-104.
- FREI Henri (1929). *La Grammaire des fautes*, Paris-Genève-Leipzig, Paul Geuthner-Kundig-Otto Harrassowitz.
- FREI Henri (1953). *Le Livre des deux mille phrases*, Genève, Droz.
- FUCHS Catherine (1993). *Linguistique et Traitements automatiques des langues*, Paris, Larousse.
- FUCHS Catherine (1996). *Les Ambiguïtés du français*, Gap-Paris, Ophrys.
- GADET Françoise (1990). « Les outils grammaticaux au risque de l'oral non standard », *Travaux de Linguistique* 21, pp. 13-23.
- GADET Françoise (1992). *Le Français populaire*, Paris, Presses universitaires de France (coll. « Que sais-je ? »).
- GADET Françoise (1997). *Le Français ordinaire*, Paris, Armand Colin.
- GADET Françoise (1998). « Le 'français avancé' à l'épreuve de ses données », M. Bilger, K. van den Eynde et Fr. Gadet (Éds.), *Analyse linguistique et approches de l'oral. Recueil d'études offert en hommage à Claire Blanche-Benveniste*, *Orbis Supplementa* 10, Leuven, Peeters, pp. 59-68.
- GADET Françoise (2000a). « On n'en a pas fini avec les problèmes de recueil de corpus », H.L. Andersen et A.B. Hansen (Éds.), *Le Français parlé*.

-
- Corpus et résultats. Actes du colloque international, Université de Copenhague du 29 au 30 octobre 1998, Études Romanes 47*, pp. 29-44.
- GADET Françoise (2000b). « Derrière les problèmes méthodologiques du recueil des données », M. Bilger (Éd.), *Linguistique sur corpus. Études et réflexions, Cahiers de l'Université de Perpignan 31*, pp. 59-75.
- GADET Françoise (2003). *La Variation sociale en français*, Paris, Ophrys.
- GARNHAM Alan, SHILLCOCK Richard C., BROWN Gordon D. A., MILL Andrew I. D., CUTLER Anne (1981). « Slips of the tongue in the London-Lund corpus of spontaneous conversation », A. Cutler (Éd.), *Slips of the tongue and language production, Linguistics 19 (7-8) (245-246)*, pp. 804-817.
- GARRIGUES Mylène (1992). « Dictionnaires hiérarchiques du français », *Langue française 96*, pp. 88-100.
- GARRIGUES Mylène (1993). *Méthode de paramétrage des dictionnaires et grammaires électroniques. Application à des systèmes interactifs en langue naturelle*. Université Paris-7, Thèse non publiée.
- GARRIGUES Mylène (1997). « Une méthode de désambiguïsation locale Nom/Adjectif pour l'analyse automatique de textes », *Langages 126*, pp. 60-78.
- GARSDIE Roger (1995). « Grammatical tagging of the spoken part of the British National Corpus : a progress report », G. Leech, G. Myers, J. Thomas (Éds.), *Spoken English on Computer. Transcription, Mark-up and Application*, New York, Longman, pp. 161-167.
- GARSDIE Roger, LEECH Geoffrey, MCENERY Tony (Éds.) (1997). *Corpus annotation : linguistic information from Computer Text Corpora*, White Plains (NY)-London, Longman.
- GARY-PRIEUR Marie-Noëlle (1985). *De la grammaire à la linguistique. L'étude de la phrase*, Paris, Armand Colin.
- GASIGLIA Nathalie (2005). « Stratégie de constitution de corpus oraux transcrits : arguments pour un corpus plurithématique à haut rendement », G. Williams (Éd.), *La Linguistique de corpus*, Rennes, Presses universitaires de Rennes, pp. 219-232.
- GAULMYN Marie-Madeleine de (1987). « Les régulateurs verbaux : le contrôle des récepteurs », J. Cosnier, C. Kerbrat-Orecchioni (Éds.), *Décrire la conversation*, Lyon, Presses Universitaires de Lyon (coll. « Linguistique et Sémiologie »), pp. 203-223.
- GEDO (1997). « Transcription de l'oral et interprétation. Illustration de quelques difficultés », *Recherches sur le français parlé 14*, Université de Provence, pp. 57-86.

-
- GENDNER Véronique, ADDA-DECKER Martine (2002). « Analyse comparative de corpus oraux et écrits français : mots, lemmes et classes morpho-syntaxiques », *Actes des 24^{es} Journées d'Étude sur la Parole* (24-27 juin 2002, Nancy), pp. 13-16.
- GIBBON Daffyd, MOORE Roger, WINSKI Richard (1997). *Handbook of Standards and Ressources for Spoken Language Systems*, Berlin, Mouton de Gruyter.
- GIOVANNONI Dominique-Catherine (1990). « Contribution à une bibliographie sur les problèmes d'écoute et de perception du langage », *Recherches sur le français parlé* 10, Université de Provence, pp. 39-50.
- GIOVANNONI Dominique-Catherine, SAVELLI Marie-Josée (1990). « Transcrire, traduire, orthographier le français parlé. De l'impossible copie à la falsification des données orales », *Recherches sur le français parlé* 10, Université de Provence, pp. 19-37.
- GOODY Jack (1979). *La Raison graphique. La domestication de la pensée sauvage*, Paris, Éditions de Minuit.
- GOODY Jack, WATT Ian (1963). « The Consequences of Literacy », *Comparative Studies in Society and History* 5, pp. 304-345.
- GOOSSE André (1977). « Qu'est-ce qu'un belgicisme ? », *Bulletin de l'Académie royale de Langue et Littérature française de Belgique*, pp. 345-367.
- GOOSSE André (1991). *La « Nouvelle » Orthographe. Exposé et commentaires*, Paris-Louvain-la-Neuve, Duculot.
- GOOSSE André (1995). « Locutions régionales de Belgique. Origine et vitalité », J. Labelle et Ch. Leclère (Éds.), *Lexiques-grammaires comparés en français. Actes du colloque international de Montréal, Lingvisticae Investigationes Supplementa* 17, Amsterdam-Philadelphia, John Benjamins, pp. 29-36.
- GOOSSE André (2000). « Le bon usage comme référence », M. Francard, G. Geron et R. Wilmet (Éds.), *Le Français de référence. Constructions et appropriations d'un concept*, *Cahiers de l'Institut de linguistique de Louvain* 26 (1-4), pp. 75-85.
- GOUGENHEIM Georges, MICHÉA René, RIVENC Paul, SAUVAGEOT Aurélien (1964). *L'Élaboration du français fondamental (1^{er} degré). Étude sur l'établissement d'un vocabulaire et d'une grammaire de base* (nouvelle édition refondue et augmentée), Paris, Didier.
- GOULIAN Jérôme, ANTOINE Jean-Yves, POIRIER Franck (2002). « Compréhension automatique de la parole et TAL : une approche syntaxico-sémantique pour le traitement des inattendus structuraux du français parlé », *Actes de TALN 2002* (24-27 juin, Nancy).

-
- GREENBAUM Sidney (1992). « A New Corpus of English : ICE », J. Svartvik (Éd.) *Directions in Corpus Linguistics : Proceedings of Nobel Symposium 82* (Stockholm), Berlin, Mouton de Gruyter, pp. 171-179.
- GREENE Barbara B., RUBIN Gerald M. (1971). *Automatic grammatical tagging of English*, Providence, Brown University, Department of Linguistics.
- GREFENSTETTE Gregory (1999). « Tokenization », H. Van Halteren (Éd.), *Syntactic Wordclass Tagging*, Dordrecht, Kluwer Academic Publishers, pp. 117-133.
- GREFENSTETTE Gregory, TAPANAINEN Pasi (1994). « What is a word, what is a sentence ? Problems of Tokenization », *Proceedings of the 3rd Conference on Computational Lexicography and Text Research (COMPLEX '94)* (September 26-28, Budapest).
- GROBET Anne (1997). « La ponctuation prosodique dans les dimensions périodique et informationnelle du discours », *Cahiers de Linguistique française* 19, pp. 83-123.
- GROBET Anne (1998). « Le rôle des ponctuateurs dans le marquage des unités périodiques, à la lumière d'un exemple tiré de *Fin de partie* de Samuel Beckett », J.-M. Defays, L. Rosier et Fr. Tilkin (Éds.), *À qui appartient la ponctuation ?*, Actes du colloque international et interdisciplinaire de Liège (13-15 mars 1997), Louvain-la-Neuve, De Boeck & Larquier (coll. « Champs linguistiques »), pp. 99-116.
- GROSJEAN François, DESCHAMPS Alain (1972). « Analyse des variables temporelles du français spontané », *Phonetica* 26, pp. 129-156.
- GROSJEAN François, DESCHAMPS Alain (1973). « Analyse des variables temporelles du français spontané. II. Comparaison du français oral dans la description avec l'anglais (description) et avec le français (interview radiophonique) », *Phonetica* 28, pp. 191-226.
- GROSJEAN François, DESCHAMPS Alain (1975). « Analyse contrastive des variables temporelles de l'anglais et du français : vitesse de parole et variables composantes, phénomènes d'hésitation », *Phonetica* 31, pp. 144-184.
- GROSS Gaston (1996a). *Les Expressions figées en français*, Paris, Ophrys.
- GROSS Gaston (1996b). « Degré de figement des noms composés », *Langages* 90, pp. 57-72.
- GROSS Maurice (1982). « Une classification des phrases "figées" du français », *Revue québécoise de linguistique* 11 (2), pp. 151-185.
- GROSS Maurice (1985). « Sur les déterminants dans les expressions figées », *Langages* 79, pp. 89-117.

-
- GROSS Maurice (1986 [1977]). *Grammaire transformationnelle du français. Syntaxe du nom*, Paris, Larousse.
- GROSS Maurice (1989). « La construction de dictionnaires électroniques », *Annales des télécommunications* 44 (1-2), pp. 4-19.
- GROSS Maurice (1995). « Representation of Finite Utterances and the Automatic Parsing of Texts », *Language Research* 31 (2), Seoul, Language Research Institute, pp. 291-307.
- GROSS Maurice (1997a). « The Construction of Local Grammars », E. Roche et Y. Schabes (Éds.), *Finite-State Language Processing*, Cambridge, Mass.-London, MIT Press, pp. 329-354.
- GROSS Maurice (1997b). *Syntaxe générale et Syntaxe française. Les notions de base de la syntaxe. Notes de cours d'informatique linguistique*, syllabus, 2^e semestre 1997.
- GROSS Maurice (1999). « Lemmatization of compound tenses in English », C. Fairon (Éd.), *Analyse lexicale et syntaxique : le système INTEX, Lingvisticæ Investigationes* 22 (1-2), Amsterdam-Philadelphia, John Benjamins, pp. 71-122.
- GROSS Maurice (2001). « Les ambiguïtés », (fascicule spécial *Description et levée des ambiguïtés*, É. Laporte (Éd.), *Lingvisticæ Investigationes* 24 (1), Amsterdam/Philadelphia, John Benjamins, pp. 3-41.
- GUAÏTELLA Isabelle (1991). *Rythme et Parole : comparaison critique du rythme de la lecture oralisée et de la parole spontanée*, Aix-en-Provence, Université de Provence, Thèse non publiée.
- GUÉNOT Marie-Laure (2002). *Vers une grammaire électronique du français parlé : Prolégomènes*, Université de Provence, Mémoire inédit de DEA.
- GUÉNOT Marie-Laure (2005). « Parsing de l'oral : traiter les disfluences », *Actes de TALN 2005* (6-10 juin, Dourdan).
- GUÉNOT Marie-Laure (2006). « La coordination considérée comme un entassement paradigmatique : description, représentation et intégration », P. Mertens, C. Fairon, A. Dister et P. Watrin (Éds.), *Verbum ex machina, Actes de TALN 2006, Cahiers du Cental* 2 (2), Presses universitaires de Louvain, Louvain-la-Neuve.
- GUILLET Alain (1995). « Les verbes français : formalisation des différences », J. Labelle et Ch. Leclère (Éds.), *Lexiques-grammaires comparés en français. Actes du colloque international de Montréal, Lingvisticæ Investigationes Supplementa* 17, Amsterdam-Philadelphia, John Benjamins, pp. 59-70.
- GUIRAUD Pierre (1954). *Les Caractères statistiques du vocabulaire*, Paris, Presses universitaires de France.

-
- GUIRAUD Pierre (1960). *Problèmes et Méthodes de la statistique linguistique*, Paris, Presses universitaires de France.
- GUMPERZ John J., BERENZ Norine (1993). « Transcribing Conversational Exchanges », J.A. Edwards and M.D. Lampert (Éds.), *Talking Data. Transcription and Coding in Discourse Research*, Hillsdale, Lawrence Erlbaum Associates, pp. 91-120.
- HABERT Benoît (1999). « Un corpus clé pour le français actuel », Présentation du projet Corpus CLEF [en ligne], disponible sur : <http://www.biomath.jussieu.fr/CLEF>.
- HABERT Benoît (2000). « Des corpus représentatifs : de quoi, pour quoi, comment ? », M. Bilger (Éd.), *Linguistique sur corpus. Études et réflexions*, Cahiers de l'Université de Perpignan 31, pp. 11-58.
- HABERT Benoît (2005a). « Face à la disette dans la profusion », *SCOLIA* 19, pp. 41-61.
- HABERT Benoît (2005b). *Instruments et ressources électroniques pour le français*, Paris, Ophrys.
- HABERT Benoît, SALEM André (1995). « L'utilisation de catégorisations multiples pour l'analyse quantitative », *TAL* 36 (1-2), pp. 249-276.
- HABERT Benoît, NAZARENKO Adeline, SALEM André (1997). *Les Linguistiques de corpus*, Paris, Armand Colin.
- HABERT Benoît, FABRE Cécile, ISAAC Fabrice (1998). *De l'écrit au numérique : constituer, normaliser et exploiter les corpus électroniques*, Paris, InterÉditions.
- HALLIDAY Michael A.K. (1985). *Spoken and Written Language*, Oxford, Oxford University Press.
- HAMBYE Philippe (2005). *La Prononciation du français contemporain en Belgique. Variation, normes et identités*, Université catholique de Louvain, Thèse non publiée.
- HANSE Joseph, DOPPAGNE Albert, BOURGEOIS-GIELEN Hélène (1971). *Chasse aux belgicisms*, Bruxelles, Office du bon langage (Fondation Plisnier).
- HANSE Joseph, DOPPAGNE Albert, BOURGEOIS-GIELEN Hélène (1974). *Nouvelle chasse aux belgicisms*. Bruxelles, Office du bon langage (Fondation Plisnier).
- HARRIS Zellig H. (1969). « Analyse du discours », J. Dubois et J. Stumpf (Éds.) *L'Analyse du discours*, *Langages* 13, pp. 8-45.

-
- HENRY Sandrine (2001). « Étude des répétitions en français parlé spontané pour les technologies de la parole », *Actes de RÉCITAL* (Nancy, 24-27 juin 2001).
- HENRY Sandrine (2005). « Quelles répétitions à l'oral ? Esquisse d'une typologie », G. Williams (Éd.), *La Linguistique de corpus*, Rennes, Presses universitaires de Rennes, pp. 81-92.
- HENRY Sandrine, CAMPIONE Estelle, VÉRONIS Jean (2004). « Répétitions et pauses (silencieuses et remplies) en français spontané », *Actes des 25^{es} Journées d'Étude sur la Parole (JEP'2004)* (19-22 avril, Fès), pp. 261-264.
- HENRY Sandrine, PALLAUD Berthille (2003). « Word fragments and repeats in spontaneous spoken French », R. Eklund (Éd.), *Proceedings of DISS'03. Disfluency in Spontaneous Speech Workshop, (5-8 Septembre 2003, Göteborg University, Sweden), Gothenburg Papers in Theoretical Linguistics* 90, pp. 77-80.
- HINDLE Donald (1983). « Deterministic parsing of syntactic non-fluencies », *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics*, Cambridge, Massachusetts, pp. 123-128.
- HOUBEINE Anne-Marie (1979). « Pour qui, pourquoi et comment transcrire ? », *Le Français dans le monde* 145, pp. 28-35.
- IDE Nancy, VÉRONIS Jean (1994). « MULTTEXT : Multilingual Text Tools and Corpora », *Proceedings of the 15th International Conference on Computational Linguistics (COLING'94)* (August 5-9, Kyoto, Japan).
- JACOBSON Michel (2002). « Les outils modernes pour la transcription de corpus de parole », *Revue Parole* 22-23-24, pp. 213-229.
- JEANJEAN Colette (1984). « Les ratés c'est ça fabuleux. Étude syntaxique des amorces et des répétitions », *LINX* 10, pp. 171-177.
- JEFFERSON Gail (1974). « Error correction as an interactional resource », *Language in Society* 2, pp. 181-199.
- JOHANSSON Stig (1995). « The approach of the Text Encoding Initiative to the encoding of spoken discourse », G. Leech, G. Myers, J. Thomas (Éds.), *Spoken English on Computer. Transcription, Mark-up and Application*, New York, Longman, pp. 82-98.
- JOHANSSON Stig, LEECH Geoffrey, GOODLUCK Helen (1978). *Manual of information to accompany the lancaster-oslo/bergen corpus of british english, for use with digital computers*. <http://khnt.hit.uib.no/icame/manuals/lob/INDEX.HTM>.
- KASTBERG Sjöblom Margareta (2004). « Analyse grammatico-métrique d'une monographie "multi-générique" : le substantif », G. Purnelle,

-
- C. Fairon et A. Dister (Éds.), *Le Poids des mots. Actes des 7^{es} Journées internationales d'Analyse statistique des Données textuelles*, Louvain-la-Neuve, Presses universitaires de Louvain, pp. 672-681.
- KASTBERG Sjöblom Margareta (2006). « Ce que l'analyse lexicométrique d'un dictionnaire français suédois nous enseigne sur le discours dictionnaire », *Actes des JADT 2006, 8^{es} Journées Internationales d'Analyse des Données Textuelles* (19-21 avril, Besançon) pp. 559-570.
- KENNEDY Graeme (1998). *An Introduction to Corpus Linguistics*, London-New York, Longman.
- KERLEROUX Françoise (1996). *La Coupure invisible. Études de syntaxe et de morphologie*. Villeneuve d'Ascq, Presses universitaires du Septentrion.
- KEVERS Laurent, GILLES Peter, SIMON Anne Catherine (2006). « [moca], un système de gestion et d'annotation de données orales », communication présentée à la 3^e rencontre fribourgeoise de la linguistique sur corpus appliquée aux langues romanes (14-17 septembre 2006, Freiburg-im-Breisgau).
- KIDA Tsuyoshi (2001). « À propos du dynamisme dans le système de prise de parole », *Travaux Interdisciplinaires de Parole et langage d'Aix-en-Provence* 19, pp. 73-89.
- KILGARRIFF Adam, GREFENSTETTE Gregory (2003). « Introduction to the Special Issue on the Web as a corpus », *Computational Linguistics* 29 (3), pp. 333-347.
- KLEIN Jean-René, LAMIROY Béatrice (1994). « Lexique-grammaire du français de Belgique : les expressions figées », *Lingvisticæ Investigationes* 18 (2), Amsterdam-Philadelphia, John Benjamins, pp. 285-320.
- KLEIN Jean-René, LAMIROY Béatrice (1995a). « Les expressions figées du français de Belgique », J. Labelle et Ch. Leclère (Éds.), *Lexiques-grammaires comparés en français. Actes du colloque international de Montréal, Lingvisticæ Investigationes Supplementa* 17, Amsterdam-Philadelphia, John Benjamins, pp. 37-52.
- KLEIN Jean-René, LAMIROY Béatrice (1995b). « “Quand il éclaire, les vaches bisent”. Lexique-grammaire des verbes simples du français de Belgique », J. Labelle (Éd.) *Lexiques-grammaires comparés et traitements automatiques*, Montréal, Université du Québec à Montréal, pp. 257-267.
- KLEIN Jean-René, LENOBLE-PINSON Michèle (1997). « Lexique », D. Blampain, A. Goosse, J.-M. Klinkenberg et M. Wilmet (Dir.), *Le Français en Belgique*, Louvain-la-Neuve, Duculot, pp. 187-206.

-
- KOSKENNIEMI Kimmo (1990), « Finite-State parsing and disambiguation », *Proceedings of COLING '90* (August 20-25, Helsinki) pp. 229-232.
- KOZIMA Hideki, FURUGORI Teiji (1993). « Similarity between words computed by spreading activation on an English Dictionary », *Proceedings of EACL-93* (April 21-23, Utrecht), pp. 232-239.
- KUCERA Henry, FRANCIS W. Nelson (1967). *Computational Analysis of Present-day American English*, Providence, Brown University Press.
- LABBÉ Dominique (1990). *Le Vocabulaire de François Mitterrand*, Paris, Presses de la Fondation nationale des Sciences Politiques.
- LABOV William (1973). « Some principles of linguistic methodology », *Language in Society* 1, pp. 97-120.
- LABOV William (1977). *Sociolinguistique*, Paris, Éditions de Minuit.
- LAFOREST Marty (1992). *Le Back-channel en situation d'entrevue*, Québec, CIRAL (coll. « Recherches sociolinguistiques »).
- LAMIROY Béatrice (1998). « Le lexique-grammaire. Essai de synthèse », *Travaux de linguistique* 37, pp. 7-23.
- LAMIROY Béatrice, LECLÈRE Christian, KLEIN Jean-René, LABELLE Jacques (2003). « Les expressions verbales figées dans quatre variétés de français. Le projet "BFQS" », *Cahiers de lexicologie* 83 (2), pp. 153-172.
- LAPORTE Éric (1988). « La reconnaissance des expressions figées lors de l'analyse automatique », *Langages* 90, pp. 117-126.
- LAPORTE Éric (1994a), « Experiment in Lexical Disambiguation Using Local Grammars », F. Kiefer, G. Kiss et J. Pajzs (Éds.), *Papers in Computational Lexicography. COMPLEX '94*, Budapest, Research Institute for Linguistics of the Hungarian Academy of Sciences, pp. 163-172.
- LAPORTE Éric (1994b). « Levée d'ambiguïtés par grammaires locales », J. Labelle et Ch. Leclère (Éds.), *Lexiques-grammaires comparés en français. Actes du colloque international de Montréal, Lingvisticæ Investigationes Supplementa* 17, Amsterdam-Philadelphia, John Benjamins, pp. 97-114.
- LAPORTE Éric (1996). « L'évaluation des résultats de la levée d'ambiguïtés lexicales », *Lexique, syntaxe et analyse automatique des textes. Hommage à Jean Dubois, LINX* 34-35, pp. 291-305.
- LAPORTE Éric (2000). « Mots et niveau lexical », J.-M. Pierrel (Éd.), *Ingénierie des langues*, Paris, Hermès, pp. 25-49.

-
- LAPORTE Éric, MONCEAUX Anne (1999). « Elimination of lexical ambiguities by grammars : the ELAG system », C. Fairon (Éd.), *Analyse lexicale et syntaxique : le système INTEX*, *Linguisticae Investigationes* 22 (1-2), Amsterdam-Philadelphia, John Benjamins, pp. 341-367.
- LAPORTE Éric, SILBERZTEIN Max (1996). « Ambiguity rates. Automatic analysis of French text corpora and computation of ambiguity rates for different tagsets », É. Laporte (Éd.), *GRAMLEX deliverables, October 1995 – June 1996*, Paris, LADL.
- LARIVIÈRE Louise-Laurence (2004). « Nature de l'incluant dans la définition des noms communs de personnes comportant la marque grammaticale n.(nom indifférencié en genre) dans les dictionnaires », communication présentée au Congrès de l'ACFAS, Montréal.
- LECOMTE Josette (1998). *Le Catégoriseur BRILL14-JL5/WinBrill-0.3.*, Nancy, INaLF-CNRS. http://www.inalf.cnrs.fr/winbrill/BRILL14-JL5_WinBrill.doc.
- LEECH Geoffrey (1997a). « A brief users' guide to the grammatical tagging of the British National Corpus ». <http://www.natcorp.ox.ac.uk/what/gramtag.html>.
- LEECH Geoffrey (1997b). « Introduction to corpus annotation », R. Garside, G. Leech, T. Mc Enery (Éds.), *Corpus annotation : Linguistic Information from Computer Text Corpora*, White Plains (NY)-London, Longman, pp. 1-18.
- LEECH Geoffrey, GARSIDE Roger, BRYANT Michael (1994). « CLAWS4 : The tagging of the British National Corpus », *Proceedings of the 15th International Conference on Computational Linguistics (COLING'94)* (August 5-9, Kyoto, Japan), pp. 622-628.
- LEECH Geoffrey, MYERS Greg, THOMAS Jenny (Éds.) (1995). *Spoken English on computer. Transcription, Mark-up and Application*, New York, Longman.
- LEECH Geoffrey, WILSON Andrew (1999). « Standards for tagsets », H. Van Halteren (Éd.), *Syntactic Wordclass Tagging*, Dordrecht, Kluwer Academic Publishers, pp. 55-80.
- LEEMAN Danielle (2002). *La Phrase complexe. Les subordonnées*, Bruxelles, Duculot.
- LEFEUVRE Florence (1999). *La Phrase averbale en français*, Paris-Montréal, L'Harmattan.
- LÉON Pierre R. (1968). « Aspects phonostylistiques des niveaux de langue », *La Grammaire du français parlé, Le Français dans le Monde* 47, juin 1968, pp. 68-72.

-
- LEROY Christine (1985). « La notation de l'oral », M.-A. Morel (Dir.), *L'Oral du débat, Langue française* 65, pp. 6-16.
- LEROY Sarah (2005). « D'un torrent de larmes à un Himalaya de bêtises. Sur certains déterminants nominaux métaphoriques en français », *Travaux de linguistique* 50 (1), pp. 97-112.
- LEVELT Willem J.M. (1989). *Speaking : from intention to articulation*. Cambridge, MIT Press.
- LICKLEY Robin (2001). « Dialogue Moves and Disfluency Rates », *Proceedings of DISS'01* (August 29-31, Edinburgh), pp. 93-96.
- LLISTERRI Joaquim (1996). *EAGLES. Preliminary recommendations on spoken texts*. <http://www.ilc.cnr.it/EAGLES96/spokentx/spokentx.html>.
- LOSIER Line, KASPARIAN Sylvia, CHEVALIER Gisèle, GAUVIN Karine (2002). *Guide de présentation de mémoires et de thèses en linguistique et conventions pour la transcription de conversations*, Université de Moncton.
- LOUFRANI Claude (1981). « Le Locuteur collectif. Typologie des configurations discursives », *Recherches sur le français parlé* 6, Université de Provence, pp. 169-193.
- LUZZATI Daniel (1985). « Recherches sur la structure du discours oral spontané », *L'Information grammaticale* 27, pp. 39-41.
- LUZZATI Françoise, LUZZATI Daniel (1986a). « Oral et familier », *L'Information grammaticale* 28, pp. 5-10.
- LUZZATI Françoise, LUZZATI Daniel (1986b). « Oral et familier, du lexique à l'organisation du discours », *L'Information grammaticale* 30, pp. 23-28.
- MARCUS Mitchell P., SANTORINI Beatrice, MARCINKIEWICZ Mary Ann (1993). « Building a large annotated corpus of English : the Penn Treebank », *Computational Linguistics* 19, pp. 313-330.
- MAROUZEAU Jacques (1950 [1921]). *La Linguistique ou science du langage*, Paris, Librairie orientaliste Paul Geuthner.
- MARTINET André (1996). *Éléments de linguistique générale*, Paris, Masson & Armand Collin (4^e édition).
- MARTINIE Bruno (2001). « Remarques sur la syntaxe des énoncés réparés en français parlé », *Recherches sur le français parlé* 16, Université de Provence, pp. 189-206.
- MATHIEU-COLAS Michel (1995). « Syntaxe du trait d'union : structures complexes », *Linguisticæ Investigationes* 19 (1), Amsterdam-Philadelphia, John Benjamins, pp. 153-171.

-
- MAUREL Denis (1990). « Adverbes de date, étude préliminaire à leur traitement automatique », *Linguisticæ Investigationes* 14 (1), Amsterdam-Philadelphia, John Benjamins, pp. 31-63.
- MCENERY Tony, WILSON Andrew (2001 [1996]). *Corpus Linguistics*, Edinburgh, Edinburgh University Press.
- MCKELVIE David (1998). *SDP - Spoken Dialogue Parser*. HCRC Technical Report H C R C / R P - 9 6 , M a y 1 9 9 8 .
<http://www.hcrc.ed.ac.uk/publications/tr-96.ps>.
- MEJRI Salah (2000). « Syntaxe et figement », *Lexique, syntaxe et sémantique* (Mélanges offerts à Gaston Gross), *BULAG* n° hors série, pp. 333-342.
- MEJRI Salah (2002). « Le figement lexical : nouvelles tendances ». *Cahiers de lexicologie* 80, pp. 213-225.
- MEJRI Salah (2003). « Le figement lexical », *Cahiers de lexicologie* 82, 23-39.
- MELIS Ludo (2000). « Le français parlé et le français écrit, une opposition à géométrie variable », *Romanesque*, 2000 (3).
- MENYHÁRT Krisztina (2003). « Age-dependent types and frequency of disfluencies », R. Eklund (Éd.), *Proceedings of DISS'03. Disfluency in Spontaneous Speech Workshop, (5-8 Septembre 2003, Göteborg University, Sweden), Gothenburg Papers in Theoretical Linguistics* 90, pp. 45-48.
- MERIALDO Bernard (1995), « Méthodes probabilistes et étiquetage automatique », *TAL* 36, pp. 7-22.
- MERTENS Piet (2002). « Les corpus de français parlés ELICOP : consultation et exploitation », J. Binon, P. Desmet, J. Elen, P. Mertens, L. Sercu (Éds.), *Tableaux Vivants. Opstellen over taal-en-onderwijs aangeboden aan Mark Debrock*, Leuven, Universitaire Pers.
- METEER Marie *et al.* (1995). *Dysfluency Annotation Stylebook for the Switchboard Corpus*. <http://www.cis.upenn.edu/~treebank/>.
- MONDADA Lorenza (2000). « Les effets théoriques des pratiques de transcription », *LINX* 42, revue de l'Université de Paris X-Nanterre, pp. 131-150.
- MOON Rosamund (1998). *Fixed Expressions and Idioms in English : a Corpus-based Approach*, Oxford, Clarendon press.
- MOREAU Marie-Louise (1971). « L'homme que je crois qui est venu. *Qui, que* : relatifs et conjonctifs », M. Gross et J. Stéfanini (Éds.), *Syntaxe transformationnelle du français*, *Langue française* 11, pp. 77-90.

-
- MOREAU Marie-Louise (1997). « Le bon français de Belgique. D'un divorce entre norme et discours sur la norme », D. Blampain, A. Goosse, J.-M. Klinkenberg et M. Wilmet (Dir.), *Le Français en Belgique*, Louvain-la-Neuve, Duculot, pp. 391-399.
- MOREAU Marie-Louise, BRICHARD Huguette, DUPAL Claude (1999). *Les Belges et la Norme. Analyse d'un complexe linguistique*, Bruxelles-Louvain-la-Neuve, Service de la Langue française-Duculot (coll. « Français & Société », n° 9).
- MOREL Marie-Annick (1985). « Présentation », M.-A. Morel (Dir.), *L'Oral du débat*, *Langue française* 65, pp. 3-5.
- MOTA Cristina (1999). « Enhancing the INTEX Morphological Parser with Lexical Constraints », *Linguisticae Investigationes* 22, Amsterdam-Philadelphia, John Benjamins, pp. 413-423.
- MULLER Charles (1977). *Principes et Méthodes de la statistique lexicale*, Paris, Hachette.
- MULLER Charles (1979). « Le MOT, unité de texte et unité de lexique en statistique lexicologique », *Langue française et Linguistique quantitative. Recueil d'articles*, Genève, Slatkine, pp. 125-143.
- MULLER Charles (1992 [1973]). *Initiation aux méthodes de la statistique linguistique*, Paris, Champion.
- NAETS Hubert (2001). *Rapport de stage* [sur le codage XML des corpus de Valibel], DEC2 en ingénierie linguistique, Université catholique de Louvain.
- NAKAMURA Takuya (2003), « Analysing texts in a specific domain with local grammars : the case of stock exchange market reports », Y. Kawaguchi et alii (Eds.), *Proceedings of the First International Conference on Linguistic Informatics*, UBLI 1, Tokyo University of Foreign Studies, Amsterdam-Philadelphia, John Benjamins.
- NASR Alexis, VOLANSCHI Alexandra (2004). « Couplage d'un étiqueteur morpho-syntaxique et d'un analyseur partiel représentés sous la forme d'automates finis pondérés », *Actes de TALN 2004* (19-22 avril, Fès).
- NELSON Gerald (1996a). « The Design of the Corpus », S. Greenbaum (Éd.) *Comparing English worldwide. The International Corpus of English*, Oxford, Clarendon Press, pp. 27-35.
- NELSON Gerald (1996b). « Markup systems », S. Greenbaum (Éd.) *Comparing English worldwide. The International Corpus of English*, Oxford, Clarendon Press, pp. 36-53.
- NIVRE Joakim, GRÖNQVIST Leif (2001). « Tagging a Corpus of Spoken Swedish », *International Journal of Corpus Linguistics* 6 (1), pp. 47-78.

-
- OCHS Elinor (1979). « Transcription as theory », E. Ochs et B. B. Schieffelin (Éds.), *Developmental pragmatics*, New York-San Francisco-London, Academic Press, pp. 43-72.
- OOSTDIJK Nelleke (2003). « Normalization and disfluencies in spoken language data », S. Granger et St. Petch-Tyson (Éds.), *Extending the scope of corpus-based research. New applications, new challenges*, Amsterdam-New York, Rodopi, pp. 59-70.
- ÖSTMAN Margareta (1998). « Quatre (un ?) corpus – Quatre genres. Quelques résultats », *Actes des JADT'98, 4^{es} Journées Internationales d'Analyse des Données Textuelles* (19-21 février, Nice), pp. 487-494.
- PALLAUD Berthille (1999). « Lapsus et phénomènes voisins dans la langue parlée. Problèmes d'identification », *Recherches sur le français parlé* 15, Université de Provence, pp. 1-33.
- PALLAUD Berthille (2002a). « Erreurs d'écoute dans la transcription de données orales », *Revue Parole* 22-23-24, pp. 267-294.
- PALLAUD Berthille (2002b). « Les amorces de mots comme faits autonymiques en langage oral », *Recherches sur le français parlé* 17, Université de Provence, pp. 79-101.
- PALLAUD Berthille (2003). « Achoppements dans les énoncés de français oral et sujets syntaxiques », J.-M. Merle (Éd.), *Le Sujet*, Paris, Ophrys.
- PALLAUD Berthille (2004). « La transgression et la variation », *Marges Linguistiques* 8, pp. 76-87.
- PALLAUD Berthille, HENRY Sandrine (2004a). « Amorces de mots et répétitions dans les énoncés oraux », *Recherches sur le français parlé* 18, pp. 201-229.
- PALLAUD Berthille, HENRY Sandrine (2004b). « Amorces de mots et répétitions : des hésitations plus que des erreurs en français parlé », G. Purnelle, C. Fairon et A. Dister (Éds.), *Le Poids des mots. Actes des 7^{es} Journées internationales d'Analyse statistique des Données textuelles*, Louvain-la-Neuve, Presses universitaires de Louvain, pp. 849-858.
- PAROUBEK Patrick (2005). « Action Transversale COPTE : Corpus Parole Texte Évaluation », *Rapport d'activité*, LIMSI, pp. 151-156.
- PAROUBEK Patrick, ADDA Gilles, MARIANI Joseph, RAJMAN Martin (1997). « Les procédures de mesure automatique de l'action GRACE pour l'évaluation des assignateurs de parties du discours pour le français », *Actes des Journées Scientifiques et Techniques du Réseau Francophone de l'Ingénierie de la Langue de l'AUFELF-UREF*, Avignon, pp. 245-252.

-
- PAROUBEK Patrick, RAJMAN Martin (2000). « Étiquetage morpho-syntaxique », J.-M. Pierrel (Éd.), *Ingénierie des langues*, Paris, Hermès, pp. 131-150.
- PAUMIER Sébastien (2003), *De la reconnaissance de formes linguistiques à l'analyse syntaxique*, Université de Marne-la-Vallée, Thèse non publiée.
- PAUMIER Sébastien (2006). *Unitex 1.2. Manuel d'utilisation*. <http://www-igm.univ-mlv.fr/~unitex/manuel.html>.
- PAUMIER Sébastien, ULLAND Harald (2005). « Analyse automatique de mots polylexicaux en norvégien », *Lingvisticæ Investigationes* 28 (2), Amsterdam-Philadelphia, John Benjamins.
- PAYNE Jonathan (1995). « The COBUILD spoken corpus : transcription conventions », G. Leech, G. Myers, J. Thomas (Éds.), *Spoken English on Computer. Transcription, Mark-up and Application*, New York, Longman, pp. 203-207.
- PEYTARD Jean (1968). « Pour une typologie des messages oraux », *La Grammaire du français parlé, Le français dans le Monde* 47, juin 1968, pp. 73-80.
- PINCEMIN Bénédicte (2004). « Lexicométrie sur corpus étiquetés », G. Purnelle, C. Fairon et A. Dister (Éds.), *Le Poids des mots. Actes des 7^{es} Journées internationales d'Analyse statistique des Données textuelles*, Louvain-la-Neuve, Presses universitaires de Louvain, pp. 865-873.
- PIRON Maurice (1973). « Les belgicisms lexicaux : essai d'un inventaire », R. Martin (Dir.), *Mélanges de linguistique française et de philologie et littérature médiévales offerts à monsieur Paul Imbs à l'occasion de son 65^e anniversaire, le 4 mai 1973 par ses collègues, ses élèves et ses amis*, Paris, Klincksieck, pp. 295-304.
- PIRON Maurice (1979). « Le français de Belgique », Al. Valdman (Dir.), *Le Français hors de France*, Paris, Champion, pp. 201-221.
- PLOUX Sabine (2005). « Qu'est-ce que ça veut dire ? Un modèle interprétatif pour la composition sémantique », *Working papers*, Institut des sciences cognitives, Université Claude Bernard Lyon-1.
- POHL Jacques (1984). « Le statalisme », *Travaux de linguistique et de littérature* 22 (1), pp. 251-264.
- POIBEAU Thierry (2003). *Extraction automatique d'information. Du texte brut au web sémantique*, Paris, Hermès.
- PURNELLE Gérard (1998). « Théorie et typographie : une synthèse des règles typographiques de la ponctuation », J.-M. Defays, L. Rosier et Fr. Tilkin (Éds.), *À qui appartient la ponctuation ?*, Actes du colloque

-
- international et interdisciplinaire de Liège (13-15 mars 1997)*, Louvain-la-Neuve, De Boeck & Larcier (coll. « Champs linguistiques »), pp. 211-221.
- RAHMAN Anna, SAMPSON Geoffrey (2000). « Extending Grammar Annotation Standards to Spontaneous Speech », *Language and Computers : studies in practical linguistics* 30, pp. 295-311.
- RAINGEARD Martine, LORSCHIEDER Ute (1977). « Édition d'un corpus de français parlé », *Recherches sur le français parlé* 1, Université de Provence, pp. 14-29.
- RAJMAN Martin (1996). « Format de description lexicale pour le français. Partie 1 : Concepts généraux », Réf. GRACE GTR-2-1.2, document non publié.
- RAJMAN Martin, LECOMTE Josette, PAROUBEK Patrick (1996). « Format de description lexicale pour le français. Partie 2 : Description morpho-syntaxique », Réf. GRACE GTR-3-2.1, document non publié.
- RANCHHOD Elisabete, MOTA Cristina, BAPTISTA Jorge (1999). « A Computational Lexicon of Portuguese for Automatic Text Parsing », *SIGLEX99 : Standardizing Lexical Resources, 37th Annual Meeting of the ACL* (June 20-26, College Park, Maryland, USA), pp. 74-80. <http://label2.ist.utl.pt/LabelEL/siglex99.ps>.
- RAYSON Paul, GARSIDE Roger (2000). « Comparing corpora using frequency profiling », *Proceedings of the workshop on Comparing Corpora, ACL 2000* (October 1-8, Hong Kong), pp. 1-6.
- RENOUF Antoinette (1992). « A Word in Time : first findings from the investigation of dynamic text », J. Aarts, P. de Haan et N. Oostdijk (Éds.), *English Corpora Design, Analysis and Exploitation : Papers from the 13th ICAME Conference, Nijmegen 1992*, Amsterdam, Rodopi.
- RENOUF Antoinette, KEHOE Andrew (2002). « WebCorp : Applying the Web to Linguistics and Linguistics to the Web », *World Wide Web 2002 Conference* (May 7-11, Honolulu, Hawaii).
- RIGAULT André (1968). « Introduction », *La Grammaire du français parlé, Le français dans le Monde* 47, juin 1968, pp. 6-8.
- RITTAUD-HUTINET Chantal (1986). « Corpus oraux et édition : quelle transcription pour quelle lecture ? », *TRANEL 11, Actes du 2^e colloque régional de linguistique* (2-3 octobre, Neuchâtel), pp. 197-221.
- RIVENC Paul (1971). « Lexique et langue parlée », A. Rigault (Dir.), *La Grammaire du français parlé*, Paris, Hachette, pp. 51-69.
- ROCHE Emmanuel (1992). « Text Disambiguation by Finite-State Automata. An Algorithm and Experiments on Corpora », *Proceedings of the 14th*

-
- International Conference on Computational Linguistics (COLING'92)* (20-28 juillet, Nantes).
- ROCHE Emmanuel, SCHABES Yves (1995). « Deterministic part of speech tagging with finite state transducers », *Computational Linguistics* 21 (2), pp. 227-253.
- ROCHE Emmanuel, SCHABES Yves (Éds.) (1997). *Finite-State Language Processing*, Cambridge, Mass.-London, MIT Press.
- ROMARY Laurent, SALMON-ALT Susanne, FRANCOPOULO Gil (2004). « Standards going concrete : from LMF to Morphalou », *Workshop on Electronic Dictionaries, COLING 2004* (August 23-27, Geneva).
- ROSIER Laurence (1998). « La ponctuation et ses acteurs », J.-M. Defays, L. Rosier et Fr. Tilkin (Éds.), *À qui appartient la ponctuation ?*, *Actes du colloque international et interdisciplinaire de Liège (13-15 mars 1997)*, Louvain-la-Neuve, De Boeck & Larcier (coll. « Champs linguistiques »), pp. 31-43.
- ROSSI Mario (1981). *L'Intonation : de l'acoustique à la sémantique*, Paris, Klincksieck.
- ROUBEAU Marie-Noëlle (2004a). « Du bon usage des amorces dans la transcription des corpus », *Recherches sur le français parlé* 18, Université de Provence, pp. 163-184.
- ROUBEAU Marie-Noëlle (2004b). « Du bon usage des notations phonétiques dans les corpus aphasiques », *Recherches sur le français parlé* 18, Université de Provence, pp. 241-254.
- ROULET Eddy (1994). « La phrase : unité de langue ou unité de discours ? », J. Cerquiglini-Toulet et O. Collet (Éds.), *Mélanges de philologie et de littérature médiévales offerts à Michel Burger*, Genève, Droz, pp. 101-110.
- ROULET Eddy (2002). « Le problème de la définition des unités à la frontière entre le syntaxique et le textuel », M. Charolles, P. Le Goffic et M.-A. Morel (Éds.), *Y a-t-il une syntaxe au-delà de la phrase ?*, *Verbum* 24 (1-2), pp. 161-178.
- RUWET Nicolas (1967). *Introduction à la grammaire générative*, Paris, Plon.
- SACKS Harvey, SCHEGLOFF Emanuel A., JEFFERSON Gail (1974). « A simplest systematics for the organization of turn-taking for conversation », *Language* 50 (4:1), pp. 696-735.
- SAGOT Benoît (2006). *Analyse automatique du français : lexiques, formalismes, analyseurs*, Université de Paris-7, Thèse non publiée.
- SALMON-ALT Susanne, DURAND Gaëlle (2003). « Représentation normalisée de corpus linguistiques : de l'étiquetage morpho-syntaxique

- et structurel à l'annotation anaphorique », 3^{es} *Journées de linguistique de corpus* (11-13 septembre, Lorient).
- SANTORINI Beatrice (1990). *Part-of-speech tagging guidelines for the Penn Treebank Project*. Technical report MS-CIS-90-47, Department of Computer and Information Science, University of Pennsylvania.
- SAUSSURE Ferdinand de (1966). *Cours de linguistique générale*, Paris, Payot.
- SAUVAGEOT Aurélien (1962). *Français écrit, français parlé*, Paris, Larousse.
- SAUVAGEOT Aurélien (1971). « L'articulation du discours », A. Rigault (Dir.), *La Grammaire du français parlé*, Paris, Hachette, pp. 137-147.
- SCHEGLOFF Emanuel A. (1981). « Discourse and conversation : *Uh huh* and other things that come between sentences », D. Tannen (Éd.), *Analyzing discourse : Text and talk*, Washington, Georgetown University Press, pp. 71-93.
- SCHEGLOFF Emanuel A., JEFFERSON Gail, SACKS Harvey (1977). « The preference for self-correction in the organization of repair in conversation », *Language* 53 (2), pp. 361-382.
- SCHMID Helmut (1994). « Part-of-Speech Tagging with Neural Networks », *Proceedings of the 15th International Conference on Computational Linguistics (COLING'94)* (August 5-9, Kyoto, Japan), pp. 172-176.
- SCHOENTJES Pierre (1998). « Ponctuer l'ironie », J.-M. Defays, L. Rosier et Fr. Tilkin (Éds.), *À qui appartient la ponctuation ?*, *Actes du colloque international et interdisciplinaire de Liège (13-15 mars 1997)*, Louvain-la-Neuve, De Boeck & Larcier (coll. « Champs linguistiques »), pp. 303-316.
- SENEILLART Pierre, BLONDEL Vincent (2003). « Automatic discovery of similar words », M. Berry (Éd.), *Survey of Text Mining. Clustering, classification and retrieval*, New York, Springer-Verlag, pp. 25-44.
- SHRIBERG Elizabeth (1994). *Preliminaries to a Theory of Speech Disfluencies*, Université de Berkeley, Thèse non publiée.
- SHRIBERG Elizabeth (2001). « To 'errr' is human : ecology and acoustics of speech disfluencies », *Journal of the International Phonetic Association* 31 (1), pp. 153-169.
- SILBERZTEIN Max (1990). « Le dictionnaire électronique des mots composés », *Dictionnaires électroniques du français, Langue française* 87, pp. 71-83.
- SILBERZTEIN Max (1993). *Dictionnaires électroniques et analyse automatique de textes. Le système INTEX*, Paris, Masson.

-
- SILBERZTEIN Max (1995). « Dictionnaires électroniques et comptage de mots », *Actes des 3^{es} Journées Internationales d'Analyse Statistique des Données Textuelles (JADT 1995)* (11-13 décembre, Rome).
- SILBERZTEIN Max (1996). « Expérience d'étiquetage avec levée d'ambiguïtés », *Actes des Premières Journées INTEX* (21-22 mars 1996, Paris), LADL, Université Paris-7, CNRS, pp. 17-31.
- SILBERZTEIN Max (1997). « The lexical analysis of natural languages », E. Roche et Y. Schabès (Éds.), *Finite-state language processing*, Cambridge, Mass.-London, MIT Press, pp. 175-203.
- SILBERZTEIN Max (1998a). « Transducteurs pour le traitement automatique des textes », *Travaux de linguistique* 37, Louvain-la-Neuve, Duculot, pp. 127-142.
- SILBERZTEIN Max (1998b). « Normalisation des textes », *Actes des 4^{es} Journées Internationales d'Analyse Statistique des Données Textuelles (JADT 1998)* (19-21 février, Nice), pp. 601-614.
- SIMON Anne Catherine (1999). « Une adaptation de l'approche modulaire au phénomène des ponctuations du discours », *Travaux neuchâtelois de linguistique* 30, pp. 7-28.
- SIMON Anne Catherine (2000). « Français de référence et segmentation du discours oral. l'exemple de la phrase », M. Francard, G. Geron et R. Wilmet (Éds.), *Le Français de référence. Constructions et appropriations d'un concept*, *Cahiers de l'Institut de linguistique de Louvain* 26 (1-4), pp. 345-358.
- SIMON Anne Catherine (2004). *La Structuration prosodique du discours en français. Une approche multidimensionnelle et expérimentelle*, Berne, Peter Lang.
- Simple Metadata Annotation Specification*, Version 6.2. (2004). Linguistic Data Consortium. www.ldc.upenn.edu/Projects/MDE.
- SINCLAIR John (1995). « From theory to practice », G. Leech, G. Myers, J. Thomas (Éds.), *Spoken English on Computer. Transcription, Mark-up and Application*, New York, Longman, pp. 99-109.
- SINCLAIR John (1996). *EAGLES, Preliminary Recommendations on Text Typology* EAG---TCWG---TTYP/P Version of June 1996.
- SINCLAIR John (1996). *Preliminary recommendations on Corpus Typology. Eagles document* EAG-TCWG-CTYP/P, May 1996. <http://www.ilc.cnr.it/EAGLES96/corpus/typ/corpus.html>.
- SINCLAIR John (1997 [1991]). *Corpus, concordance, collocation*, Oxford, Oxford University Press.

-
- STÉFANINI Jean (1981). « Sur la notion de phrase et son histoire », *Recherches sur le français parlé* 3, Université de Provence, pp. 7-18.
- STRASSEL Stephanie (2003). *Simple Metadata Annotation Specification*. Version 5.0.
- TANGUY Ludovic, HATHOUT Nabil (2002). « Webaffix est un outil d'acquisition morphologique dérivationnelle à partir du web », *Actes de TALN 2002* (24-27 juin, Nancy).
- TANNEN Deborah (1981). « Review of *Therapeutic discourse : Psychotherapy as conversation*, W. Labov and David Fanshel », *Language* 57 (2), pp. 481-486.
- TEI (2002) *Text Encoding Initiative*. <http://www.tei-c.org.uk/>.
- TESNIÈRE Lucien (1962). *Éléments de syntaxe structurale*, Paris, Klincksieck.
- THIBEAULT Mélanie (2004). *La Catégorisation grammaticale automatique : adaptation du catégoriseur de Brill au français et modification de l'approche*, Université Laval, Thèse de maîtrise non publiée.
- VALLI André, VÉRONIS Jean (1999). « Étiquetage grammatical des corpus de parole : problèmes et perspectives », *Revue française de linguistique appliquée* 4 (2), pp. 113-133.
- VAN HALTEREN Hans (Éd.) (1999). *Syntactic Wordclass Tagging*, Dordrecht, Kluwer Academic Publishers.
- VAN HALTEREN Hans (2000). « The Detection of Inconsistency in Manually Tagged Text », A. Abeillé, T. Brants, H. Uszkoreit (Éds.), *Proceedings of the 2nd Workshop on Linguistically Interpreted Corpora*, (August 6, Luxembourg).
- VANNIER Gérard, LACHERET-DUJOUR Anne, VERGNE Jacques (1999). « Pauses location and duration calculated with syntactic dependencies and textual considerations for TTS system », *Actes de ICPHS*. <http://users.info.unicaen.fr/~jvergne/ICPhS.pdf>.
- VASILESCU Ioana, CANDEA Maria, ADDA-DECKER Martine (2005). « Perceptual salience of language-specific acoustic differences in autonomous fillers across eight languages », *Actes de Interspeech 2005*, (September 4-8, Lisbonne).
- VÉDÉNINA Ludmila (1989). *Pertinence linguistique de la présentation typographique*, Paris, Peeters-Selaf.
- VENDRYÈS Joseph (1968 [1923]). *Le Langage. Introduction linguistique à l'histoire*, Paris, Albin Michel.

-
- VERGNE Jacques (1999). *Étude et modélisation de la syntaxe des langues à l'aide de l'ordinateur. Analyse syntaxique automatique non combinatoire. Synthèse et résultats*, Université de Caen, Habilitation à diriger des thèses.
- VERGNE Jacques (2001). « Analyse syntaxique automatique de langues : du combinatoire au calculatoire », *Actes de TALN 2001*, Tours.
- VERGNE Jacques, GIGUET Emmanuel (1998). « Regards théoriques sur le tagging », *Actes de TALN 1998* (10-12 juin, Paris).
- VÉRONIS Jean, KHOURI Liliane (1995). « Étiquetage grammatical multilingue : le projet MULTTEXT », *TAL* 36, pp. 233-248.
- VÉRONIS Jean (2000). « Annotation automatique de corpus : panorama et état de la technique », J.-M. Pierrel (Éd.), *Ingénierie des langues*, Paris, Hermès, pp. 111-129.
- VERSTRAETE Estelle (2000). « Digressions dans les interactions verbales entre élèves travaillant en groupe », *Recherches* 33, pp. 151-179.
- VIETRI Simonetta, ELIA Annibale (2000). « Electronic Dictionaries and Linguistic Analysis of Italian Large Corpora », *Actes des JADT 2000, 5^{es} Journées Internationales d'Analyse des Données Textuelles* (9-11 mars, Lausanne), pp. 179-188.
- VILNAT Anne, MONCEAUX Anne, PAROUBEK Patrick, ROBBA Isabelle, GENDNER Véronique, ILLOUZ Gabriel, JARDINO Michèle (2004). « Annoter en constituants pour évaluer des analyseurs syntaxiques », *Actes de TALN 2004* (19-22 avril, Fès).
- VINAY Jean-Paul (1971). « Les cadres de la phrase », A. Rigault (Dir.), *La Grammaire du français parlé*, Paris, Hachette, pp. 119-135.
- VINCENT Diane (1993). *Les Ponctuels de la langue et autres mots du discours*, Québec, Nuit Blanche Éditeur.
- VINCENT Diane, MARTEL Guylaine (2001). « Particules métadiscursives et autres modes langagières : des cas de changement linguistique », M. Matthey (Éd.), *Le Changement linguistique. Évolution, variation, hétérogénéité*, *Travaux neuchâtelois de linguistique* 34/35, pp. 141-152.
- VIOLLET Catherine (1986). « Interaction verbale et pratiques d'interruption », *Paroles inachevées. Études linguistiques pragmatiques et processuelles de conversations, brouillons et textes*, *DRLAV* 34-35, pp. 183-193.
- VOUTILAINEN Atro (1999). « A Short History of Tagging », H. Van Halteren (Éd.), *Syntactic Wordclass Tagging*, Dordrecht, Kluwer Academic Publishers, pp. 3-7.

-
- WATRIN Patrick (2002). *?C@T! Un outil d'analyse de corpus*, rapport de stage, Université de Louvain.
- WILLEMS Dominique (1998a). « La problématique des données et la place de l'exception en syntaxe contemporaine », *Langue française* 66, pp. 86-98.
- WILLEMS Dominique (1998b). « Données et théories en linguistique : réflexions sur une relation tumultueuse et changeante », M. Bilger, K. van den Eynde et Fr. Gadet (Éds.), *Analyse linguistique et approches de l'oral. Recueil d'études offert en hommage à Claire Blanche-Benveniste*, *Orbis Supplementa* 10, Leuven, Peeters, pp. 79-87.
- WILLEMS Dominique (1999). « Pour une terminologie grammaticale européenne. Défense et illustration », M.-J. Béguelin, J. Fr. de Pietro et A. Năf (Éds.), *La Terminologie grammaticale à l'école : perspectives interlinguistiques*, *TRANEL* 31, pp. 129-142.
- WILMET Marc (1997a). *Grammaire critique du français*, Paris, Hachette-Duculot.
- WILMET Marc (1997b). « Morphologie et syntaxe », D. Blampain, A. Goosse, J.-M. Klinkenberg et M. Wilmet (Dir.), *Le Français en Belgique*, Louvain-la-Neuve, Duculot, pp. 175-186.
- YAGUELLO Marina (Dir.) (1993). *Le Grand Livre de la langue française*, Paris, Seuil.
- ZELLNER Brigitte (1992). « Le bé bégayage et euh... l'hésitation en français spontané », *Actes des 19^{es} Journées d'Étude sur la Parole (JEP'92)* (19-22 juin, Bruxelles), pp. 481-487.
- ZELLNER Brigitte (1998). *Caractérisation et prédiction du débit de parole en français. Une étude de cas*, Université de Lausanne, Thèse non publiée.

TABLE DES FIGURES ET TABLEAUX

TABLEAU 1.	Conventions de transcription VALIBEL	59
FIGURE 1.	Représentation schématique des catégories de textes oraux	72
FIGURE 2.	Corpus VALIBEL. Effectifs selon le sexe	79
FIGURE 3.	Corpus VALIBEL. Effectifs selon l'âge	79
FIGURE 4.	Corpus VALIBEL. Effectifs selon le niveau d'études	80
FIGURE 5.	Carte des provinces de Belgique	82
FIGURE 6.	Corpus VALIBEL. Effectifs selon la localisation géographique	83
TABLEAU 2.	Tableau synoptique des textes qui composent le corpus	85
FIGURE 7.	Distribution des textes selon le nombre de mots	86
TABLEAU 3.	Nombre de locuteurs par enregistrement	87
TABLEAU 4.	Formes de notre corpus classées par fréquence décroissante	93
TABLEAU 5.	Formes de <i>Corpaix</i> classées par fréquence décroissante	94
FIGURE 8.	Modélisation de la séquence disfluente (Shriberg 1994 : 8)	102
TABLEAU 6.	Nombre de normalisations selon la catégorie de textes	105
TABLEAU 7. (1972)	Répartition des types de disfluences chez Grojean et Deschamps	108
TABLEAU 8.	Répartition des types de disfluences chez Candéa (2000a)	108
FIGURE 9.	Hergé, <i>Le Crabe aux pinces d'or</i> , p. 37, 7-9	119
TABLEAU 9.	Répartition selon le nombre de répétés	134
TABLEAU 10.	Répartition selon la longueur du répétable	136
TABLEAU 11. (valeur absolue)	Répartition selon la longueur du répétable et le nombre de répétés	137
TABLEAU 12. (pourcentage)	Répartition selon la longueur du répétable et le nombre de répétés	137
TABLEAU 13.	Tableau synthétique des hapax	141
TABLEAU 14.	Répartition des répétitions directes et associées selon le nombre de répétés	143
TABLEAU 15.	Insertions dans les répétitions	144
TABLEAU 16.	Insertions dans les répétitions associées	145
FIGURE 10.	Graphique de dispersion des marques insérées	146
TABLEAU 17.	Répartition du contexte des marques	151

TABLEAU 18.	Répartition des répétitions directes et associées en fonction des marques	151
TABLEAU 19.	Fréquence des marques dans les séquences mixtes	154
TABLEAU 20.	Catégorie grammaticale des répétables de longueur 1 (valeur absolue)	158
TABLEAU 21.	Catégorie grammaticale des répétables de longueur 1 (pourcentage)	159
TABLEAU 22.	Autocorrections du déterminant défini	171
TABLEAU 23.	Autocorrections des déterminants indéfinis	176
TABLEAU 24.	Autocorrections du pronom sujet	180
TABLEAU 25.	Répartition des types d'amorces chez Pallaud (2002b)	199
TABLEAU 26.	Catégorie grammaticale des amorces complétées	202
FIGURE 11.	Graphique de dispersion des amorces et des répétitions	203
TABLEAU 27.	Types d'amorces à l'intérieur d'une séquence répétition	207
TABLEAU 28.	Cooccurrence du <i>euh</i> et d'autres marques	217
TABLEAU 29.	Répartition du <i>euh</i>	228
TABLEAU 30.	Répartition du <i>euh</i> intrasyntagme et du <i>euh</i> à l'emplacement d'une rupture syntaxique	231
TABLEAU 31.	Répartition des trois degrés de pauses	259
TABLEAU 32.	Répartition des pauses dans le sous-corpus (valeur absolue)	265
TABLEAU 33.	Répartition des pauses dans le sous-corpus (pourcentage)	266
TABLEAU 34.	Types de marques contiguës à la pause	269
TABLEAU 35.	Contexte des marques combinées à la pause	271
TABLEAU 36.	Répartition des pauses intraconstituant et des pauses à l'emplacement d'une rupture	273
TABLEAU 37.	Distribution des pauses intraconstituant	278
FIGURE 12.	Exemple d'automate	298
FIGURE 13.	Étiquetage du <i>euh</i>	299
FIGURE 14.	Graphe <i>Phrase.grf</i>	301
FIGURE 15.	Graphe <i>alternanceCode.grf</i>	302
FIGURE 16.	Graphe <i>baliseIGN.grf</i>	303
FIGURE 17.	Automate idéal des multitranscriptions	304
FIGURE 18.	Graphe <i>Balises.grf</i>	305
FIGURE 19.	Graphe <i>Parentheses.grf</i>	307

FIGURE 20.	Graphe <i>NomsPropres.grf</i>	308
FIGURE 21.	Graphe <i>Locuteurs.grf</i>	309
TABLEAU 38.	Formes du pronom personnel (BU § 634)	335
TABLEAU 39.	Tableau synoptique du jeu d'étiquettes	353
TABLEAU 40.	Codes flexionnels	355
FIGURE 22.	Automate de <i>carte bleue</i>	356
FIGURE 23.	Graphe de flexion du futur	362
FIGURE 24.	Graphe de flexion du présent	364
FIGURE 25.	Graphe de flexion des verbes du 1 ^{er} groupe	365
TABLEAU 41.	Répartition des homographies chez Bolasco (1993)	383
FIGURE 26.	Automate du texte <i>la petite brise la glace</i>	391
FIGURE 27.	Automate du texte <i>la petite brise la glace</i>	391
FIGURE 28.	Automate explosé du texte <i>la petite brise la glace</i>	392
FIGURE 29.	Automate du texte <i>la petite brise la glace</i>	393
FIGURE 30.	Grammaire de désambiguïsation : la suite <PRO> <V>	396
FIGURE 31.	Automate de <i>la petite brise la glace</i> après une première désambiguïsation	397
FIGURE 32.	Grammaire de désambiguïsation : la suite <DET> <N>	397
FIGURE 33.	Automate de <i>la petite brise la glace</i> après une deuxième désambiguïsation	398
FIGURE 34.	Grammaire <i>en-N.grf</i>	398
FIGURE 35.	Grammaire de désambiguïsation : la suite <PRO> <V>, autre représentation	399
FIGURE 36.	Ambiguïté de <i>avocat</i> dans l'automate	401
FIGURE 37.	Automate non ambigu	402
FIGURE 38.	Automate de <i>parce que</i>	405
FIGURE 39.	Grammaire qui supprime les boîtes contenant des symboles XI	406
FIGURE 40.	Automate de <i>parce que</i> , avec <i>parce</i> mot inconnu	407
FIGURE 41.	Automate de <i>parce que</i> non ambigu	407
FIGURE 42.	Grammaire <i>s_elide.grf</i>	409
FIGURE 43.	Grammaire <i>s_elideBis.grf</i>	410
FIGURE 44.	Grammaire <i>DnumNom.grf</i>	411
FIGURE 45.	Grammaire <i>Tiret-ci-la.grf</i>	414

FIGURE 46.	Automate de <i>celui-là</i>	415
FIGURE 47.	Automate de <i>autres-là</i>	415
FIGURE 48.	Grammaire <i>PrepVAnne.grf</i>	416
FIGURE 49.	Automate de <i>à brûle-pourpoint</i>	416
FIGURE 50.	Automate de la séquence <i>à pitas</i>	417
FIGURE 51.	Grammaire <i>je.grf</i> de l'écrit	418
FIGURE 52.	Grammaire <i>je.grf</i> de l'oral	420
FIGURE 53.	Automate de <i>je n/ ne fais</i>	423
FIGURE 54.	Grammaire <i>ça marche</i>	425
FIGURE 55.	Grammaire <i>ça marche</i> généralisée aux catégories N et V	426
TABLEAU 42.	Résultats de la désambiguïsation selon le type d'étiquettes	432
FIGURE 56.	Automate de <i>le chose</i>	434
FIGURE 57.	Automate partiellement désambiguïisé de <i>le chose</i>	435

TABLE DES MATIÈRES

INTRODUCTION.....	7
CHAPITRE 1. LA TRANSCRIPTION DE DONNÉES ORALES.....	11
1 INTRODUCTION.....	12
2 DE LA NÉCESSITÉ D'ENREGISTRER	12
3 DE LA NÉCESSITÉ DE TRANSCRIRE.....	13
3.1 Le mythe du copiste.....	13
3.2 Le statut du texte transcrit.....	14
4 DIFFICULTÉS DE TRANSCRIRE : LES DIFFICULTÉS D'ÉCOUTE LIÉES À DES ASPECTS TECHNIQUES.....	16
5 « TRANSCRIPTION AS THEORY »	17
5.1 Segmenter et catégoriser.....	21
5.2 Que transcrire ?	21
5.3 Finesse de la transcription	23
5.4 La recherche du sens.....	24
5.4.1 Les homophonies.....	25
Illustration : les amorces de morphème	26
5.4.2 Les formes « sans signifié »	33
5.4.3 Les croyances du transcripateur.....	33
6 TRANSCRIPTION PHONÉTIQUE VS TRANSCRIPTION ORTHOGRAPHIQUE.....	33
6.1 Transcription phonétique	33
6.2 Transcription orthographique	36
6.2.1 Les trucages orthographiques.....	36
6.2.2 Les inconvénients des trucages.....	40
6.2.3 Les recommandations EAGLES et NERC.....	41
6.2.4 Cas particuliers : l'apocope et l'aphérèse.....	42
6.2.5 Les notations phonétiques qui accompagnent les transcriptions en orthographe standard.....	43
6.2.6 Ponctuation	44
6.2.7 L'unité de transcription	45
7 LA REPRÉSENTATION SPATIO-VISUELLE DES DONNÉES. LE TOUR DE PAROLE	45
7.1 Les chevauchements	49
7.2 Le chevauchement : deux autres représentations.....	50
7.3 Les continueurs	51
7.4 La transcription en partition	52
7.5 Les informations associées aux tours de parole.....	54
8 LE CORPUS COMME DONNÉES PARTAGÉES	55
9 LES CONVENTIONS DE TRANSCRIPTION DU CENTRE DE RECHERCHE VALIBEL	57
10 CONCLUSIONS	60
CHAPITRE 2. LE CORPUS.....	61
1 VOUS AVEZ DIT CORPUS ?	61
2 CORPUS AU SENS LARGE	62
3 CORPUS AU SENS RESTREINT.....	63
4 DIFFÉRENTS TYPES DE CORPUS	64
4.1 Le corpus clos	64

4.2	Le corpus ouvert.....	64
4.3	Le corpus dynamique.....	65
4.4	Le corpus de référence.....	66
4.4.1	Le Corpus de référence du français parlé.....	68
4.4.2	La représentativité.....	69
4.4.3	L'échantillonnage.....	69
4.4.4	La taille.....	70
4.4.5	L'équilibre.....	71
4.5	La banque de données VALIBEL.....	72
4.6	Constitution de notre corpus.....	74
4.6.1	Sous-corpus et enregistrements.....	75
4.6.2	Qualité sonore des enregistrements.....	76
4.6.3	Situation de parole.....	76
4.6.4	Les locuteurs.....	78
4.6.4.1	Le sexe.....	79
4.6.4.2	L'âge.....	79
4.6.4.3	Le degré de scolarité.....	80
4.6.4.4	La localisation géographique.....	81
4.6.5	Tableau synoptique des textes qui composent le corpus.....	85
4.6.6	Répartition des textes.....	86
4.6.6.1	Nombre de locuteurs par enregistrement.....	87
4.6.6.2	Le poids des locuteurs.....	87
4.6.6.3	Les tours de parole.....	89
4.7	Les formes du corpus.....	91
4.8	Conclusions.....	96
CHAPITRE 3. SECTION 1. LES « DISFLUENCES » DE L'ORAL.....		97
1	INTRODUCTION.....	97
2	MODÉLISATION DES DISFLUENCES.....	101
3	LES DISFLUENCES ET LE TAL.....	102
4	FRÉQUENCE DES DISFLUENCES.....	105
4.1	Selon le type d'oral et la situation de communication.....	105
4.2	Selon le type de disfluences.....	107
4.3	Selon les locuteurs.....	109
4.4	Quatre types de disfluences.....	110
4.5	Un exemple d'annotation de corpus oraux : le projet COPTE.....	111
CHAPITRE 3. SECTION 2. LES RÉPÉTITIONS.....		113
1	QUELS TYPES DE RÉPÉTITIONS ?.....	113
2	LES RÉPÉTITIONS « FAITS DE LANGUE ».....	116
2.1	La répétition exigée par la syntaxe.....	116
2.2	La répétition à effet oratoire.....	117
2.2.1	L'intensité.....	117
2.2.2	La gradation.....	119
2.2.3	La modification du sens.....	119
2.2.4	La mise en évidence.....	120
3	LES RÉPÉTITIONS « FAITS DE PAROLE ».....	121
4	UN PARTAGE DIFFICILE ENTRE DISFLUENCE ET FAIT DE LANGUE.....	122
5	LES RÉPÉTITIONS DANS LE CORPUS.....	125
5.1	Le repérage des répétitions dans le corpus.....	125
5.1.1	Méthode de repérage.....	125
5.1.2	Conséquences de la méthode.....	128

5.2	Résultats globaux : répartition des répétitions dans le corpus.....	130
5.2.1	Fréquence générale.....	130
5.2.2	La répartition par locuteur.....	131
5.2.3	Répétitions et parole superposée.....	133
5.2.4	Nombre de répétés (n).....	133
5.2.5	Longueur du répétable (l).....	136
5.2.6	Distribution des répétables.....	138
5.2.6.1	Formes les plus fréquentes.....	138
5.2.6.2	Les hapax.....	139
5.3	Forme de la séquence répétition : répétition directe, répétition associée, séquence mixte.....	141
5.3.1	Répétition directe vs répétition associée.....	142
5.3.2	Les insertions les plus fréquentes dans les répétitions associées.....	144
5.3.3	Séquence mixte.....	149
5.3.4	Le schéma des séquences mixtes.....	150
5.3.5	Les marques les plus fréquentes dans les séquences mixtes.....	153
5.4	Analyse morphosyntaxique des répétitions.....	155
5.4.1	Catégorie grammaticale des répétables.....	157
5.4.2	Le répétable est une séquence de mots grammaticaux.....	162
5.5	Continuité ou rupture syntaxique de l'énoncé.....	164
6	CONCLUSIONS.....	165
CHAPITRE 3. SECTION 3. L'AUTOCORRECTION IMMÉDIATE		167
1	DÉFINITION.....	167
2	LES DIFFÉRENTS TYPES D'AUTOCORRECTIONS RETENUS.....	170
2.1	Les déterminants.....	171
2.1.1	Les déterminants définis : le, la, les, l'.....	171
2.1.2	Les déterminants indéfinis.....	176
2.1.3	les déterminants possessifs.....	177
2.1.4	Les déterminants démonstratifs.....	178
2.1.5	Les changements de sous-catégorie de déterminants.....	178
2.2	Les pronoms sujets.....	179
2.3	Les pronoms relatifs.....	181
2.4	Les prépositions.....	183
3	EMPILEMENT DE DISFLUENCES : INTERACTION DE L'AUTOCORRECTION ET DE LA RÉPÉTITION.....	185
4	CONCLUSIONS.....	187
CHAPITRE 3. SECTION 4. L'AMORCE DE MORPHÈME		189
1	DÉFINITION.....	189
2	REPÉRAGE AUTOMATIQUE DES AMORCES.....	191
3	RÉSULTATS GLOBAUX : RÉPARTITION DES AMORCES DANS LE CORPUS.....	191
3.1	Fréquence générale.....	191
3.2	Répartition par locuteur.....	192
3.3	Catégories d'amorces.....	193
3.4	Typologie des amorces.....	194
3.4.1	Les amorces complétées.....	195
3.4.1.1	Le mot amorcé est complété sans reprise de la partie amorcée.....	195
3.4.1.2	Le mot amorcé est complété avec reprise de la partie déjà énoncée.....	195

3.4.1.3	Le mot amorcé est complété avec reprise de la partie amorcée mais aussi d'élément(s) antéposé(s) à l'amorce	196
3.4.2	Les amorces corrigées	197
3.4.3	Les amorces inachevées	198
3.5	Répartition des amorces en corpus	199
3.6	Les insertions entre <i>reparandum</i> et <i>repair</i>	200
3.7	La catégorie grammaticale des amorces	201
3.8	Amorces et répétitions	203
3.8.1	Les séquences d'amorces	204
3.8.2	L'amorce à l'intérieur d'une séquence répétition	206
4	CONCLUSIONS	208
CHAPITRE 3. SECTION 5. LE EUH DIT D'HÉSITATION		209
1	DÉFINITION	209
1.1	Le <i>euh</i> et l'allongement	210
1.2	Le <i>euh</i> et le [@] d'appui	211
1.3	<i>Euh</i> dit d'hésitation	211
2	LES EUH DANS NOTRE CORPUS	213
3	LE EUH COMBINÉ À UNE AUTRE MARQUE	214
3.1	Le <i>euh</i> et la répétition	215
3.1.1	Le <i>euh</i> dans une séquence répétition	215
3.1.2	Le <i>euh</i> objet de la répétition	216
3.2	Le <i>euh</i> et les autres marques : tableau général	216
3.3	Le <i>euh</i> et la pause silencieuse	217
3.4	Le <i>euh</i> et les ponctuants	218
3.5	Le <i>euh</i> « continuatif » : fin de tour de parole et chevauchement	219
3.5.1	Le <i>euh</i> en fin de tour	219
3.5.2	Le <i>euh</i> dans les chevauchements de parole	221
3.5.3	Le <i>euh</i> en début de tour de parole	221
3.6	Séquences de marques et <i>euh</i> seul	221
3.7	Place syntaxique du <i>euh</i>	224
4	CONCLUSIONS	233
CHAPITRE 4. LE PRÉTRAITEMENT DU CORPUS		239
1	INTRODUCTION	239
2	UNITEX ET LE PRÉTRAITEMENT	240
2.1	Normalisation de formes non ambiguës	241
2.1.1	Normalisation de formes ambiguës	241
2.2	Conséquences de la normalisation graphique	242
2.3	Le découpage en phrases	243
3	LA NOTION DE PHRASE	244
3.1	La phrase intuitive	244
3.2	La phrase intuitive : les représentations	245
3.3	La phrase linguistique	248
3.4	La phrase à l'écrit, la phrase à l'oral	249
3.5	Le découpage des productions orales	252
4	LA PAUSE SILENCIEUSE	253
4.1	La perception des pauses	256
4.2	Le degré des pauses	258
4.3	La répartition des pauses dans le corpus	259
4.4	Le silence	260

4.5	La pause longue et la pause brève	265
4.6	Étude sur un sous-corpus	265
4.6.1	Études antérieures	266
4.6.2	Combinaison de la pause et d'autres marques	269
4.6.3	Contexte des marques combinées	271
4.6.4	Pause marque de rupture et pause intraconstituant	272
4.6.5	La pause à l'emplacement d'une rupture syntaxique	273
4.6.6	La pause intraconstituant	276
4.7	Conclusions sur la pause	279
4.8	Le point d'interrogation	280
5	LE PRÉTRAITEMENT DU CORPUS AVANT UNITEX	282
5.1	La référencement des locuteurs et des textes	282
5.2	La normalisation des conventions de transcription	283
5.3	Chevauchements de parole	284
5.4	Les multitranscriptions	286
5.5	Les crochets	286
5.6	Les balises <s> et </s>	288
5.7	Le balisage des disfluences étudiées	289
5.7.1	Les répétitions	290
5.7.2	Les autocorrections	293
5.7.3	Les amorces	293
5.7.4	Les euh	295
5.8	Les parenthèses : éléments para-verbaux et passages incompréhensibles	296
5.9	Conclusions sur les marques	297
6	LE PRÉTRAITEMENT DU CORPUS AVEC UNITEX	297
6.1	Les graphes de prétraitement	297
6.2	Le graphe Sentence	300
6.2.1	Le graphe phrase	300
6.2.2	Le graphe baliseIGN	302
6.3	Le graphe LocateOral	306
6.3.1	Le graphe Parentheses	306
6.3.2	Le graphe NomsPropres	307
6.3.3	Le graphe Locuteurs	309
7	DU TEXTE « BRUT » AU TEXTE PRÉTRAITÉ	310
CHAPITRE 5. LES RESSOURCES LINGUISTIQUES		313
1	INTRODUCTION	313
2	Dictionnaires papier, dictionnaires sur support électronique ET dictionnaires électroniques	314
2.1	Systématisation de l'information donnée	315
2.2	Nombre d'entrées	316
3	LE DICTIONNAIRE ÉLECTRONIQUE DES FORMES FLÉCHIES DU FRANÇAIS : DELAF	317
3.1	Principes de présentation des données	317
3.2	Comparaison de nos ressources avec d'autres lexiques	322
3.2.1	Parallèles avec le DELAF	323
3.2.2	Divergences avec le DELAF	323
4	LA NORME LEXICOGRAPHIQUE : NOS CHOIX LINGUISTIQUES DE CATÉGORIES GRAMMATICALES	325
4.1	Les mots lexicaux	327

4.1.1	Les noms propres : NPr.....	327
4.1.2	Les noms (communs) : N.....	329
4.1.3	Les verbes : V.....	329
4.1.4	Les adjectifs : A.....	330
4.1.5	Les adverbes : ADV.....	331
4.2	Les mots grammaticaux.....	331
4.2.1	Les préfixes : PFX.....	331
4.2.2	Les pronoms : PRO.....	333
4.2.2.1	Les pronoms dits personnels : PRO+Ppers.....	333
4.2.2.2	Les pronoms relatifs : PRO+Prel.....	337
4.2.2.3	Les pronoms démonstratifs : PRO+Pdem.....	338
4.2.2.4	Les pronoms possessifs : PRO+Ppos.....	338
4.2.2.5	Les pronoms indéfinis : PRO+Pind.....	340
4.2.2.6	Les pronoms interrogatifs.....	340
4.2.3	Les déterminants : DET.....	341
4.2.3.1	Les déterminants définis : DET+Ddef.....	342
4.2.3.2	Les déterminants indéfinis : DET+Dind.....	343
4.2.3.3	Les déterminants possessifs : DET+Dpos.....	344
4.2.3.4	Les déterminants démonstratifs : DET+Ddem.....	345
4.2.3.5	Les déterminants numéraux : DET+Dnum.....	345
4.2.3.6	Les déterminants relatifs.....	347
4.2.3.7	Les déterminants exclamatifs et interrogatifs : DET+Dexi.....	347
4.2.3.8	Les conjonctions.....	347
4.2.4	Les conjonctions de subordination : CONJS.....	348
4.2.5	Les conjonctions de coordination : CONJC.....	348
4.2.6	Les prépositions : PREP.....	349
4.2.7	Les interjections : INTJ.....	350
4.3	La recatégorisation.....	351
4.4	Tableau synoptique du jeu d'étiquettes.....	351
4.5	Mots simples / mots composés.....	355
4.6	Conclusions sur nos ressources lexicales.....	359
4.7	Le Delfrabel : le dictionnaire électronique du français en Belgique.....	359
4.8	Les particularités lexicales du français de Belgique : enquête de vitalité et de diffusion.....	360
4.9	Constitution du DELFRABEL.....	361
4.9.1	Les entrées construites manuellement.....	361
4.9.2	La flexion automatique des verbes.....	362
4.9.3	Conclusions.....	366
CHAPITRE 6. L'ÉTIQUETAGE MORPHOSYNTAXIQUE.....		369
1	QU'EST-CE QU'ÉTIQUETER UN TEXTE ?.....	369
2	NOTRE ÉTIQUETAGE : L'ÉTIQUETAGE MORPHO-SYNTAXIQUE.....	371
3	LES CORPUS ANNOTÉS MORPHOSYNTAXIQUEMENT.....	372
3.1	Brown Corpus.....	373
3.2	The Lancaster-Oslo-Bergen Corpus.....	374
3.3	British National Corpus.....	374
3.4	Penn Treebank.....	376
3.5	Le Corpus d'Orléans.....	376
3.6	Le Corpus Paris 7.....	377
4	GRANULARITÉ DU JEU D'ÉTIQUETTES ET CONCEPT D'AMBIGÜITÉ.....	377
5	AMBIGÜITÉS FRÉQUENTES.....	383

6	LES MÉTHODES D'ÉTIQUETAGE	384
7	LA CAMPAGNE D'ÉVALUATION GRACE	386
8	FORMAT DE L'ÉTIQUETAGE.....	387
8.1	Une seule étiquette par forme.....	387
8.2	Plusieurs étiquettes par forme.....	388
9	MOTS INCONNUS	390
10	TRAVAILLER SUR L'AUTOMATE DU TEXTE	390
11	ELAG : ELIMINATION OF LEXICAL AMBIGUITIES BY GRAMMARS	394
11.1	Grammaire locale	395
11.2	Syntaxe d'une grammaire ELAG	395
11.2.1	Le nombre de branches alors.....	398
11.2.2	Le point de synchronisation	399
12	LE JEU D'ÉTIQUETTES.....	399
12.1	Le jeu d'étiquettes : le fichier tagset.def.....	400
12.2	La description du jeu d'étiquettes et la construction de l'automate du texte	401
12.3	Les masques lexicaux.....	402
12.4	Les symboles formels.....	403
13	MOT SIMPLE / MOT COMPOSÉ.....	404
14	INTÉRÊT DE SPÉCIFIER DES SOUS-CATÉGORIES	408
14.1	L'ambiguïté du s élidé.....	408
14.2	L'ambiguïté des déterminants numériques et des noms homographes	410
15	INDÉPENDANCE DES GRAMMAIRES	411
16	LES UNITÉS DE TRAITEMENT	412
17	UNITÉS DE TRAITEMENT REJETÉES	414
17.1	Coquilles et fautes d'orthographe.....	414
17.2	Incompatibilité liée aux données des dictionnaires.....	416
17.3	Les inachèvement syntaxiques.....	417
18	LES ÉTIQUETTES <IGN>.....	422
19	SUGGESTION POUR UNE AMÉLIORATION DU FORMALISME	425
20	ÉVALUATION : RÉSULTATS PRÉLIMINAIRES	427
20.1	Nos grammaires.....	428
20.2	Les unités de traitement rejetées.....	429
20.3	Résultats de l'étiquetage	432
20.3.1	Six catégories d'étiquettes.....	432
20.3.2	Le rappel	433
20.3.3	La précision.....	435
21	PERSPECTIVES	436
	CONCLUSION	439
	BIBLIOGRAPHIE	447
	TABLE DES FIGURES ET TABLEAUX	485
	TABLE DES MATIÈRES	489

ANNEXE 1 :

Le tableau ci-dessous présente les données sociologiques sur les locuteurs. L'ordre de présentation est l'ordre alphabétique du code des locuteurs.

Code	Sexe	Âge	Localisation géographique	Niveau d'étude
accBF1	féminin	21	3	1
accCP1	masculin	28	3	1
accCT0	féminin	22	3	1
accFJ1	masculin	47	3	1
accGP1	masculin	48	3	2
accPH1	féminin	63	3	7
blaAD0	féminin	32	5	1
blaAD1	féminin	32	5	1
blaAT1	féminin	24	5	1
blaBD1	masculin	34	5	1
blaCD1	féminin	51	5	1
blaGD1	masculin	62	5	1
blaJD1	féminin	76	5	3
blaJS1	masculin	36	5	1
blaJV1	masculin	21	5	3
blaNB1	féminin	36	5	1
blaND1	féminin	24	5	1
chaBP1	masculin	72	5	5
chaBR1	masculin	54	5	7
chaCL1	masculin	54	5	1
chaGG0	féminin	25	5	1
chaHI1	féminin	67	5	6
chaSR1	féminin	73	5	6
famDA1	féminin	50	9	1
famED1	masculin	53	9	2
famES1	masculin	19	9	1
famHC1	féminin	44	4	4
famMO1	féminin	15	2	5
famRM1	féminin	53	9	1
famVI1	masculin	26	9	1
famVJ1	masculin	54	9	2
famVV1	féminin	19	2	1
ilcBJ2	masculin	63	5	5
ilcCF1	masculin	45	1	1
ilcDA1	masculin	66	1	1
ilcMF0	féminin	26	3	1
ilcNJ2	masculin	29	4	1

ilcSP1	masculin	58	3	1
ileAO1	masculin	31	1	1
ileDC1	masculin	43	3	2
ileFN1	féminin	27	3	1
ileGF0	féminin	27	9	1
ileGG1	masculin	43	4	1
ilePA2	féminin	59	5	2
iljBB1	masculin	36	4	2
iljDV1	féminin	24	3	1
iljGE1	masculin	35	3	1
iljGF0	féminin	27	9	1
iljNJ1	masculin	30	5	1
iljPF1	féminin	42	9	1
iljVM1	féminin	34	9	1
ilpBP1	masculin	48	4	2
ilpCM1	masculin	38	2	1
ilpDJ1	masculin	51	6	1
ilpGV0	masculin	22	2	1
ilpHJ1	masculin	44	5	4
ilpMJ1	féminin	34	3	1
ilpSA1	féminin	63	1	1
ilpSG1	masculin	54	4	2
ilpSS0	féminin	22	2	1
ilpTN0	féminin	21	3	1
ilrCP1	masculin	20	1	4
ilrDT1	masculin	18	6	5
ilrGS2	féminin	20	6	3
ilrLD3	masculin	20	3	3
ilrLP1	féminin	18	1	5
ilrMA1	féminin	20	4	3
ilrMF0	féminin	26	3	1
ilrMS1	féminin	19	5	3
ilrPC1	féminin	18	3	3
ilrTC1	masculin	20	5	4
ilrTF1	féminin	19	6	4
ilrTJ1	masculin	19	4	3
magFM0	masculin	48	6	1
magMM1	féminin	67	6	6
magPM1	féminin	35	6	5
norFA0	féminin	21	6	1
norGA1	masculin	61	9	1
norHJ1	masculin	85	4	1
norIN1	féminin	36	6	1
norKJ1	masculin	43	5	1

norPM1	féminin	44	9	1
norRM1	féminin	37	3	1
souGJ1	masculin	33	3	2
souGJ2	masculin	1	3	
souGJ3	féminin	8	3	
souGS1	masculin	5	3	
souLC1	féminin	32	3	3
souTJ1	masculin	31	9	1
souVC1	féminin	29	3	1

Légende :

LOCALISATION

1	Bruxelles
2	Brabant wallon
3	Hainaut
4	Namur
5	Liège
6	Luxembourg
9	Mobilité

NIVEAU D'ÉTUDES

1	Universitaire
2	Supérieur non universitaire
3	Secondaire professionnel
4	Secondaire technique
5	Secondaire général
6	Secondaire niveau inférieur
7	Primaire

ANNEXE 2

Le tableau ci-dessous présente pour chaque texte du corpus la répartition du nombre de mots par locuteur et par texte. L'ordre de présentation est l'ordre alphabétique du code des textes.

Texte	Locuteur	Nombre de mots
accBF1r	accBF1	2987
accBF1r	accCT0	3754
Total accBF1r		6741
accCP1r	accCP1	8011
accCP1r	accCT0	3478
Total accCP1r		11489
accFJ1r	accCT0	3215
accFJ1r	accFJ1	8859
Total accFJ1r		12074
accGP1r	accCT0	4527
accGP1r	accGP1	6524
Total accGP1r		11051
accPH1r	accCT0	3211
accPH1r	accPH1	4900
Total accPH1r		8111
blaAT1l	blaAT1	6589
blaAT1l	blaND1	1577
blaAT1l	blaPH0	11
Total blaAT1l		8177
blaGD1l	blaBD1	1858
blaGD1l	blaGD1	1641
Total blaGD1l		3499
blaJD1l	blaAD0	48
blaJD1l	blaCD1	889
blaJD1l	blaJD1	3580

Total blaJD1l		4517
blaJV1l	blaJV1	4959
blaJV1l	blaND1	1413
Total blaJV1l		6372
blaNB1l	blaAD1	2692
blaNB1l	blaJS1	106
blaNB1l	blaNB1	3816
Total blaNB1l		6614
chaBP1r	chaBP1	5125
chaBP1r	chaGG0	588
chaBP1r	chbi	0
Total chaBP1r		5713
chaBR1r	chaBR1	8902
chaBR1r	chaGG0	536
chaBR1r	chaXX2	160
Total chaBR1r		9598
chaCL1r	chaCL1	6754
chaCL1r	chaGG0	542
Total chaCL1r		7296
chaHI1r	chaGG0	429
chaHI1r	chaHI1	1955
Total chaHI1r		2384
chaSR1r	chaGG0	322
chaSR1r	chaSR1	964
Total chaSR1r		1286
famES1r	famED1	3457
famES1r	famES1	6680

famES1r	famHC1	4716
Total famES1r		14853
famRM1r	famMO1	1065
famRM1r	famRM1	3187
Total famRM1r		4252
famVV1r	famDA1	629
famVV1r	famVI1	2008
famVV1r	famVJ1	46
famVV1r	famVV1	1537
Total famVV1r		4220
ilcBJ2r	ilcBJ2	2207
ilcBJ2r	ilcMF0	917
Total ilcBJ2r		3124
ilcCF1r	ilcCF1	2764
ilcCF1r	ilcMF0	1033
Total ilcCF1r		3797
ilcDA1r	ilcDA1	7682
ilcDA1r	ilcMF0	1636
Total ilcDA1r		9318
ilcNJ2r	ilcMF0	867
ilcNJ2r	ilcNJ2	3211
Total ilcNJ2r		4078
ilcSP1r	ilcMF0	817
ilcSP1r	ilcSP1	1720
ilcSP1r	ilcXX1	14
Total ilcSP1r		2551
ileAO1r	ileAO1	6229
ileAO1r	ileGF0	1876
Total ileAO1r		8105
ileDC1r	ileDC1	9021
ileDC1r	ileGF0	791

Total ileDC1r		9812
ileFN1r	ileFN1	4291
ileFN1r	ileGF0	1724
ileFN1r	ileXX8	0
Total ileFN1r		6015
ileGG1r	ileGF0	1894
ileGG1r	ileGG1	7096
Total ileGG1r		8990
ilePA2r	ileGF0	927
ilePA2r	ilePA2	9695
ilePA2r	ilexx9	13
Total ilePA2r		10635
iljBB1r	iljBB1	4394
iljBB1r	iljGF0	859
iljBB1r	iljXX3	9
Total iljBB1r		5262
iljDV1r	iljDV1	5561
iljDV1r	iljGF0	771
Total iljDV1r		6332
iljGE1r	iljGE1	5739
iljGE1r	iljGF0	997
iljGE1r	iljXX1	5
Total iljGE1r		6741
iljNJ1r	iljGF0	1093
iljNJ1r	iljNJ1	5132
Total iljNJ1r		6225
iljPF1r	iljGF0	1641
iljPF1r	iljPF1	11338
iljPF1r	iljXX2	2
iljPF1r	iljXX4	2
iljPF1r	iljXX5	25

Total iljPF1r		13008
iljVM1r	iljGF0	775
iljVM1r	iljVM1	4354
Total iljVM1r		5129
ilpBP1r	ilpBP1	8079
ilpBP1r	ilpTN0	712
ilpBP1r	ilpWW1	15
ilpBP1r	ilpXX1	7
ilpBP1r	ilpXX2	26
ilpBP1r	ilpXX3	3
Total ilpBP1r		8842
ilpCM1r	ilpCM1	8231
ilpCM1r	ilpGV0	1426
Total ilpCM1r		9657
ilpDJ1r	ilpDJ1	4040
ilpDJ1r	ilpGV0	13
ilpDJ1r	ilpSS0	1632
Total ilpDJ1r		5685
ilpHJ1r	ilpHJ1	2441
ilpHJ1r	ilpTN0	760
Total ilpHJ1r		3201
ilpMJ1r	ilpGV0	1208
ilpMJ1r	ilpMJ1	6988
ilpMJ1r	ilpSS0	11
ilpMJ1r	ilpXX1	16
Total ilpMJ1r		8223
ilpSA1r	ilpSA1	2220
ilpSA1r	ilpTN0	474
Total ilpSA1r		2694
ilpSG1r	ilpSG1	3353
ilpSG1r	ilpTN0	700

Total ilpSG1r		4053
ilrCP1r	ilrCP1	8098
ilrCP1r	ilrMF0	1168
ilrCP1r	ilrXX7	8
Total ilrCP1r		9274
ilrDT1r	ilrDT1	3409
ilrDT1r	ilrMF0	1860
Total ilrDT1r		5269
ilrGS2r	ilrGS2	6350
ilrGS2r	ilrMF0	1768
Total ilrGS2r		8118
ilrLD3r	ilrLD3	11552
ilrLD3r	ilrMF0	1170
ilrLD3r	ilrXX3	9
Total ilrLD3r		12731
ilrLP1r	ilrLP1	4271
ilrLP1r	ilrMF0	1176
ilrLP1r	ilrXX6	14
Total ilrLP1r		5461
ilrMA1r	ilrMA1	6847
ilrMA1r	ilrMF0	1253
Total ilrMA1r		8100
ilrMS1r	ilrMF0	1377
ilrMS1r	ilrMS1	8866
ilrMS1r	ilrXX5	22
ilrMS1r	ilrXX8	6
Total ilrMS1r		10271
ilrPC1r	ilrMF0	1721
ilrPC1r	ilrPC1	6894
Total ilrPC1r		8615
ilrTC1r	ilrMF0	1935

ilrTC1r	ilrTC1	6664
ilrTC1r	ilrXX4	17
Total ilrTC1r		8616
ilrTF1r	ilrMF0	1703
ilrTF1r	ilrTF1	5533
Total ilrTF1r		7236
ilrTJ1r	ilrMF0	1054
ilrTJ1r	ilrTJ1	5287
Total ilrTJ1r		6341
magPM1r	magFM0	2805
magPM1r	magMM1	1390
magPM1r	magPM1	5815
Total magPM1r		10010
norGA1r	norFA0	964
norGA1r	norGA1	6612
Total norGA1r		7576
norHJ1r	norFA0	3074
norHJ1r	norHJ1	14189
Total norHJ1r		17263
norIN1r	norFA0	1738

norIN1r	norIN1	4997
Total norIN1r		6735
norKJ1r	norFA0	1809
norKJ1r	norKJ1	9624
Total norKJ1r		11433
norPM1r	norFA0	2004
norPM1r	norPM1	6087
Total norPM1r		8091
norRM1r	norFA0	1059
norRM1r	norRM1	4615
Total norRM1r		5674
souVC1r	souGJ1	1831
souVC1r	souGJ2	23
souVC1r	souGJ3	301
souVC1r	souGS1	174
souVC1r	souLC1	1367
souVC1r	souTJ1	993
souVC1r	souVC1	1820
Total souVC1r		6509
Total		443047

ANNEXE 3 :

Le tableau ci-dessous présente le nombre moyen de tours de parole par texte, par ordre décroissant.

Entretien	Nombre de mots	Nombre de tours	longueur moyenne des tours (n ^{bre} mots / n ^{bre} tours)
chaCL1r	7296	104	70,15
chaBR1r	9598	183	52,45
ilpBP1r	8842	169	52,32
chaBP1r	5713	112	51,01
iljDV1r	6332	154	41,12
iljPF1r	13008	321	40,52
norGAr	7576	192	39,46
ilrLD3r	12731	362	35,17
ileDC1r	9812	289	33,95
ilePA2r	10635	316	33,66
ilpCM1r	9657	311	31,05
ilrCP1r	9274	306	30,31
norKJ1r	11433	382	29,93
iljBB1r	5262	177	29,73
accCP1r	11489	387	29,69
iljGE1r	6741	229	29,44
chaHI1r	2384	83	28,72
iljNJ1r	6225	228	27,30
iljVM1r	5129	188	27,28
norRM1r	5674	213	26,64
ilpSA1r	2694	102	26,41
ilpMJ1r	8223	333	24,69
ilrLP1r	5461	223	24,49
ilpDJ1r	5685	234	24,29
blaAT1l	8177	378	21,63
norPM1r	8091	382	21,18
ilrMS1r	10271	511	20,10
norIN1r	6735	336	20,04
ilrMA1r	8100	413	19,61
ilpHJ1r	3201	167	19,17
accFJ1r	12074	655	18,43
norHJ1r	17263	948	18,21
ilrTF1r	7236	406	17,82
blaJV1l	6372	367	17,36
ileCF1r	3797	227	16,73
ilpSG1r	4053	244	16,61

magPM1r	10010	611	16,38
chaSR1r	1286	80	16,08
blaGD1l	3499	222	15,76
ilrPC1r	8615	559	15,41
ilcNJ2r	4078	265	15,39
ilcBJ2r	3124	208	15,02
blaJD1l	4517	304	14,86
ilrTJ1r	6341	427	14,85
ileFN1r	6015	407	14,78
accBF1r	6741	462	14,59
accGP1r	11051	782	14,13
ilcDA1r	9318	683	13,64
famRM1r	4252	319	13,33
ilrGS2r	8118	612	13,26
ileAO1r	8105	618	13,11
blaNB1l	6614	516	12,82
ilrTC1r	8616	685	12,58
famVV1r	4220	370	11,41
ileGG1r	8990	791	11,37
accPH1r	8111	721	11,25
famES1r	14853	1392	10,67
ilcSP1r	2551	240	10,63
ilrDT1r	5269	560	9,41
souVC1r	6509	921	7,07

ANNEXE 4 :

Le tableau ci-dessous le nombre moyen de tours de parole par locuteur, par ordre décroissant.

Entretien	Locuteurs	Nombre de mots	Nombre de tours	Longueur moyenne des tours (= n^{bre} de mots / n^{bre} de tours)
chaCL1r	chaCL1	6754	53	127,43
ilpBP1r	ilpBP1r	8079	75	107,72
chaBR1r	chaBR1	8902	90	98,91
chaBP1r	chaBP1	5125	56	91,52
iljPF1r	iljPF1	11338	145	78,19
norGA1r	norGA1	6612	88	75,14
iljDV1r	iljDV1	5561	78	71,29
ilePA2r	ilePA2	9695	143	67,80
ilrLD3r	ilrLD3	11552	176	65,64
ilpCM1r	ilpCM1	8231	128	64,30
ileDC1r	ileDC1	9021	145	62,21
ilrCP1r	ilrCP1	8098	148	54,72
ilpMJ1r	ilpMJ1	6988	137	51,01
norKJ1r	norKJ1	9624	190	50,65
iljGE1r	iljGE1	5739	119	48,23
chaHI1r	chaHI1	1955	42	46,55
ilpSA1r	ilpSA1	2220	48	46,25
accCP1r	accCP1	8011	175	45,78
iljBB1r	iljBB1	4394	96	45,77
iljNJ1r	iljNJ1	5132	115	44,63
iljVM1r	iljVM1	4354	98	44,43
norRM1r	norRM1	4615	105	43,95
ilrLP1r	ilrLP1	4271	108	39,55
blaAT1l	blaAT1	6589	172	38,31
ilrMS1r	ilrMS1	8866	247	35,89
accFJ1r	accFJ1	8859	254	34,88
ilpDJ1r	ilpDJ1	4040	117	34,53
ilrMA1r	ilrMA1	6847	202	33,90
norHJ1r	norHJ1	14189	440	32,25
norPM1r	norPM1	6087	193	31,54
ilpHJ1r	ilpHJ1	2441	80	30,51
norIN1r	norIN1	4997	169	29,57
blaJV1l	blaJV1	4959	169	29,34
ilpSG1r	ilpSG1	3353	118	28,42
ilrTF1r	ilrTF1	5533	201	27,53
ilcNJ2r	ilcNJ2	3211	119	26,98
ilrTJ1r	ilrTJ1	5287	205	25,79

ilrPC1r	ilrPC1	6894	274	25,16
blaJD1l	blaJD1	3580	143	25,03
ilcCF1r	ilcCF1	2764	112	24,68
chaSR1r	chaSR1	964	40	24,10
ileFN1r	ileFN1	4291	180	23,84
ilcDA1r	ilcDA1	7682	335	22,93
famRM1r	famRM1	3187	140	22,76
magPM1r	magPM1	5815	256	22,71
ileAO1r	ileAO1	6229	276	22,57
ilcBJ2r	ilcBJ2	2207	103	21,43
magPM1r	magMM1	1390	66	21,06
ilrGS2r	ilrGS2	6350	305	20,82
ilrTC1r	ilrTC1	6664	323	20,63
ileGG1r	ileGG1	7096	379	18,72
accGP1r	accGP1	6524	349	18,69
blaGD1l	blaBD1	1858	110	16,89
accBF1r	accCT0	3754	226	16,61
accCP1r	accCT0	3478	212	16,41
famVV1r	famVI1	2008	135	14,87
blaNB1l	blaNB1	3816	257	14,85
accPH1r	accPH1	4900	333	14,71
ilcSP1r	ilcSP1	1720	117	14,70
blaGD1l	blaGD1	1641	112	14,65
ilpDJ1r	ilpSS0	1632	114	14,32
accBF1r	accBF1	2987	236	12,66
ilrDT1r	ilrDT1	3409	282	12,09
famES1r	famES1	6680	572	11,68
blaNB1l	blaAD1	2692	241	11,17
iljBB1r	iljGF0	859	80	10,74
famVV1r	famVV1	1537	144	10,67
chaCL1r	chaGG0	542	51	10,63
norPM1r	norFA0	2004	189	10,60
chaBP1r	chaGG0	588	56	10,50
ilrLP1r	ilrMF0	1176	112	10,50
chaHI1r	chaGG0	429	41	10,46
accGP1r	accCT0	4527	433	10,45
norIN1r	norFA0	1738	167	10,41
iljDV1r	iljGF0	771	76	10,14
famES1r	famHC1	4716	467	10,10
magPM1r	magFM0	2805	283	9,91
souVC1r	souGJ1	1831	185	9,90
norRM1r	norFA0	1059	108	9,81
famES1r	famED1	3457	353	9,79
iljNJ1r	iljGF0	1093	113	9,67

iljPF1r	iljGF0	1641	171	9,60
norKJ1r	norFA0	1809	192	9,42
norGAR	norFA0	964	104	9,27
iljGE1r	iljGF0	997	109	9,15
ilcCF1r	ilcMF0	1033	115	8,98
ilpSA1r	ilpTN0	474	54	8,78
ilpHJ1r	ilpTN0	760	87	8,74
ilcBJ2r	ilcMF0	917	105	8,73
iljVM1r	iljGF0	775	90	8,61
ilrTF1r	ilrMF0	1703	205	8,31
ilpBP1r	ilpTN0	712	86	8,28
accPH1r	accCT0	3211	388	8,28
chaSR1r	chaGG0	322	40	8,05
accFJ1r	accCT0	3215	401	8,02
ilpCM1r	ilpGV0	1426	183	7,79
blaAT1l	blaND1	1577	203	7,77
famVV1r	famDA1	629	81	7,77
ileFN1r	ileGF0	1724	227	7,59
ilrCP1r	ilrMF0	1168	154	7,58
souVC1r	souVC1	1820	247	7,37
ilcSP1r	ilcMF0	817	111	7,36
chaBR1r	chaGG0	536	75	7,15
blaJV1l	blaND0	1413	198	7,14
ilrDT1r	ilrMF0	1860	278	6,69
ilrLD3r	ilrMF0	1170	184	6,36
souVC1r	souTJ1	993	158	6,28
ilpMJ1r	ilpGV0	1208	193	6,26
souVC1r	souGJ3	301	49	6,14
norHJ1r	norFA0	3074	508	6,05
ilrPC1r	ilrMF0	1721	285	6,04
souVC1r	souLC1	1367	227	6,02
famRM1r	famMO1	1065	178	5,98
ilrMA1r	ilrMF0	1253	211	5,94
ilcNJ2r	ilcMF0	867	146	5,94
blaNB1l	blaJS1	106	18	5,89
blaJD1l	blaCD1	889	152	5,85
ilrGS2r	ilrMF0	1768	307	5,76
ilpSG1r	ilpTN0	700	126	5,56
ilpMJ1r	ilpSS0	11	2	5,50
ileDC1r	ileGF0	791	144	5,49
ileAO1r	ileGF0	1876	342	5,49
ilePA2r	ileGF0	927	171	5,42
ilrTC1r	ilrMF0	1935	357	5,42
ilrMS1r	ilrMF0	1377	256	5,38

blaJD1l	blaAD0	48	9	5,33
ilrTJ1r	ilrMF0	1054	222	4,75
ilcDA1r	ilcMF0	1636	348	4,70
famVV1r	famVJ1	46	10	4,60
ileGG1r	ileGF0	1894	412	4,60
souVC1r	souGS1	174	40	4,35
ilpDJ1r	ilpGV0	13	3	4,33
souVC1r	souGJ2	23	15	1,53

ANNEXE 5 :

Le tableau ci-dessous présente la proportion de répétées par locuteur, par ordre décroissant.

Entretien	Locuteurs	Proportion de répétés
norHJ1r	norFA0	10,18%
blaJD1l	blaCD1	6,30%
accFJ1r	accFJ1	5,81%
ileGG1r	ileGG1	5,79%
ilcNJ2r	ilcNJ2	5,67%
iljDV1r	iljDV1	5,54%
ilpMJ1r	ilpMJ1	5,54%
ileDC1r	ileDC1	5,38%
souVC1r	souGJ3	5,32%
ilpSA1r	ilpSA1	5,27%
ilcDA1r	ilcDA1	5,21%
famES1r	famHC1	5,15%
ileFN1r	ileGF0	5,10%
ilrDT1r	ilrDT1	4,96%
ilrLP1r	ilrLP1	4,80%
blaNB1l	blaNB1	4,77%
ilrCP1r	ilrCP1	4,74%
ileGG1r	ileGF0	4,59%
blaJD1l	blaJD1	4,53%
ileAO1r	ileGF0	4,42%
norPM1r	norFA0	4,39%
accGP1r	accGP1	4,32%
magPM1r	magMM1	4,32%
ilpCM1r	ilpCM1	4,08%
famRM1r	famRM1	4,05%
famES1r	famED1	3,99%
famRM1r	famMO1	3,94%
iljNJ1r	iljNJ1	3,94%
ilePA2r	ilePA2	3,85%
iljVM1r	iljVM1	3,84%
ilpCM1r	ilpGV0	3,79%
norKJ1r	norKJ1	3,75%
norHJ1r	norHJ1	3,72%
iljBB1r	iljBB1	3,71%
ilrTJ1r	ilrTJ1	3,69%
blaAT1l	blaAT1	3,66%
ilePA2r	ileGF0	3,56%
iljPF1r	iljPF1	3,55%

norPM1r	norPM1	3,52%
accPH1r	accPH1	3,47%
ilpDJ1r	ilpDJ1	3,42%
accBF1r	accBF1	3,38%
norKJ1r	norFA0	3,32%
chaHI1r	chaHI1	3,27%
ilcDA1r	ilcMF0	3,24%
ilpHJ1r	ilpHJ1	3,24%
ileFN1r	ileFN1	3,22%
norIN1r	norFA0	3,16%
norIN1r	norIN1	3,12%
norRM1r	norRM1	3,06%
ilrTC1r	ilrTC1	3,05%
chaBR1r	chaBR1	2,97%
iljGE1r	iljGE1	2,96%
blaNB1l	blaJS1	2,83%
norGAr	norFA0	2,80%
ileAO1r	ileAO1	2,76%
ilrPC1r	ilrMF0	2,73%
blaNB1l	blaAD1	2,71%
ilpBP1r	ilpTN0	2,67%
accCP1r	accCP1	2,66%
ilcCF1r	ilcCF1	2,64%
chaBR1r	chaGG0	2,61%
ilpHJ1r	ilpTN0	2,50%
chaBP1r	chaBP1	2,48%
iljPF1r	iljGF0	2,44%
blaGD1l	blaBD1	2,37%
norRM1r	norFA0	2,36%
magPM1r	magFM0	2,35%
ilrMS1r	ilrMS1	2,28%
ileDC1r	ileGF0	2,28%
ilrMA1r	ilrMA1	2,25%
blaJV1l	blaJV1	2,20%
blaGD1l	blaGD1	2,13%
famES1r	famES1	2,13%
ilrGS2r	ilrMF0	2,09%
famVV1r	famVV1	2,08%
ilcSP1r	ilcSP1	1,92%
ilrPC1r	ilrPC1	1,90%
ilrMS1r	ilrMF0	1,89%
ilrDT1r	ilrMF0	1,88%
chaHI1r	chaGG0	1,86%
ilpSG1r	ilpTN0	1,86%

ilpDJ1r	ilpSS0	1,84%
ilpBP1r	ilpBP1r	1,83%
ilcBJ2r	ilcBJ2	1,81%
ilrCP1r	ilrMF0	1,71%
norGAr	norGA1	1,69%
ilpSA1r	ilpTN0	1,69%
ilpSG1r	ilpSG1	1,67%
accFJ1r	accCT0	1,65%
ilrLD3r	ilrLD3	1,64%
blaJV1l	blaND0	1,63%
accPH1r	accCT0	1,62%
ilrGS2r	ilrGS2	1,61%
souVC1r	souVC1	1,59%
chaSR1r	chaGG0	1,55%
ilrTF1r	ilrMF0	1,53%
chaCL1r	chaCL1	1,51%
magPM1r	magPM1	1,50%
ilpMJ1r	ilpGV0	1,49%
souVC1r	souGJ1	1,47%
ilrTF1r	ilrTF1	1,46%
ilcCF1r	ilcMF0	1,45%
accBF1r	accCT0	1,44%
famVV1r	famDA1	1,43%
ilrTC1r	ilrMF0	1,40%
souVC1r	souLC1	1,39%
chaSR1r	chaSR1	1,35%
iljGE1r	iljGF0	1,30%
accGP1r	accCT0	1,30%
iljVM1r	iljGF0	1,29%
souVC1r	souTJ1	1,21%
blaAT1l	blaND1	1,20%
ilcSP1r	ilcMF0	1,10%
ilrLD3r	ilrMF0	1,03%
iljNJ1r	iljGF0	1,01%
chaBP1r	chaGG0	0,85%
ilrLP1r	ilrMF0	0,85%
ilcNJ2r	ilcMF0	0,81%
ilrMA1r	ilrMF0	0,80%
accCP1r	accCT0	0,75%
iljBB1r	iljGF0	0,70%
iljDV1r	iljGF0	0,65%
chaCL1r	chaGG0	0,55%
ilcBJ2r	ilcMF0	0,55%
ilrTJ1r	ilrMF0	0,47%

famVV1r	famVI1	0,10%
----------------	--------	-------

ANNEXE 6 :

Le tableau ci-dessous présente la proportion d'amorces par locuteur, par ordre décroissant. La dernière colonne indique le nombre moyen de mots qui sépare deux amorces.

Locuteurs	Entretien	Nbre d'amorces	Proportion d'amorces	Une amorce tous les x mots
accPH1	accPH1r	101	2,06%	49
norHJ1	norHJ1r	277	1,95%	51
blaAT1	blaAT1l	121	1,84%	54
blaJV1	blaJV1l	90	1,81%	55
blaJD1	blaJD1l	60	1,68%	60
blaAD1	blaNB1l	43	1,60%	63
norKJ1	norKJ1r	152	1,58%	63
ilcNJ2	ilcNJ2r	50	1,56%	64
norPM1	norPM1r	90	1,48%	68
blaCD1	blaJD1l	13	1,46%	68
magMM1	magPM1r	19	1,37%	73
norFA0	norHJ1r	40	1,30%	77
ilpMJ1	ilpMJ1r	88	1,26%	79
ileDC1	ileDC1r	111	1,23%	81
blaNB1	blaNB1l	46	1,21%	83
ileFN1	ileFN1r	49	1,14%	88
blaND0	blaJV1l	16	1,13%	88
ilePA2	ilePA2r	109	1,12%	89
ilrLP1	ilrLP1r	48	1,12%	89
accGP1	accGP1r	72	1,10%	91
famED1	famES1r	37	1,07%	93
famHC1	famES1r	50	1,06%	94
ileGG1	ileGG1r	72	1,01%	99
iljNJ1	iljNJ1r	52	1,01%	99
norFA0	norPM1r	20	1,00%	100
ilrDT1	ilrDT1r	34	1,00%	100
souGJ3	souVC1r	3	1,00%	100
norFA0	norKJ1r	18	1,00%	101
ilpSA1	ilpSA1r	22	0,99%	101
famVV1	famVV1r	15	0,98%	102
famDA1	famVV1r	6	0,95%	105
blaND1	blaAT1l	15	0,95%	105
norIN1	norIN1r	47	0,94%	106
famMO1	famRM1r	10	0,94%	107
iljDV1	iljDV1r	51	0,92%	109
ilpGV0	ilpCM1r	13	0,91%	110
ilrCP1	ilrCP1r	73	0,90%	111

iljPF1	iljPF1r	100	0,88%	113
blaBD1	blaGD1l	16	0,86%	116
ilrPC1	ilrPC1r	59	0,86%	117
accFJ1	accFJ1r	75	0,85%	118
accCT0	accPH1r	27	0,84%	119
famES1	famES1r	55	0,82%	121
famVI1	famVV1r	16	0,80%	126
iljGE1	iljGE1r	45	0,78%	128
iljBB1	iljBB1r	34	0,77%	129
chaHI1	chaHI1r	15	0,77%	130
magPM1	magPM1r	44	0,76%	132
norFA0	norIN1r	13	0,75%	134
norGA1	norGAr	49	0,74%	135
ilcDA1	ilcDA1r	55	0,72%	140
ilrGS2	ilrGS2r	45	0,71%	141
ilrTC1	ilrTC1r	47	0,71%	142
norFA0	norRM1r	7	0,66%	151
famRM1	famRM1r	21	0,66%	152
ileGF0	ilePA2r	6	0,65%	155
ilpDJ1	ilpDJ1r	26	0,64%	155
ilrTJ1	ilrTJ1r	34	0,64%	156
chaSR1	chaSR1r	6	0,62%	161
norFA0	norGAr	6	0,62%	161
ileAO1	ileAO1r	38	0,61%	164
ilpCM1	ilpCM1r	50	0,61%	165
souLC1	souVC1r	8	0,59%	171
norRM1	norRM1r	27	0,59%	171
ileGF0	ileFN1r	10	0,58%	172
ilpHJ1	ilpHJ1r	14	0,57%	174
ilpTN0	ilpBP1r	4	0,56%	178
blaGD1	blaGD1l	9	0,55%	182
iljGF0	iljPF1r	9	0,55%	182
ilpSG1	ilpSG1r	18	0,54%	186
accBF1	accBF1r	16	0,54%	187
ilrLD3	ilrLD3r	60	0,52%	193
accCP1	accCP1r	41	0,51%	195
iljVM1	iljVM1r	22	0,51%	198
ilrMF0	ilrPC1r	8	0,46%	215
ileBJ2	ileBJ2r	10	0,45%	221
ileGF0	ileGG1r	8	0,42%	237
ilrMF0	ilrTF1r	7	0,41%	243
ilpBP1r	ilpBP1r	32	0,40%	252
ilrMF0	ilrGS2r	7	0,40%	253
ilrMS1	ilrMS1r	35	0,39%	253

magFM0	magPM1r	11	0,39%	255
iljGF0	iljDV1r	3	0,39%	257
ilrMA1	ilrMA1r	26	0,38%	263
ilpSS0	ilpDJ1r	6	0,37%	272
ilcMF0	ilcSP1r	3	0,37%	272
accCT0	accGP1r	16	0,35%	283
iljGF0	iljBB1r	3	0,35%	286
ilrMF0	ilrCP1r	4	0,34%	292
ilrMF0	ilrLP1r	4	0,34%	294
chaBP1	chaBP1r	17	0,33%	301
ilpGV0	ilpMJ1r	4	0,33%	302
souGJ1	souVC1r	6	0,33%	305
accCT0	accFJ1r	10	0,31%	322
iljGF0	iljGE1r	3	0,30%	332
ilrMF0	ilrMS1r	4	0,29%	344
ilrTF1	ilrTF1r	16	0,29%	346
accCT0	accCP1r	10	0,29%	348
ilpTN0	ilpSG1r	2	0,29%	350
ilrMF0	ilrDT1r	5	0,27%	372
iljGF0	iljVM1r	2	0,26%	388
ilcCF1	ilcCF1r	7	0,25%	395
chaCL1	chaCL1r	17	0,25%	397
ilcMF0	ilcDA1r	4	0,24%	409
chaGG0	chaHI1r	1	0,23%	429
chaBR1	chaBR1r	20	0,22%	445
ilpTN0	ilpSA1r	1	0,21%	474
souTJ1	souVC1r	2	0,20%	497
ilcMF0	ilcCF1r	2	0,19%	517
ilrMF0	ilrTJ1r	2	0,19%	527
chaGG0	chaBR1r	1	0,19%	536
chaGG0	chaCL1r	1	0,18%	542
ilcSP1	ilcSP1r	3	0,17%	573
ilrMF0	ilrLD3r	2	0,17%	585
souVC1	souVC1r	3	0,16%	607
ileGF0	ileAO1r	3	0,16%	625
accCT0	accBF1r	6	0,16%	626
ilpTN0	ilpHJ1r	1	0,13%	760
ileGF0	ileDC1r	1	0,13%	791
ilcMF0	ilcBJ2r	1	0,11%	917
ilrMF0	ilrTC1r	2	0,10%	968
iljGF0	iljNJ1r	1	0,09%	1093
ilrMF0	ilrMA1r	1	0,08%	1253
blaJS1	blaNB1l	0	0,00%	0
chaGG0	chaBP1r	0	0,00%	0

chaGG0	chaSR1r	0	0,00%	0
ilcMF0	ilcNJ2r	0	0,00%	0

ANNEXE 7 : LE FICHIER TAGSET.DEF

NAME french

adjectifs
POS A

flex:
genre= m f
nombre = s p
.

##adverbes
POS ADV
.

#alternance de code
POS ALT
.

#amorces (celles qui ne sont pas balisées par Gérard, càd les non
bégayantes)
POS AMO
.

conjonction de coordination
POS CONJC
.

##conjonctions de subordination
POS CONJS

cat:
souscat1 = Eli
.

déterminants
POS DET

flex:
genre= m f
nombre = s p

cat:

souscat = Ddef Dind Dpos Ddem Dnum Dexi
possesseur = Dpos1s Dpos2s Dpos3s Dpos1p Dpos2p Dpos3p
contraction = Prepdet
.

##enregistrement
POS ENR
.

```
##portions balisées à ignorer lors de l'application des grammaires
POS IGN
IGNORE
```

```
cat:
```

```
souscat = Euh ChF ChO Loc Phon Rep Amo Cor Mult MulO MulF
```

```
.
```

```
#passages incompréhensibles
POS INC
```

```
.
```

```
#point d'interrogation
POS INT
```

```
.
```

```
##interjections
POS INTJ
```

```
.
```

```
## noms
POS N
```

```
flex:
genre = m f
nombre = s p
```

```
cat:
```

```
souscat = Nnum
```

```
.
```

```
#locuteurs
POS NLoc
```

```
.
```

```
# locuteurs dans chevauchements internes (qui ont été extraits)
POS NLocChev
```

```
.
```

```
##noms propres
POS NPr
```

```
.
```

```
#la pause brève est étiquetée PAB et ignorée par les grammaires Elag
POS PAB
IGNORE
```

```
.
```

```
#pause longue //
POS PAL
```

```
.
```

#ce qui est entre parenthèses, le paraverbal
POS PAR

.

##préfixes
POS PFX
#{porto,porto.PFX+z2}

.

##prépositions
POS PREP

flex:
genre = m f
nombre = s p

cat:

souscat1 = eff
souscat2 = Prepdet
souscat3 = Preppro

.

#silences
POS SIL

.

#balises de son
POS SON

.

pronoms
POS PRO

flex:
pers = 1 2 3
genre= m f
nombre = s p

cat:

souscat = Pind Pdem Ppers Ppos Prel
possesseur = Ppos1s Ppos2s Ppos3s Ppos1p Ppos2p Ppos3p
contraction = Preppro
elision = Eli

.

verbes
POS V

flex:
temps= C F I J K P S T W Y G X B
pers = 1 2 3
genre= m f
nombre = s p

.

POS X
flex:

genre = m f
nombre = s p

.

POS XI

.

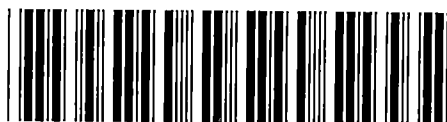
#pour le t euphonique, et uniquement lui. Mettre X ou XI n'est pas possible puisque ces codes sont enlevés #par une grammaire Elag

POS XX

.



Université catholique de Louvain



10.500.508