Analyse syntaxique automatique de l'oral : étude des disfluences

Rémi Bove

Université Aix-Marseille I – Université de Provence
U.F.R Lettres, Arts, Communication et Sciences du Langage
(L.A.C.S)

N ° attribué par la bibliothèque

|_|_|_|_|_|_|_|_|_|_|_|_|_|

Thèse

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITE D'AIX-MARSEILLE I

Formation Doctorale :

Langage et Parole (mention Traitement Automatique des Langues)

Présentée et soutenue publiquement
Par

Rémi BOVE

le 25 novembre 2008

Analyse syntaxique automatique de l'oral : étude des disfluences

TOME I

Directeur de thèse : Jean Véronis

Jury

Mme. Martine ADDA-DECKER	(Université de Paris-Sud, rapporteur)
M. Jean-Yves ANTOINE	(Université de Tours, rapporteur)
M. Henri-José DEULOFEU	(Université de Provence, examinateur)
M. Jacques VERGNE	(Université de Caen, examinateur)
M. Jean VÉRONIS	(Université de Provence, directeur)

"[H]esitation phenomena [. . .] provide good evidence that speaking is not a matter of regurgitating material already stored in the mind in linguistic form, but that it is a creative art, relating two media, thought and language, which are not isomorphic but require adjustments and readjustments to each other. A speaker does not follow a clear, well traveled path, but must find his way through territory not traversed before, where pauses, changes of direction, and retracing of steps are quite to be expected. The fundamental reason for hesitating is that speech production is an act of creation."

(W.L. Chafe, 1980, p. 170.)

Remerciements

Je tiens, en premier lieu, à remercier chaleureusement Jean Véronis pour m'avoir encadré durant cette thèse, et plus largement durant toute ma formation, dès ma licence. Je lui suis extrêmement reconnaissant pour sa disponibilité, ses nombreux conseils, remarques, encouragements et la confiance qu'il m'a accordé durant toutes ces années.

Je remercie aussi sincèrement Martine Adda-Decker et Jean-Yves Antoine qui m'ont fait l'honneur de rapporter mon travail. Je leur suis profondément reconnaissant pour leur lecture attentive, critique et constructive de mon manuscrit.

Je tiens également à témoigner toute ma gratitude envers Jacques Vergne et Henri-José Deulofeu d'avoir accepté d'être examinateurs dans ce jury.

J'adresse tout particulièrement mes remerciements à Nuria Gala, Laure Brieussel et Corinne Zaoui pour leur aide, leur gentillesse, nos discussions fructueuses et leur disponibilité malgré toutes les fois où je suis venu les "embêter" :-P. Mon travail doit énormément à toutes les trois, et je ne saurai assez leur exprimer ma gratitude.

Un grand merci à toute l'équipe du Centre Informatique pour les Lettres et Sciences Humaines (CILSH) au sein de laquelle j'ai été successivement étudiant, tuteur puis moniteur. Merci à tous : doctorants, ingénieurs, secrétaires et autres personnels. J'ai une pensée particulière pour Jean-Luc Péris qui m'a fait découvrir la programmation au début de ma formation, et pour Gérard Della Ragione qui m'a encadré durant mon monitorat.

Un merci du fond du coeur à toute ma famille et mes proches, et particulièrement à mes parents Jean-Marie et Françoise, mon frère Hubert et ma belle-soeur Maria Emilia, et mon oncle Joseph pour leur inconditionnel soutien et qui ont toujours eu foi en moi. Un grand merci aussi à mon entraîneur et ami Mohamed. Ils m'ont sans cesse aidé à me dépasser et à avancer sans forcément toujours le savoir.

Et enfin, last but not least, à Marie pour son amour et sa patience. Merci de m'avoir écouté, porté, supporté mes remises en cause, et d'avoir été présente pour tout au quotidien. Il me reste toute la vie pour t'en remercier.

Sommaire

Table des figures

Liste des tableaux

Introduction

Contexte

Le traitement automatique de l'oral constitue depuis quelques années un domaine de recherche de plus en plus actif. En effet, à l'heure actuelle d'importants progrès ont été réalisés que ce soit dans les domaines de la reconnaissance vocale, de la traduction automatique ou encore des systèmes questions-réponses, etc. S'inscrivant dans le cadre général de la linguistique informatique, les recherches menées dans ce domaine ont pour objectif de développer des outils informatiques permettant l'automatisation des différents traitements linguistiques : analyse lexicale, syntaxique et sémantique, traduction, etc.

Pourtant, l'intérêt pour l'oral ne s'est déclaré que tardivement en raison des multiples préjugés dont il a été la cible durant de nombreuses années, au point que ce champ d'étude n'est pas encore totalement entériné. Reflétant une sociologie simpliste ou des conceptions pédagogiques naïves, l'oral a souvent été assimilé au parlé « populaire » ou considéré comme l' « anti-écrit ». De ce fait, l'oral a longtemps été mis à l'écart des études de la langue française qui ont préféré se focaliser sur les données écrites.

Par ailleurs, les systèmes développés dans le domaine des technologies de la parole n'atteignent pour l'instant leurs meilleures performances que dans des conditions très contraintes : mots isolés, domaine à vocabulaire limité ou à large vocabulaire mais avec une qualité sonore élevée et un environnement non bruité. De plus,

la qualité de la reconnaissance se dégrade très rapidement sur la parole spontanée, comme dans le cas de la transcription de flux audio en vue de l'indexation, d'applications de dialogue homme-machine ou encore de traduction parole-parole visant le grand public. Ces applications ont le plus souvent été abordées à l'aide de méthodes purement statistiques, mais il nous semble qu'un couplage avec des méthodes linguistiques pourrait être porteur d'améliorations sensibles.

Le traitement de la parole et le traitement automatique des langues (TAL) ont cependant des traditions encore relativement séparées. Le traitement de la parole est plutôt concentré sur le domaine acoustique-phonétique, tandis que les outils développés en TAL, et en particulier les analyseurs syntaxiques, sont principalement destinés à analyser l'écrit. Les recherches portant sur l'interface entre les deux domaines sont pour l'instant peu nombreuses, et seuls quelques travaux, tels que ceux de [Antoine *et al.*, 2003] ont tenté d'enchaîner la reconnaissance et les techniques du TAL, comme l'analyse robuste (chunking) ou la construction de dépendances sémantiques guidée par des besoins applicatifs.

Problématique

L'analyse syntaxique de l'oral est un projet généralement plus ambitieux que celle réalisée sur les textes écrits. Certains phénomènes rencontrés lors du traitement de l'écrit tels que le phénomène de listes apparaissent également à l'oral :

Ecrit

- *Faible voltage de circuit*
- *Faible résistance*
- *Situation à l'intérieur du compartiment moteur*
- *Non écrantement par du métal.*

([Gala Pavia, 2003])

Oral

> *alors évidemment il y a les parents il y a les évènements il y a l'école il y a la société il y a tout ce que vous voulez*
>
> (Corpus de Référence du Français Parlé)[1]

S'ajoutent à ce type de phénomène des problèmes propres au traitement de l'oral. En effet, l'une des particularités de l'oral est la présence importante de phénomènes de production (hésitations, amorces, répétitions, constructions interrompues, anacoluthes, etc.). Ces phénomènes sont souvent appelés « disfluences », et bien que ce terme semble évoquer une anormalité, les phénomènes concernés font pourtant partie des modes de production tout à fait normaux de l'oral, comme en témoignent les deux exemples suivants :

> *de se de se hum de se trouver ouais de se trouv- de se trouver*
> *le le tour de enfin le tour du pays Dogon une partie de une exploration un peu du pays Dogon*

La plupart des analyseurs syntaxiques conçus pour le traitement du langage naturel sont construits pour des données écrites qui répondent généralement à une grammaire clairement définie (la grammaire « standard », « normée »). Les entrées prévues pour ces analyseurs doivent donc être dépourvues de tout phénomène ne répondant pas aux structures syntaxiques classiques.

Cette constatation a mené les chercheurs à adopter, peu ou prou, deux grands types d'approches. D'un côté, des approches sélectives ont été développées, favorisant un prétraitement des données orales. Il s'agit alors de détecter et de supprimer les disfluences pour obtenir des données d'entrée proches de l'écrit. Et d'un autre côté, des systèmes ont été conçus pour prendre en compte les disfluences dans leur

[1]Sauf indication contraire, tous les exemples qui suivent seront extraits de ce corpus.

analyse tout en développant des procédures spécifiques à leur traitement. Le travail réalisé dans ce mémoire s'inscrit davantage dans cette seconde approche.

En relation avec cette problématique, le but de cette thèse est d'étudier de façon détaillée l'impact des disfluences sur l'analyse syntaxique automatique de l'oral, de proposer une analyse linguistique de ces phénomènes, et de fournir des mécanismes d'analyse syntaxique automatique permettant de les intégrer dans cette analyse.

Cette thèse vise ainsi à s'inscrire comme une étude intermédiaire entre des approches purement linguistiques d'une part ([McKelvie, 1998] ; [Candéa, 2000b]), et des approches informatiques d'autre part ([Carbonell et Hayes, 1983] ; [Goulian, 2002]).

Plan de la thèse

La première partie de cette thèse sera consacrée aux particularités intrinsèques de l'oral (chapitre 1). Nous proposerons ensuite une typologie détaillée des phénomènes de disfluence observés à l'oral. Divers auteurs ont abordé l'étude de disfluences particulières en français - par exemple [Pallaud, 2002] pour les amorces, [Henry, 2002b] pour les répétitions, [Candéa, 2000b], [Campione et Véronis, 2004] pour les pauses et phénomènes d'hésitation, etc. - mais une vision d'ensemble, montrant notamment l'interaction des différents phénomènes, fait pour l'instant défaut (chapitre 2). Dans le même ordre d'idées, le chapitre 3 sera dédié au recensement (non exhaustif) des corpus oraux existants, de façon à rendre compte de la différence entre les ressources disponibles pour l'oral face à celles disponibles pour l'écrit. Ce chapitre permettra également de décrire le corpus de travail utilisé dans notre étude. La présentation des différentes modélisations théoriques dont ont fait l'objet les disfluences dans de nombreux domaines (psycholinguistique, linguistique, informatique, etc.) au cours de ces dernières années terminera cette première partie (chapitre 4).

Nous introduirons ensuite la deuxième partie de ce travail par un état de la tech-

nique du traitement automatique de l'oral, en présentant les principaux outils développés jusqu'à présent (chapitre 5). Le chapitre 6 est destiné à la fois à positionner notre approche et notre vision structurelle des disfluences. Il s'agira ensuite de présenter de la banque de données de représentations arborescentes créée (chapitre 7). Dans ce même chapitre, le corpus de travail servira de base d'observation pour l'analyse quantitative des différents phénomènes de production.

La dernière partie sera consacrée au développement d'un analyseur syntaxique de surface pour l'oral, essentiellement focalisé sur les disfluences. Il s'agira tout d'abord de rappeler quelques travaux ultérieurs que nous avons menés sur un analyseur existant (chapitre 8), avant de présenter les différentes procédures d'analyse automatique mises en oeuvre (chapitre 9). Nous détaillerons les trois phases principales qui composent notre analyse : l'analyse morpho-syntaxique du corpus à l'aide d'un étiqueteur existant que nous avons adapté au cas des transcriptions orales, la détection « bruteé » des disfluences en préalable à la dernière étape de regroupement en syntagmes minimaux (ou chunks). Le corpus final sera ainsi segmenté en chunks de l'écrit d'une part, à côté des chunks disfluents d'autre part. Enfin, nous évaluerons de manière quantitative les résultats obtenus sur le corpus d'étude.

Première partie

Contexte de l'étude

Chapitre 1

Particularités générales de l'oral

1.1 Introduction

La langue écrite a longtemps été la seule préoccupation des études linguistiques. Néanmoins, cette tendance s'inverse depuis quelques années puisque l'oral et le dialogue prennent une place de plus en plus importante, tant dans le domaine de la linguistique que dans celui du traitement automatique des langues (technologies de la parole, apprentissage des langues, etc.). Les productions orales se caractérisent par un certain nombre de phénomènes et de particularités qui lui sont propres. Cette partie du mémoire a pour objectif d'en dresser un inventaire en y apportant des précisions terminologiques sur les concepts fondamentaux, avant de nous intéresser plus précisément au cas des phénomènes de production appelés également « disfluences ». Précisons, en outre, que nous privilégions dans notre étude un contexte énonciatif résolument monologique et narratif ; les exemples présentés s'incrivent donc systématiquement dans ce registre.

1.2 La notion de variation à l'oral

On peut constater que la variation est inhérente à l'oral, et de la même manière qu'il existe différents types de textes écrits, il existe également des types de productions orales. La variation linguistique, telle que nous la concevons ici, réside dans le fait

qu'une même réalité linguistique peut être exprimée de différentes manières.

1.2.1 Facteurs de la variation

Un certain nombre de facteurs conditionnent ainsi la variation à l'intérieur des productions orales. En effet, à l'instar de l'écrit qui présente une grande variabilité au niveau des formes d'écritures (romanesque, documentaire, etc.) et des types de productions (essai, lettre, article, roman, blog, sms, etc.), les productions orales ne sont pas toutes semblables. À la suite de [Deulofeu, 2004], on distingue le plus souvent quatre types de variations :

- La variation **diatopique** renvoie à la différence de prononciation en fonction de la région des locuteurs.
- La variation **diachronique** s'observe au niveau de l'âge des locuteurs et concerne essentiellement les aspects lexicaux et syntaxiques du langage.
- La variation **diastratique** correspond à la différence de classe sociale. À ce propos, [Deulofeu, 2004] précise qu'il est possible d'avoir deux conceptions de cette variation. La première tend à considérer que la classe sociale représente le facteur principal de variation ; la seconde envisage plutôt le nombre d'années de pratique d'une langue comme facteur de variabilité.
- La variation **discursive** renvoie à la situation discursive dans laquelle se trouve le locuteur. En effet, dans une situation d'énonciation publique ou privée et en fonction du rapport entretenu avec son interlocuteur, la production du locuteur sera différente.

Notons que les quatre types de variation illustrés ci-dessus ne sont pas exclusifs et peuvent parfaitement se combiner.

1.2.2 Degré de spontanéité des productions orales

Hormis les facteurs de variation identifiés plus haut, on peut remarquer que le degré de spontanéité observé dans certaines situations à l'oral peut également varier. En effet, toutes les productions orales ne se caractérisent pas par le même degré de spontanéité. Certaines contraintes ou situations définies à l'avance influencent le locuteur lors de la construction de son discours. Par exemple, lorsque la situation d'énonciation est déterminée au préalable, les productions orales vont présenter les caractéristiques d'un oral « orienté » et plus contraint. Ainsi, C. Blanche-Benveniste préfère utiliser dans ses travaux le terme de « français parlé » pour désigner les productions d'oral spontané. Elle emploie cette dénomination par opposition au terme générique d'« oral » relevant plus d'une taxinomie des productions orales : juridique, politique, pédagogique, etc.

Dans le cadre de ce travail, et afin d'éviter toute confusion terminologique, nous emploierons sans distinction particulière les termes d'« oral » et de « français parlé » pour désigner toute production de parole, quel que soit le degré de spontanéité observé.

1.3 Segmentation

1.3.1 Notion de « phrase »

Plusieurs études menées sur l'oral tendent vers un même constat : la délimitation des unités de base à l'oral est problématique. La notion de phrase telle qu'elle apparaît dans les définitions couramment proposées, ne repose pas sur ce que font réellement les locuteurs quand ils parlent, mais plutôt sur ce que [Beguelin, 2000] appelle des « *représentations construites et idéalisées de ces activités* ». Ainsi, le concept de « phrase » à l'oral (comme à l'écrit, d'ailleurs) relèverait plus d'une intuition que partagent les locuteurs en ce qui concerne l'unité phrastique. Pour illustrer cela, observons quelques exemples extraits de [Benzitoun, 2004]. Dans le

premier énoncé, l'unité phrastique d'une longueur importante inclut plusieurs unités indépendantes qui ne sont pas séparées par une ponctuation forte (uniquement des virgules). À l'inverse, dans le deuxième énoncé, la ponctuation et la segmentation en « phrases » qui en découle, impose un découpage des unités qui ne tient pas compte des relations de dépendance (*dans ce combat* étant sous la dépendance du verbe *accompagner*).

1)

> *Les armes c'est comme les voitures, ça coûte des vies humaines chaque année, mais c'est bien de les avoir quand même, voudriez vous vous passer des voitures, non et bien les armes c'est pareil, si les citoyens étaient armés les criminels ne seraient plus aussi nombreux, beaucoup d'entre eux sont des lâches, une balle ça fait réfléchir, les bons sentiments ça les fait rigoler, une balle c'est le début de la sagesse, je sais c'est triste je compatis avec vous, mais la réalité c'est ça, pouvoir se défendre quelque soit le prix a payer, sinon on va avoir droit à la jungle et ce système aussi il tue des innocents chaque année.*
>
> ([Benzitoun, 2004])

2)

> *C'est une grande épreuve pour toute la famille et je l'***accompagne** *encore aujourd'hui quotidiennement.* **Dans ce combat** *permanent pour la recherche de l'équilibre glycémique.*
>
> ([Benzitoun, 2004])

À la lumière de ces deux exemples, nous remarquons que l'unité représentée par la « phrase » peut poser des problèmes à l'écrit. Ce constat nous laisse entrevoir les difficultés que l'on rencontrerait si l'on transposait cette unité au français parlé. En effet, l'oral se caractérise souvent par des constructions inachevées, de multiples

retours en arrière, des commentaires épilinguistiques, etc. qui rendent difficile la segmentation en « phrases » telle qu'elle apparaît souvent à l'écrit. À ce propos, si on élargit la notion de phrase, [Gala Pavia, 2003] explique que cette unité peut être représentée à l'écrit par des structures aussi diverses que les listes ou encore les titres :

> "*Phrase : entité complète au sein du texte, c'est-à-dire entité ne pouvant pas être subdivisée en plusieurs suites ayant chacune une fonction particulière dans le texte.*" ([Gala Pavia, 2003] p. 56)

De même, la notion de phrase en syntaxe appelle volontiers celle de grammaticalité : les éléments qui composent une phrase doivent répondre aux règles de la grammaire par l'ordre des constituants tout en respectant la cohérence générale. Or, à l'oral, le locuteur construit son énoncé au fur et à mesure de son énonciation et y intègre des segments non prévisibles qui rendent parfois ses phrases agrammaticales (selon l'idée avancée ci-dessus).

De plus, on observe une inadéquation entre les signes conventionnels de l'écrit (majuscules, points, etc.) et la délimitation des segments syntaxiquement cohérents à l'oral. En effet, comment ponctuer l'énoncé suivant sans imposer de parti pris :

> *une balle qui n'avançait pas + comme à l'accoutumée*
>
> ([Blanche-Benveniste, 1990])

L'exemple renvoie ici au jeu de tennis de Yannick Noah. Si on insère un signe de ponctuation qui correspond à la pause (virgule, point-virgule, etc. codée ici avec le « + ») située juste après la négation, on devra comprendre au moment de la lecture que la négation porte sur le verbe « avançait » ; en d'autres termes, que les balles n'avançaient jamais. Or, il faut comprendre ici que la négation porte sur

« comme à l'accoutumée » : d'habitude les balles de Noah « avancent ».

La pause n'est donc pas dans ce cas précis une indication de limite syntaxique et aucune ponctuation conventionnelle ne peut la représenter.

Le recours à la prosodie semble être une alternative plus efficace pour délimiter les unités syntaxiques à l'oral que le recours aux signes de ponctuation de l'écrit. En effet, [Blanche-Benveniste, 1990] considère à ce propos que :

> "*La présence de tons finals dominants, avec effet de regroupement, aux frontières syntaxiques majeures, indique une correspondance entre structure syntaxique et structure intonative.*" (p. 173)

Ainsi, plusieurs auteurs ([Blanche-Benveniste, 1990] ; [Delais-Roussarie et Choi-Jonin, 2004] ; etc.) travaillant sur l'oral et sur sa syntaxe sont unanimes sur le fait que la notion de phrase n'a pas sa place dans ce contexte d'étude. La phrase, représente dans le meilleur des cas une approximation graphique résultant d'un compromis entre structure syntaxique, intonation et mise en page.

Pour ces différentes raisons et en rejoignant le cadre théorique formulé par Blanche-Benveniste, nous préférerons introduire le concept d' « unité maximale » pour désigner une unité de base en syntaxe de l'oral.

1.3.2 Notion d'Unité Maximale

Le concept d'unité maximale (UM) s'applique mieux à l'oral, tant dans les situations monologiques (monologues narratifs ou explicatifs) que dans les conversations où les structures de référence sont souvent non verbales (non, oui, d'accord, la semaine prochaine, etc.). Les « unités maximales » sont définies par [Benzitoun *et al.*, 2004] de la manière suivante :

> " *Constructions verbales, nominales, adjectivales ou encore adverbiales regroupant un élément tête* (...) [qui ne soit sous la dépendance d'aucune unité] *et les éléménts qui le suivent*".

Exemple :

> *j'ai tout refait à ma sauce pour que ça me corresponde et puis que puis que je sache où je vais en fait parce que bon euh j'ai pris des outils euh qui sont très bien faits*
>
> ([Benzitoun *et al.*, 2004])

Dans l'exemple ci-dessus, la conjonction de subordination *parce que* n'a pas sa fonction habituelle. Elle marque le début d'une nouvelle UM qui n'entretient aucun rapport syntaxique avec la précédente.

La plupart des travaux en traitement automatique des langues pour l'écrit considèrent la phrase comme une unité naturelle. Toutefois, nous l'avons vu ci-dessus, la pertinence linguistique de cette notion fait pourtant l'objet d'un vaste débat[1]. Même à l'écrit, phrases et unités maximales ne se correspondent pas toujours. [Benzitoun *et al.*, 2004] donnent un exemple illustrant cette idée :

> *Soyons direct,* || *après l'avoir fréquenté depuis des années, après l'avoir écouté pendant des heures au long de monologues sans fin, on aime Claude Got. — Et peut-être encore plus aujourd'hui qu'hier, alors qu'on l'accuse de vouloir imposer une « société sanitaire », sans plaisirs ni risques, ennuyeuse à mourir.*

Dans cet exemple, tiré du journal *Libération*, la première phrase contient deux unités linguistiques indépendantes, marquées par || (ces unités peuvent cependant être considérées comme dépendantes par « juxtaposition »). Par contre, là où le

[1] Voir à ce sujet : [Blanche-Benveniste, 1990] ; [Kleiber, 2003]

texte marque une rupture entre deux phrases (signe —), il n'y a qu'une seule unité linguistique en jeu (*encore plus aujourd'hui qu'hier* est le complément du verbe *aime*).

1.4 Organisation discursive des productions orales

L'organisation du discours constitue une autre particularité des productions orales. Elle mêle divers phénomènes déjà présents à l'écrit (énumérations, les constructions parenthétiques, etc.) et d'autres inhérents à l'oral (marqueurs discursifs).

1.4.1 Listes

L'organisation sous forme de « listes » formulée par [Blanche-Benveniste, 1990] est omniprésente à l'oral. À l'instar de [Benzitoun *et al.*, 2004], nous entendons par « liste » un ensemble de constituants qui occupent la même place syntaxique dans une unité maximale.

Exemple :

> *c'était complètement différent il y avait pas les voitures il y avait pas les les deux roues il y avait rien + il y avait pas cette modernité-là*

L'énumération des différents syntagmes verbaux (introduits par *il y avait*) présents dans l'énoncé ci-dessus peut être mise en valeur à l'aide de la représentation suivante dite de « mise en grille » que nous détaillerons plus loin dans notre étude (cf. 6.2) :

> *c'était complètement différent* **il y avait pas les voitures**
> **il y avait pas les les deux roues**
> **il y avait rien +**
> **il y avait pas cette modernité-là**

Ce phénomène de liste est largement répandu à l'oral mais existe aussi à l'écrit où il apparaît comme une configuration du discours observée dans divers domaines : scientifique, économique, technique, ou encore juridique (cf. [Gala Pavia, 2003]).

[Benzitoun *et al.*, 2004] rappellent qu'à l'oral les listes peuvent avoir diverses valeurs, et qu'une analyse en termes de coordination risquerait d'être trop aléatoire. Dans certains cas, elles peuvent également être des rattrapages de production (le locuteur répare), des ajouts d'information (le locuteur précise), des jeux de modalité (le locuteur oppose ou compare), des conclusions (le locuteur résume), etc. De plus, des « marqueurs » (cf. 1.4.3) viennent souvent introduire certains termes des listes (*enfin*, *pas*, *mais*, *et*, *je veux dire*, *donc*, *sinon*, etc.), comme le montrent les exemples suivants.

je dis les jeunes mineures
mais *aussi les jeunes majeures*

qui leur donnent une réelle indépendance + matérielle
je parle pas de *l'indépendance au niveau de + de la possibilité pour elles de*
mais au moins *de la de d'une indépendance matérielle*

ce n'est que de la ficelle
ou *de la corde*
ou *+ bon du*
soit en coton

La coordination apparaît comme un cas particulier de liste. Nous verrons plus loin qu'étant donnée leur fréquence, les listes ne peuvent être considérées comme un épiphénomène sans importance, mais au contraire comme un principe structurant de même nature que les relations de dépendance.

1.4.2 Constructions parenthétiques

La « construction parenthétique » ou « incise », phénomène également répandu à l'oral, est définie par [Roulet, 2003] comme "*l'insertion d'un segment de sens complet, au milieu d'un autre dont il interrompt la suite, avec ou sans rapport au sujet*".

D'autres auteurs tels que [Apotheloz et Zay, 1999] précisent qu'une construction parenthétique représente le "*lieu d'un traitement en parallèle de deux énonciations indépendantes*". Ces définitions, complémentaires, peuvent être illustrées par les exemples suivants :

> *par l'éducation que j'ai reçue que bon euh +* **pour parler un peu crûment** *et je dirais euh + le plus simplement possible je pense que tout le monde est bien là-haut*
>
> *et puis aussi euh gagner* **même si c'est pas grand chose** *c'est avoir été marqué de manière bénéfique par euh par le destin*
>
> *il y avait pas eu de réduction du temps de travail + hein + ben ça après c'est + euh +* **ça après bon on peut l'interpréter comme on le veut hein** *+ donc sur ce dossier là donc ce que vous pouvez noter*

[Blanche-Benveniste, 2000] – parlant d' « incidentes » pour désigner ce phénomène – signale à quel point il est étonnant de voir de quelle manière les locuteurs sont capables d'interrompre le fil syntaxique de leur discours, mettre en mémoire une partie déjà énoncée, placer la parenthétique et reprendre le fil de leur discours. L'usage massif de parenthétiques peut parfois donner l'impression que les locuteurs mènent simultanément plusieurs énoncés qui s'entrecroisent et où le discours ne représente plus une construction linéaire.

Il est également à noter que ces formulations peuvent souvent prendre la forme de commentaires méta-linguistiques qui illustrent le regard que porte le locuteur sur

son propre discours.

> *on aurait pu faire mieux* **comme on dit** *+ pour y aller*
>
> *on (n') a pas un métier de banquier* **au sens euh + administratif du terme** *on a un métier où on doit contacter les gens*
>
> *c'est là qu'on s'en rend compte + qu'on pourrait euh là encore* **je le mets entre guillemets** *+ presque conditionner + euh les jeunes*

Pour faire le parallèle avec les productions écrites, [Forget, 2000] considère qu'une construction parenthétique est généralement marquée à l'aide de marques typographiques : en général des tirets ou des parenthèses. À l'oral, en l'absence de ponctuation, elle se caractérise par un contour intonatif spécifique, mais peut aussi être signalée à l'aide de marqueurs dénominatifs (*i.e* des unités qui commentent l'énoncé dans lequel elles s'insèrent) comme dans les exemples suivants :

> **c'est quand même tout à fait intéressant + à ra- rappeler + en passant** *+ eh bien là-dedans il y a effectivement quelque chose qui est structuré*
>
> *il savait pas conduire quoi* **faut dire les choses claires et nettes** *+ c'était c'était pas ça*

Il est à noter que les points d'insertions des constructions parenthétiques correspondent souvent aux frontières des constituants syntaxiques majeurs. Les exemples qui suivent, illustrent bien cette correspondance entre l'introduction de constructions parenthétiques et les frontières syntaxiques :

– Entre le **syntagme nominal** et le **verbe**

> *et* [l'équipe de Cognac] **je crois que c'était en 1924 je n'en suis pas sûr mais je crois** [est arrivée] *en demi-finale des Jeux Olympiques pour les avirons*

– Entre le **verbe** et son **complément**

tout le monde [faisait] **j'en ai fait moi même** [de l'aviron] *j'avais été* [désigné] **je le savais** [comme un otage]

– Entre la **rection** et le **verbe recteur**

il disait [déjà à l'époque] **c'est-à-dire au début du siècle** [il disait] *pourquoi + au lieu de faire la guerre* (...)

1.4.3 Marqueurs discursifs

Une autre caractéristique de l'oral est l'omniprésence de mots ou locutions tels que *hein*, *bon*, *ben*, *quoi*, *tu vois*, *tu sais*. D'un point de vue terminologique, on retrouve ces unités dans la littérature sous des dénominations aussi diverses que « particules discursives », « inserts », « petits mots du discours », « petits mots de l'interaction », « connecteurs », « ligateurs », « ponctuants », etc.

Dans son article, [Chanet, 2003], regroupe sous la dénomination de « marqueur discursif » à la fois ce que la littérature pragmatique nomme les « connecteurs » (*mais*, *donc*, *parce que*, etc.) mais également les « particules discursives » (*bon*, *ben*, *voilà*, etc.). Ces deux types d'unités ont en commun :

> " *le fait de constituer des unités non référentielles et le fait d'agir sur les représentations cognitives construites par le discours, et dans la construction de ces représentations* ". (p.85)

A l'instar de l'auteur, nous préférons employer le terme « marqueur discursif » qui englobe différentes unités linguistiques comme nous le verrons ci-après.

Il est délicat de définir d'emblée ce que l'on entend par « marqueur discursif » car une telle définition pose le problème de la méthodologie adoptée et varie selon l'étude. Loin de faire l'objet d'un consensus terminologique, le terme de « marqueur discursif » englobe également à une multitude de formes et de catégories qui en

font une classe d'unités au contour flou et catégoriellement hétérogène.

D'un point de vue syntaxique, les marqueurs apparaissent comme des unités qui, dans le discours, n'entrent dans aucune construction syntaxique, tout en étant néanmoins rattachés prosodiquement au syntagme (ou plus largement à l'unité maximale) dans lequel ils s'insèrent.

> **donc** *euh* **ben** *ça ça se fait pas* **en fait**
>
> *un beau jour* **bon ben** *finalement je vais acheter une voiture*
>
> *j'ai envie un peu de +* **tu vois** *d'être dehors et de m'amuser* **quoi**
>
> *des week-ends de fous on s'en est fait tellement +* **tu sais** *c'est des week-ends entiers*

D'après [Chanet, 2003], « connecteurs » et « particules » fourniraient des informations sur la façon dont les interactants peuvent co-construire des représentations, les modifier et les ajuster les unes aux autres.

Dans une optique davantage syntaxique, [Teston et Véronis, 2004] préfèrent parler d'éléments « non régis » pour définir les marqueurs discursifs. Ces unités relèveraient par conséquent d'une analyse syntaxique et pas seulement pragmatique. Les éléments dits « non régis » se caractérisent par le fait qu'ils ne s'intègrent ni structurellement ni fonctionnellement dans l'énoncé. N'étant pas rattachés au verbe recteur, ils n'entrent donc pas dans les relations de dépendance. En effet, on peut tout à fait les supprimer sans que cela change pour autant le sens de l'énoncé. Pour illustrer cette idée, prenons les exemples suivants :

> *non* **franchement** *j'ai pas été malheureuse*
>
> *je pense qu'à ça + donc euh +* **je veux dire** *je n'ai + je n'ai pas beaucoup de centres d'intérêt*

Les éléments figurant en gras dans les exemples ci-dessus, s'opposent par leur

comportement syntaxique aux éléments dits « régis » des exemples suivants :

> *si je vous dis ma pensée + euh* [en parlant] **franchement** *je vous dirais + depuis que la gauche en est il y a beaucoup plus de problèmes que quand il y avait la droite*
>
> *ce n'est pas en la rendant obligatoire + que nous allons réussir + à la favoriser +* **ce que je veux dire** [c'est] *que lorsqu'elle était interdite*

Nous avons vu que la définition des marqueurs discursifs soulève quelques difficultés selon l'approche adoptée : syntaxique ou pragmatique. Outre cette difficulté, une multitude de formes a un usage assimilable à celui d'un marqueur discursif. [Piu, 2006] définit les types d'unités pouvant êtres regroupées sous le terme de « marqueur discursif ». Cette typologie peut être illustrée à l'aide du schéma suivant :

– Les « connecteurs » : *mais*, *donc*, *parce que*, *puisque*, *aussi*, *car*, etc.

  Dans leur emploi discursif, les connecteurs sont considérés par [Chanet, 2003] comme des items reliant une information produite dans le discours à l'ensemble des représentations mentales antérieurement construites par ce discours. Ils n'agiraient donc pas au niveau local mais au niveau global en assurant un rôle de structuration du discours. Voici par exemple, la forme *parce que* qui sert à relier deux propositions (niveau local) en (a) et qui a un emploi de marqueur discursif (structuration du discours) en (b) :

  > (a) *on retourne à l'hôpital euh quelques heures euh voir le nouveau matériel euh se faire expliquer quand même telle ou telle chose parce que sinon on a l'impression que l'information passera pas*
  >
  > (b) *parce que bon tu en as qui apprennent sur le tas bon euh elle elle apprendra comme moi sur le tas hein*

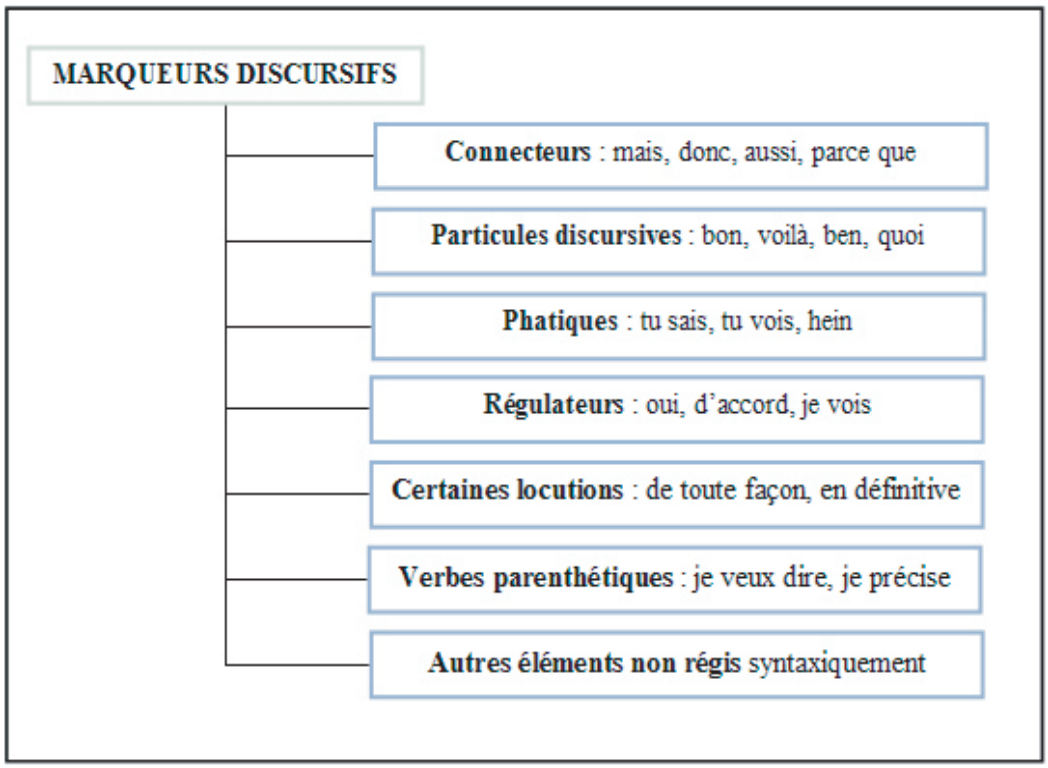


FIG. 1.1 – Types de marqueurs discursifs.

Il est à noter qu'un même connecteur peut assumer plusieurs valeurs sémantiques ou pragmatiques dans la mesure où son interprétation est négociée par le contexte dans lequel il est employé.

(c) *j'aime la cuisine donc j'aime bien manger*

(d) *euh donc euh on a une bonne entente on s'entend bien on rigole bien*

Dans les deux exemples qui précèdent, la forme donc n'a pas la même valeur sémantique. En (c), elle établit une relation de conséquence (le fait d'aimer manger est la conséquence du fait d'aimer faire la cuisine). Alors qu'en (d), la forme donc n'exprime pas de rapport de sens véritable avec ce qui précède puisqu'elle figure en position initiale, en tête d'énoncé.

– Les « particules discursives » : *bon*, *voilà*, *ben*, *là*, *quoi*, etc.

Ces unités ont pour rôle essentiel de fournir des informations sur les opérations conduites par le locuteur dans la construction de son discours. Il n'est pas toujours évident d'identifier automatiquement ces unités dans la mesure où bon nombre d'entre elles sont homonymes de catégories grammaticales. C'est le cas notamment des formes *voilà*, *bon* et *là* qui peuvent avoir plusieurs emplois différents selon le contexte :

(e) *voilà un peu comment ça s'organise*

(f) *donc c'est sûr beaucoup de la clientèle de quartier euh voilà*

Le comportement syntaxique de *voilà* est différent dans ces deux exemples : élément recteur d'une construction dans l'exemple (e), il est employé comme marqueur discursif dans l'exemple (f). Dans l'exemple (g) bon fonctionne comme un adjectif épithète tandis que son emploi en (h) est celui d'un marqueur discursif.

(g) *on essaie euh de prendre la vie du bon côté*

(h) *c'est un moyen aussi de toucher les adultes bon ça passe par les enseignants*

Il en va de même de la forme *là* dans l'exemple (i) qui est régie dans l'énoncé (construction en c'est...) et qui n'est pas régie dans l'autre exemple (on pourrait très bien supprimer cette forme sans changer le sens de l'énoncé).

(i) *c'est au niveau de l'inconscient que ça se passe et c'est là où ça s'inscrit*

(j) *en fait euh avec cette chienne je me suis sentie euh là ben pff disons capable de bouger toute seule*

– Les éléments « phatiques » et « régulateurs » : *tu sais*, *tu vois*, *hein*, etc.(phatiques) et *hm hm*, *oui d'accord*, *je vois*, etc. (régulateurs).

Les éléments « phatiques » sont émis par le locuteur et les éléments « régula-

teurs » sont quant à eux émis par l'allocutaire, le rôle de ces deux unités étant d'assurer le feed-back interactionnel.

(k) *c'est aussi former les hommes de demain hein c'est transformer les mentalités pour demain*

(l) *ah oui oui tout à fait + on essaie quand même de se partager la tâche hum de façon cohérente avec ma collègue*

– Certaines locutions : *de toute façon*, *en fin de compte*, *en définitive*, *de toute manière*, etc.

[Chanet, 2003] considère que certaines locutions peuvent avoir un emploi de marqueur discursif à l'oral. C'est le cas notamment de la locution *de toute façon* dans l'exemple qui suit :

(m) *mais je veux dire euh j'avais tellement soif de cette liberté-là et pour moi cette liberté elle passait parce voyage et ce voyage en stop parce que de toute façon bon j'avais pas d'autre possibilité sans doute*

– Certains verbes parenthétiques : *je veux dire*, *je sais pas*, *je précise*, etc.

Ces verbes à l'oral sont employés non canoniquement et forment des unités de communication complètes[2]. Il est à noter que ces verbes expriment souvent des états cognitifs tels que *penser*, *croire*, *supposer*, *savoir*, etc.

(n) *alors une théorie du désir et une théorie du lien objectal je précise*

(o) *enfin je suis partie donc euh je sais pas le vingt-cinq décembre ou le vingt-six un truc comme ça*

(p) *le : cépage de : euh la ville de Beaune je veux dire*

– Les autres éléments « non régis »

Cette catégorie regroupe les autres éléments « non régis » syntaxiquement et qui

[2]Voir plus haut : 1.4.2

n'appartiennent pas à l'une des catégories susmentionnées. Ces éléments comprennent par exemple les adverbes qui n'assurent pas leur fonction syntaxique habituelle et qui fonctionnent comme des marqueurs discursifs :

(q) *nous avons quand même des intervenants extérieurs que nous font veeuh nous faisons venir pour des sujets justement un petit peu pointus et parce que c'est intéressant aussi euh pour les élèves d'avoir un autre regard euh dessus*

(r) *en hiver euh il y a plus euh bon comme en France sans doute dans les coins touristiques hein*

1.5 Constructions syntaxiques

En ce qui concerne les études sur la parole, un certain nombre de travaux a porté sur les aspects phonétique et phonologique. Toutefois, la syntaxe, discipline fondamentale en linguistique, a longtemps privilégié l'écrit et les données textuelles.

Ce n'est qu'à partir des années 1990 que les aspects syntaxiques de l'oral ont été mis en avant de manière précise dans les études linguistiques. [Blanche-Benveniste, 2000] considère que l'oral possède une syntaxe particulière et rejette l'idée selon laquelle elle serait incohérente comparativement à celle de l'écrit. Au lieu de parler d'incohérence de la syntaxe à l'oral, il vaut mieux considérer qu'il s'agit d'une syntaxe propre, souvent moins stricte que celle de l'écrit. En effet, on ne peut pas faire porter sur l'ensemble de la langue parlée des caractères d'incohérence typiques de certaines situations comme celle des conversations à bâtons rompus par exemple. De plus, au regard de la grande diversité des genres et des registres de parole, on pourrait même parler de « syntaxes » de l'oral.

Les divers modes de production de la langue parlée représentent en fait de pré-

cieuses indications sur la structuration syntaxique. Il est effectivement possible de voir fonctionner à travers les hésitations et corrections produites par les locuteurs, certains processus généraux de fabrication des syntagmes.

1.5.1 Syntaxe verbale

Nous utilisons le terme de « rection verbale » pour qualifier des relations entre le verbe et les éléments qui se trouvent sous sa dépendance (sujets et compléments) ou plus généralement de l'ensemble des éléments régis par le verbe. Ce terme semble plus approprié pour l'étude des productions orales, comme le souligne [Blanche-Benveniste, 1990] :

> "*Comme le terme de « construction verbale » est très ambigu, nous utilisons celui de « rection verbale » pour désigner les relations entre le verbe et les éléments qu'il organise (sujet et compléments), de verbe « recteur » et d'éléments « régis ».*" (p. 40)

Exemple :

> *ils sont vraiment immenses*

Dans cet exemple, le pronom *ils*, l'adverbe *vraiment* et l'adjectif *immenses* sont sous la rection du verbe *être*.

Notons que [Blanche-Benveniste, 1990] s'appuie sur l'approche pronominale[3] afin de rendre compte des différentes rections verbales possibles. Soulignons également à la suite de cette idée, que nous considérons uniquement la valence d'un verbe[4] (éléments caractérisant la construction d'un verbe, et qui apparaissent comme « indispensables » à sa formation) comme une sous-partie de la rection.

Le domaine de la syntaxe verbale à l'oral peut faire référence à un grand nombre de procédés, appelés « dispositifs de rection ». Ceux-ci renvoient aux différentes asso-

[3] cf.[Blanche-Benveniste *et al.*, 1984]
[4] Voir à ce sujet [Willems, 1981]

ciations possibles entre le verbe recteur de la construction verbale et les éléments qu'il régit, comme nous le verrons dans ce qui suit.

Dispositif d'extraction : constructions clivées

On entend par « extraction » le fait de diviser la rection en deux parties. La première partie contient un élément de la rection du verbe isolé entre *c'est* et *que* ou *qui*.

Exemples :

c'est **eux** *qui vont venir vers toi* ⟶ **ils** *vont venir vers toi*

c'est **le travail** *que j'ai choisi* ⟶ *j'ai choisi* **ce travail**

Dans les deux cas présentés ci-avant, le pronom *eux* et le syntagme nominal *le travail* sont en position d'extraction alors qu'ils prennent respectivement la forme *ils* et *ce travail*, régis chacun par le verbe *venir* et *choisir* dans les énoncés où ils ne sont pas « extraits » (à droite).

Notons que la copule conjuguée *c'est* ne constitue pas ici un verbe recteur dans le dispositif d'extraction. De plus, cette forme varie de manière assez réduite : accord en nombre avec le terme extrait ou encore passage du présent à l'imparfait comme l'illustrent les exemples suivants.

ce sont *des vins* **qu'***on fabrique dans la région*

ce sont *les natures mortes* **qui** *me demandent le moins de de souci de construction*

c'était *une place* **qui** + *avait en son centre des étendards*

Dispositif d'extraction : constructions pseudo-clivées

Une autre construction typique de l'oral est celle de la construction dite « pseudo-clivée » qui, de même que la configuration précédente, a pour effet de diviser la formulation du locuteur en deux parties. Par exemple :

ce qui *m'a mis le pied à l'étrier justement* **c'est** + *c'est d'avoir commencé sur les malles*

ce qui *est très beau aussi dans ces deux Bruegel enfin dans ce Bruegel-là particulièrement* + **c'est** *la composition*

ce que *je préfère* **c'est** *justement de tout faire quoi*

Ces exemples montrent qu'elles ressemblent de près aux formes clivées mais elles s'en distinguent par le fait que la première partie de l'énoncé est réalisée de façon suspensive, pour créer une attente. Un des éléments régis (ici *ce qui* ... ou *ce que* ...) est réalisé sans forme lexicale qui est attendue lors d'une réalisation ultérieure – plus ou moins éloignée – sous forme de lexique (dans nos exemples : *d'avoir commencé sur les malles*, *la composition*, *justement de tout faire quoi*)

Signalons que ce procédé peut s'avérer récursif, combinant à la fois structures clivées et pseudo-clivées :

ce qui *m'a paru dommage* **c'est que** *visiblement même à la cantine*

ce qui *est stupéfiant quand même* **c'est que** *dans la famille tout le monde le savait*

ce que *je proposerais au comité de quartier* + **c'est que** *nous fassions une commission*

Aspects topologiques et double marquage

La topologie à l'oral renvoie à l'ordre selon lequel les mots sont agencés au sein de la phrase. La topologie permet généralement de connaître la fonction d'un argument selon sa position par rapport au verbe.

Par exemple, le français est une langue à ordre SVO (Sujet Verbe Objet). À l'écrit, l'ordre standard est, dans la plupart des cas, respecté. Selon les langues, cet ordre peut varier de fixe à totalement variable. De même dans certains cas à l'oral l'ordre

peut varier ([Antoine et Goulian, 2001]). Les énoncés suivants illustrent parfaitement des productions possibles dans une conversation parlée :

mon travail *je le réussis du premier coup*

⟶ antéposition d'un SN : OSV.

dans les bennes de camions *il y avait toutes les motos qui étaient entassées*

⟶ antéposition d'un SP : OSVO.

moi *Michelin je connaissais pas*

⟶ double marquage : SOSV.

([Blanche-Benveniste, 1990])

Le dernier exemple présente une structure en « double marquage » (ou dislocations) ; celle-ci est en fait un cas particulier d'élément « associés » [Blanche-Benveniste, 1990]. Les associés constituent des éléments à l'apparence similaire aux compléments régis par le verbe, mais qui pourtant n'en ont pas les propriétés : ni paradigme, ni correspondance avec les modalités du verbe, etc. et qui ne sont pas sans rappeler certaines caractéristiques des marqueurs discursifs.

Les cas les plus communs de dislocations sont les unités possédant exclusivement ce rôle d'associés comme dans les exemples suivants :

et **d'ailleurs** *ça c'est un truc dont on parle beaucoup à la fac*

évidemment *c'est déjà une ville + très pavillonnaire*

vous serez obligé de faire des manips + ou **en tout cas** *des actions de travail*

Les cas de double marquage renvoient en revanche à des associés dont la forme est tout à fait conforme à la rection du verbe auprès duquel il se trouve, alors que cette rection est déjà assurée par un pronom clitique ; il y a ainsi deux réalisations simultanées de la même rection, la rection est doublement marquée par deux éléments appartenant à des catégories différentes et dans le même paradigme.

Les cas de double marquage renvoient en revanche à des associés dont la forme

est tout à fait conforme à la rection du verbe auprès duquel il se trouve, alors que cette rection est déjà assurée par un pronom clitique ; il y a ainsi deux réalisations simultanées de la même rection, la rection est doublement marquée par deux éléments appartenant à des catégories différentes et dans le même paradigme.

Le cas le plus évident est celui où la rection d'une construction verbale est réalisée par un pronom clitique ainsi qu'une unité lexicale.

> des coutumes *il* **en** *existe encore*
>
> au bas de la rue *j'***y** *suis restée*

[Blanche-Benveniste, 1990] parle d' « étalement de paradigme » dans le sens où les deux unités appartiennent à un même paradigme. Notons que cet étalement n'a pas le statut d'une rection ordinaire.

| *en*
| *des coutumes*

Il convient de signaler qu'il existe des exemples de « semi » double marquage : la rection peut parfois ne pas être totalement réalisée par le ou les unité(s) lexicale(s).

> le tarot *qu'est-ce que vous* **y** *développez*
>
> la réforme de de la justice *vous* **en** *parlez abondamment*
>
> *bon* les troncs *je vous* **en** *parle pas*

Pour que la rection soit totalement effective, on s'attendrait dans ces exemples à avoir respectivement :

> **dans** *le tarot (qu'est-ce que vous y développez)*
>
> **de** *la réforme de de la justice (vous en parlez abondamment)*
>
> **des** *troncs (je vous en parle pas)*

Dans le même ordre d'idée, [Blasco-Dulbecco, 2004] – qui emploie le terme de « dislocation » – s'est intéressée plus particulièrement aux structures de type *moi*

(. . .) je :

> *ben je vous ai dit* **moi j'***ai eu un magnétophone à l'âge de douze ans*
>
> **moi** *qui suis au au delà du SMIC par exemple +* **je** *gagne douze mille francs*

Ici la structure de double marquage est représentée par le pronom disjoint mis en emphase (*moi*) avec le pronom conjoint sujet correspondant (*je*). Il en est de même avec d'autres pronoms, comme dans les exemples suivants :

> *et puis* **eux + ils** *se sont arrêtés de jouer moi je continue à + moi je continue euh + toutes formes de sculpture*
>
> *enfin* **toi tu** *es de Caen je crois*

Les relatives

La relative est considérée comme l'un des principaux exemples de divergence entre l'oral et l'écrit ([Kurdi, 2003]). Une proposition relative correspond à une proposition qui contient un pronom relatif enchâssé dans le syntagme nominal constituant d'une phrase dite principale. Le syntagme nominal qui sert de base à l'enchâssement est appelé antécédent. Soit l'énoncé :

> *des notions dont je me sers*

Le syntagme *des notions* est ici l'antécédent et *dont je me sers* correspond au relatif suivi de la principale.

Outre les formes dites « standards » utilisées à la fois à l'oral et à l'écrit (comme dans l'exemple ci-dessus), d'autres types de relatives peuvent être observés uniquement à l'oral comme par exemple les relatives dites de « français populaire ». [Blanche-Benveniste, 2000] rajoute que l'étude de ce phénomène a largement contribué à l'émergence de la notion de « syntaxe populaire ». L'idée étant que la structure des relatives est trop complexe pour le « peuple », qui l'utilise alors de façon

décomposée[5].

De telles relatives peuvent être réalisées de diverses manières comme dans les exemples suivants :

– Séquence avec un clitique :

c'est ma femme qu'elle s'occupe un peu de ça ([Blanche-Benveniste, 1990])
dans cette grange que je vous parle là

– Séquence avec un groupe prépositionnel

le prof que je parle de lui ([Kurdi, 2003]) *moi c'est la ma façon que j'arrive à m'investir*

– Séquence avec un pronom possessif

le prof que je parle de sa matière ([Kurdi, 2003])

Notons que l'on peut retrouver plusieurs exemples de relatives « non-standards » telles qu'illustrées ci-dessus, mais celles-ci sont généralement corrélées à l'emploi de certains verbes ; [Blanche-Benveniste, 2000] relève parmi les plus fréquents : *parler*, *avoir besoin*, *se servir*, *se souvenir*, *être content*.

1.5.2 Syntaxe nominale

Les particularités de la syntaxe nominale sont moins vastes que celles des constructions verbales mais apportent également une grande part d'informations sur la structuration du discours.

Accord en genre et en nombre

Il s'agit en français d'un mécanisme selon lequel un nom ou un pronom donné exerce une contrainte formelle sur les pronoms qui le représentent, sur les verbes dont il est sujet, sur les adjectifs ou participes passés qui se rapportent à lui ([Dubois *et al.*, 1994]). Selon les constructions, l'accord est plus ou moins respecté à

[5]Blanche-Benveniste parle de « décumul »

l'oral. Par exemple, le non-respect de l'accord entre le substantif et/ou ses adjectifs sont très rares (exemple a), alors que l'accord en genre entre l'attribut et le mot auquel il se rapporte est très fréquent (b et c). Voici une série d'exemples de non-respect de l'accord.

a) *des sommes exorbitants*

b) *dans la direction qu'il avait pris*

c) *c'est mes garçons*

Même en cas de respect de l'accord, ce respect n'a pas toujours de réalisations phonétiques perceptibles par l'auditeur de l'énoncé. Par exemple, le *e* utilisé pour marquer le genre féminin n'est associé à un phonème que dans des contextes exceptionnels comme lorsqu'il est précédé d'un *s* : *émise*, *admise*.

Syntagmes sans tête nominale

Cette catégorie renvoie aux cas où le syntagme nominal est décomposable en deux parties de la façon suivante : la première comporte une séquence ayant une valeur d'identification et de détermination, signalée par le recours aux pronoms ; la seconde partie est généralement constituée de la préposition *de* accompagnée du lexique nominal. Ces constructions peuvent s'illustrer à l'aide des exemples suivants :

tu prends **celle** aux anchois **de** pizza ([Blanche-Benveniste, 1990])

celle des élus **comme** Michèle Alliot-Marie

1.6 Conclusion

Nous avons tenté de présenter quelques traits « saillants » de l'oral tels que les problèmes de segmentation, les structures syntaxiques remarquables ou encore l'organisation discursive. Cet inventaire se veut plus illustratif qu'exhaustif dans la mesure où l'oral comprend de nombreux observables dont il aurait été difficile de rendre compte de manière complète. De plus, il convient de cibler notre propos en nous intéressant à notre objet d'étude.

Dans ce qui suit, notre tenterons d'esquisser une typologie des phénomènes de production à l'oral (ou « disfluences ») en adoptant une approche synthétique tenant compte des différentes études linguistiques menées sur le sujet.

Chapitre 2

Objet d'étude : les disfluences

Ces trente dernières années marquent le début des études linguistiques menées sur des corpus oraux [1] et par conséquent l'émergence de toute une problématique sur l'oral : à savoir comment traiter les phénomènes d'hésitation (ou disfluences) massivement présents dans les corpus oraux ? Les linguistes, les psycholinguistes ainsi que les spécialistes de la parole se sont penchés sur la question en observant la structure interne de l'oral et ses caractéristiques propres. L'objectif visé était de représenter et de modéliser ces phénomènes afin de dégager des régularités formelles remarquables.

En ce qui concerne le rôle de ces phénomènes dans la parole, plusieurs hypothèses ont été alors formulées : ces phénomènes ont été tour à tour considérés comme des hésitations, des marques du travail de formulation[2], des recherches dans la mémoire, traduisent-ils un malaise dans l'énonciation ?

Sans renier l'étude des processus cognitifs liés à la production de la parole (approche psycholinguistique), nous nous intéresserons davantage ici aux aspects syntaxiques des phénomènes de disfluences et aux contraintes auxquelles ils obéissent.

[1] On entend en général par « corpus oral » un corpus comprenant une transcription (orthographique ou phonétique) et un enregistrement sonore.

[2] [Morel et Danon-Boileau, 1998]

Malgré l'apparente irrégularité de ces phénomènes et leur diversité, ceux-ci peuvent être observés dans des corpus de même type, ou au sein d'une même langue, voire de plusieurs langues ([Clerc-Renaud *et al.*, 2004]). On peut également remarquer qu'ils répondent à des contraintes syntaxiques bien précises.

En effet, tel que nous l'avons exposé plus haut, diverses observations ont été dégagées durant ces dernières années, notamment au niveau de la variabilité linguistique à l'oral, de la notion de segmentation (la phrase n'est plus posée comme unité de base), de l'organisation discursive et de la syntaxe spécifique à l'oral, ainsi que les nombreuses disfluences qui apparaissent avec ce mode de production. Bien que sous certains aspects l'oral présente des similitudes avec l'écrit il reste néanmoins un mode de production ayant ses caractéristiques propres.

Pour comprendre de façon plus précise les difficultés auxquelles nous allons être confrontés et dans une perspective d'amélioration de l'analyse syntaxique de l'oral, nous souhaitons proposer une typologie des différents phénomènes regroupés sous le terme " disfluences " s'inspirant des nombreux travaux existants dans le domaine. Nous présenterons celles-ci en nous appuyant sur des exemples attestés relevés dans le Corpus de Référence du Français Parlé à partir duquel seront effectuées nos observations.

2.1 Problématique

Nous l'avons vu précédemment, les caractéristiques de l'oral sont nombreuses et on peut facilement imaginer à quel point leur traitement automatique peut s'avérer bien plus délicat qu'à l'écrit. Il reste alors un phénomène que nous n'avons pas encore évoqué puisqu'il constitue l'objet d'étude même de notre travail, à savoir

les disfluences.

[Blanche-Benveniste, 1990] explique qu'étudier les caractéristiques du français parlé, c'est

> " *étudier des discours généralement non préparés à l'avance. Or, lorsque nous produisons des discours non préparés, nous les composons au fur et à mesure de leur production, en laissant des traces de cette production.* ".

Le terme de « disfluences » regroupe un certain nombre de phénomènes spécifiques à l'oral : hésitations, répétitions, inachèvements, etc. Contrairement à ce que peu laisser penser la connotation quelque peu négative du terme, les disfluences sont des phénomènes tout à fait normaux et habituels de la parole spontanée, qui correspondent à la mise en œuvre en temps réel des structures de la langue. Pour [Blanche-Benveniste, 2003], elles sont souvent le reflet d'une séparation par le locuteur entre la syntaxe et le lexique : par exemple, les structures syntaxiques peuvent être déjà en place, sans que l'accès au lexique ait été totalement effectif, comme dans les exemples ci-dessous :

> *généralement* **les les les les les** *jeunes qui étaient là-bas + étaient d'un milieu assez aisé*
>
> *il y a* **beaucoup de beaucoup de** *préparatifs* **auxquels il a pas auxquels il a pas** *participé*

Ainsi, alors que les productions écrites représentent le plus souvent des produits finis, les exemples ci-dessus montrent clairement qu'une production orale s'élabore en même temps qu'elle est construite par le locuteur. Il n'existe donc pas un « avant-texte »[3] de l'oral comme à l'écrit.

Dans les échanges oraux les interlocuteurs ont peu conscience de ces phénomènes. [Blanche-Benveniste et Jeanjean, 1987] expliquent qu'on les remarque si peu que

[3]Terme emprunté à [Blanche-Benveniste et Jeanjean, 1987]

chacun d'entre nous est surpris d'en trouver dans les transcriptions de ses propres paroles. De plus, contrairement à ce qu'on pourrait penser on trouve autant de disfluences dans les transcriptions d'adultes (même chez les « professionnels de la parole » tels que journalistes, hommes politiques, etc.) que dans celles d'enfants.

Il reste difficile de savoir pourquoi nous percevons peu ces phénomènes à l'oral alors qu'ils apparaissent de façon incontournable à l'écrit. Il convient de rappeler qu'on ne les voit pas à l'écrit. En effet, le texte écrit publié, en tant que produit fini, n'a pas été mis au point du premier jet. Avant la version finale, l'écrit comporte des étapes similaires (jusqu'à un certain point) aux disfluences qu'on trouve à l'oral : les « avant-textes », ou brouillons, des auteurs.

Toutefois, à l'oral les procédés ne sont pas exactement les mêmes dans la mesure où on ne peut pas raturer, ajouter des signes typographiques pour signaler le déplacement d'un bloc de texte à la place d'un autre etc. Mais le locuteur fait ce même travail de correction de son texte oral sous d'autres formes. [Blanche-Benveniste et Jeanjean, 1987] rappellent que certaines études portant sur de grandes séquences de parole spontanée ont tenté de mettre en évidence les mécanismes de l'erreur et de sa correction.

A l'oral il pourrait paraître facile d'identifier où sont les erreurs en imaginant que, comme à l'écrit, l'auteur perfectionne son discours. Par un système de rature, il indiquerait que c'est la dernière version non raturée qui est celle à garder. Or, il est assez délicat de trouver des indices clairs qui correspondraient à une intention explicite de « rature » du locuteur. [Levelt, 1983] a d'ailleurs montré que les mécanismes formels mis en place pour corriger une erreur évidente peuvent en fait réunir plusieurs intentions. Levelt précise alors que d'un point de vue linguistique, les corrections apportées sont de même type que les répétitions ou les coordina-

tions.

[Blanche-Benveniste et Jeanjean, 1987] expliquent que l'on trouve dans des textes littéraires des « retouches » étant interprétées comme des effets de style et qui ressemblent de près aux disfluences de la langue parlée. Cependant, les disfluences sont souvent considérées comme des accidents aléatoires, des faiblesses imputables au locuteur, n'ayant aucune valeur fonctionnelle ni grammaticale.

Pour les auteurs, il est impossible dans les productions orales " *de distinguer entre l'erreur manifeste et la retouche intéressante*".

Nous avons cependant observé dans le cadre de ce travail et dans diverses études antérieures (par exemple [Henry *et al.*, 2004] ; [Campione et Véronis, 2004], que les disfluences n'interviennent pas au hasard mais apparaissent en fonction de contraintes syntaxiques très précises. [Blanche-Benveniste, 2003] soutient même que les répétitions et les hésitations ont une réelle valeur fonctionnelle, servant d'indices à la mise en place des syntagmes par le locuteur. Les disfluences apparaissent comme relevant d'une forme de syntaxe particulière que [Martinie, 2000] désigne "*réinstanciation d'une place syntaxique*".

Avant d'étudier dans le détail les multiples phénomènes de disfluence, il nous paraît indispensable de faire le point sur le nom « disfluence » lui même. En effet, ce terme n'est pas consacré dans la littérature où chaque auteur ou équipe de recherche semble avoir adopté sa propre terminologie, générant ainsi une multitude de mots qui désignent tout compte fait le même événement langagier.

2.2 Précisions terminologiques

Les phénomènes de production de l'oral ou « disfluence » ne font pas l'objet d'un consensus terminologique. La terminologie employée est extrêmement riche, tant

dans la littérature francophone qu'anglo-saxonne. L'ensemble des termes suivants (la liste n'étant pas exhaustive) font référence à ce type de phénomène :

- Bribes, turbulences, marques / phénomènes de production de l'oral ([Blanche-Benveniste *et al.*, 1984])
- Dysfonctionnements de la parole, programmes interrompus ou différents, télescopages syntaxiques, temps d'hésitation ou d'élocution, ralentissements, légère modulation temporelle d'hésitation ([Barberis et Maurer, 1998])
- Distorsions ([Boufaden *et al.*, 1998])
- Achoppements à l'oral, scories, ratés ([Pallaud, 1999])
- Marque de réparation ([Martinie, 1999])
- Marque de travail de formulation, marquage de recherche de formulation ([Morel et Danon-Boileau, 1998] ; [Candéa, 2000b])
- Reprogrammations ([Beguelin, 2000])
- Phénomènes dits d'hésitation ([Candéa, 2000b])
- Phénomènes de performance propres à la langue parlée ([Valli et Véronis, 2000])
- Énoncés réparés, marques de réparation ([Martinie, 2000])
- Inattendus structuraux du français parlé ([Goulian *et al.*, 2002])
- Extragrammaticalités ([Kurdi, 2003])
- Marques liées à la mise en discours ([Benzitoun, 2004])
- Disturbance ([Kasl et Mahl, 1965])
- Fragmentation ([Allen et Guy, 1974])
- Non-fluency ([Hindle, 1983])
- Discontinuity ([Taylor et Cameron, 1987])
- Speech management ([Allwood *et al.*, 1989])
- Disfluency ([Lickley, 1994] ; [Shriberg, 1994] ; [Heeman, 1997] ; [Core et Schubert, 1999])
- Speech disfluencies ([Shriberg, 1994])
- Performance phenomena ([Hübener *et al.*, 1996])
- Speech repairs ([McKelvie, 1998])

– Disruptions ([Core et Schubert, 1999])

et sont autant de mots qui ont été proposés et utilisés dans la littérature pour désigner les phénomènes spontanés de l'oral. Chacun de ces termes a sa motivation selon l'équipe de recherche et les travaux dans lesquels ils sont employés.

Dans un souci de clarté, et parmi ces nombreux choix terminologiques, nous avons choisi de nous limiter à certains termes. Dans cette étude, nous emploierons alternativement le terme de phénomènes de production et celui de disfluences (*dis-* : séparation, négation ; *-fluence* : lat. *fluencia* « écoulement ») utilisé notamment par [Sabio, 1996] pour désigner de façon générale les phénomènes d'élaboration des productions orales. Même si cette dénomination semble évoquer une anormalité, les phénomènes que recouvre ce terme font partie des modes de production tout à fait normaux de l'oral (à ne pas confondre avec les *dysfluences* qui correspondent à des phénomènes d'ordre pathologique). Nous utiliserons ce terme pour garder l'idée d'un endroit dans l'énoncé où " *le déroulement syntagmatique est brisé* " ([Blanche-Benveniste, 1990]).

Une fois ces considérations terminologiques passées en revue, il convient de s'interroger sur deux points : Qu'englobe réellement le terme « disfluences » ? Quelle est la nature de ces phénomènes omniprésents en français parlé ?

2.3 Que regroupe la catégorie des disfluences ?

D'une façon générale, les disfluences ont fait l'objet de peu de travaux. Les études existantes dans le domaine s'attachent souvent à décrire un phénomène précis mais une vision d'ensemble fait pour l'instant défaut. A l'intérieur de la catégorie globale désignée par chacun des termes vus précédemment, les phénomènes sont plus ou moins nombreux selon les auteurs. En ce qui nous concerne, nous souhaitons préalablement reprendre la distinction établie par [Henry, 2002b] entre les pauses silencieuses et les autres disfluences. Bien que spécifique au français parlé, le phéno-

mène de pause silencieuse (ou pause non sonore, silent pause en anglais) n'apparaît pas comme disfluent mais comme un élément qui participe au contraire à la fluence du discours. [Boomer, 1965] lui reconnaît entre autre un rôle démarcatif apparaissant à la jonction de segments du discours, et qui participent à la structuration de ceux-ci.

Les pauses silencieuses sont donc nécessaires à la fois pour la planification des énoncés par les locuteurs, et pour le traitement par les auditeurs. Elles sont notées dans notre corpus à l'aide des signes « + ».

> *c'est clair + mais la chose la plus importante c'est que moi j'ai rééquilibré ma vie +*

Nous avons vu précédemment que le terme « disfluence » possède de nombreux équivalents dans la littérature. Il en est de même concernant les différents phénomènes appartenant à la catégorie des disfluences, où la terminologie pour y référer est extrêmement variée, tel que le montre le schéma ci-après (le terme mis en gras est celui retenu pour notre typologie). [Candéa, 2000a] remarquait justement que selon les auteurs et les objectifs de chaque étude, la terminologie sera généralement explicitée en détail et chaque phénomène particulier sera défini et désigné par un terme univoque.

Cette disparité empêche donc toute harmonisation dans la terminologie utilisée par les différents linguistes francophones ou anglophones qui se sont intéressés aux disfluences. Devant la multitude de termes employés pour désigner les phénomènes de l'oral, nous avons fait le choix d'utiliser certains termes plutôt que d'autres. Ceux que nous employons sont largement inspirés des travaux de l'équipe DELIC[4]. Nous exposerons donc les choix terminologiques que nous avons fait et nous donnerons une définition aussi précise que possible de chacun des phénomènes qui seront étudiés tout au long de notre travail.

[4]Description Linguistique Informatisée sur Corpus

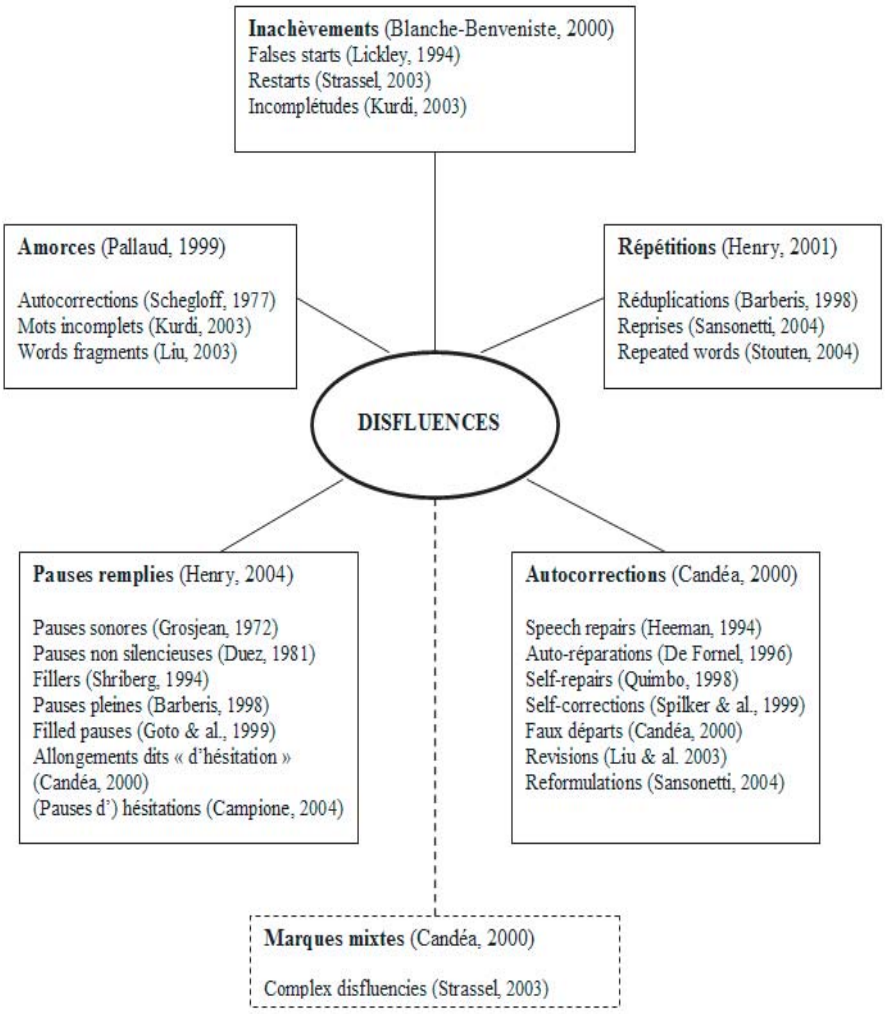


FIG. 2.1 – Phénomènes de disfluences et quelques équivalences terminologiques.

2.3.1 Pauses remplies

Les pauses remplies dans la production orale peuvent se manifester de diverses manières : soit par le recours à un morphème spécifique (le cas de la pause remplie *euh* ou *hum*), soit en prenant la forme d'un allongement de syllabe. Elles ne sont pas seulement des bruits parasites et insignifiants dans la communication : [Campione, 2001] considère au contraire qu'elles constituent des éléments à part entière du code linguistique. Elles renvoient généralement à des nécessités psy-

chologiques et cognitives pour le locuteur, à qui elles laissent le temps de trouver ses mots, une formulation, etc. Elles sont aussi un mécanisme de régulation de la communication et notamment de gestion des tours de parole.

Elles constituent en quelque sorte un signal conventionnel de la part du locuteur, lui permettant d'occuper le terrain de l'interaction et d'éviter d'être interrompu pendant le laps de temps nécessaire à la construction de la suite de son énoncé ([Campione et Véronis, 2004]).

D'après [Campione et Véronis, 2004], il y a deux sortes de pauses remplies : celles qui sont internes à un segment discursif, elles permettent de marquer une interruption suivies ou non d'une reprise et/ou d'une réparation (81% des cas) :

> **euh** *Beaune est une* **euh** *la la la euh le cé- le cépage de :* **euh** *la ville de Beaune je veux dire*

Dans 19% des cas, elle se situe en début de segment ; il ne s'agit alors pas d'une interruption liée à une difficulté de mise en place lexicale ou syntaxique ; il s'agit pour le locuteur d'occuper le terrain en attendant de trouver un suite au discours et éviter ainsi que l'interlocuteur prenne la parole.

> **euh** *et après bon ben après les choses se sont mises en place*

Le *euh* d'hésitation est une voyelle prononcée indépendamment, avant ou après un mot. Comme le précisent les études menées par [Candéa, 2000b], les marques dites d' « hésitation » n'ont commencé à être étudiées que tardivement, vers la fin des années cinquante et uniquement en anglais grâce à l'ampleur que commençait à prendre le recours à des corpus de parole naturelle. Ce phénomène apparaît comme le reflet de difficultés que rencontre ponctuellement le locuteur dans ses opérations mentales de « travail de formulation » ([Morel et Danon-Boileau, 1998]) liées à la production du discours.

La durée des voyelles peut varier fortement dans la parole d'un seul et même locu-

teur en fonction de divers facteurs. Ainsi arrive-t-il fréquemment qu'une voyelle soit significativement plus longue que les voyelles précédentes et/ou suivantes à l'intérieur d'une seule prise de parole, même après la correction des durées intrinsèques.

Les allongements syllabiques affectent généralement une voyelle en fin de mot. Selon [Candéa, 2000b], tout allongement d'un noyau vocalique anormal en position finale de mot ou d'amorce de mot, présentant un contour plat et bas ou très légèrement descendant et surtout non modulé pendant l'allongement, représente un allongement marquant le travail de formulation en cours. De plus, compte tenu de l'ordre des mots en français l'allongement final ne sera concrètement perçu comme marque de frontière ou comme pause que lorsqu'il porte sur un mot plein et jamais, par exemple, sur un prédéterminant.

On considère le plus souvent qu'il est anormalement allongé si sa durée est comprise entre 180 et 220 ms. L'allongement est généralement marqué par le signe « : » dans les conventions de transcription adoptées par [Blanche-Benveniste, 2000].

On a déjà noté que leurs propriétés sont proches de celles des *euh*, de telle sorte que les mêmes algorithmes permettent de détecter ces deux phénomènes (cf. [Goto *et al.*, 1999]). A la suite de [Goto *et al.*, 1999], nous regroupons également allongements et *euh* sous le même terme de pauses remplies du fait de leur similitude de propriétés et de fonctionnement.

Euh

> *j'en ai une autre qui est professeur +* **euh** *de physique chimie*
>
> *et en arrivant* **euh** *au sud du Portugal*
>
> *c'est un peu* **euh** *comme à la gymnastique*

Hum

> *et* **hum** *+ voilà vous êtes bordelais*

hum *en ce qui me concerne*

est-ce que **hum** *on peut faire du Judo en dehors de la compétition*

Allongements syllabiques

si c'était **:** *l'hiver ou* **:** *plus longtemps*

donc je suis partie en stop en **:** *+ en* **:** *à dix-huit ans*

ils ont fait des **:** *des blancs*

2.3.2 Amorces

Les amorces ou fragments de mots (tout comme les répétitions) constituent des événements langagiers d'une grande fréquence à l'oral. En effet, [Henry et Pallaud, 2003] montrent que sur un corpus d'environ 46 000 mots on trouve en moyenne quatre amorces tout les 1000 mots, soit pour un débit moyen de 200 mots par minute une amorce toutes les 75 secondes.

Bien que ce phénomène soit le plus souvent ignoré du locuteur et de son récepteur, - comme si ces phénomènes étaient non communicatifs - ils sont la marque d'une élaboration de l'énoncé. Leur présence témoigne d'une réflexion ou d'une activité linguistique se traduisant par une interruption de morphèmes en cours d'énonciation.

Par ailleurs [Kurdi, 2003] précise que bien qu'elles ne soient pas une fin en soi dans le traitement de la parole spontanée, les amorces (qu'il nomme « mots incomplets ») constituent un indicateur assez important pour la détection des autres disfluences (Kurdi parle ici d' « extragrammaticalités supralexicales » pour désigner les disfluences autres que les pauses silencieuses ou remplies et les amorces de mots, qu'il qualifie à l'inverse d'« extragrammaticalités lexicales »). Malheureusement, cette information n'est pas encore utilisable dans des conditions réelles, puisque les systèmes actuels de reconnaissance de la parole ne reproduisent pas les mots incomplets. Néanmoins, dans l'optique de nos traitements automatiques,

nous nous placerons dans un contexte où nous considérons notre corpus de travail comme une sortie de reconnaissance idéale.

Les amorces ne sont pas toutes semblables ; à cet égard [Henry et Pallaud, 2004] identifient trois types d'amorces. Lorsqu'il y a réduction définitive sur une place syntaxique donnée, l'amorce est dite inachevée et les unités qui suivent occupent une autre place syntaxique. Si l'élément suivant l'amorce ne change pas de place syntaxique, deux cas sont alors possibles : soit il y a piétinement sur la place syntaxique et une simple reprise puis poursuite du morphème inachevé (on parle d'amorce complétée) ; soit il y a une modification lexicale complète (on parle ici d'une amorce modifiée). Les exemples qui suivent permettent de distinguer ces trois catégories :

- Amorces sur un **même** emplacement syntaxique :
  - Complétées

    des **p-** *euh des* **particuliers**

    je suis restée trois mois en **cam-** *en* **camping**

    je **ju-** *je* **juge** *pas*

    le mental en **t-** **tout** *ce qui était dans dans la tête*

    Le mot commencé et interrompu se trouve complété. Par exemple, l'amorce se trouve complétée après la reprise du déterminant s'il s'agit d'un syntagme nominal, après la reprise de la préposition s'il s'agit d'un syntagme prépositionnel ou encore du pronom personnel sujet s'il s'agit d'un verbe. L'amorce peut également porter sur de simples unités lexicales et être complétée immédiatement sans reprise d'autre élément. Il s'agit donc des amorces où le locuteur complète finalement ce que, dans un premier temps, il n'avait qu'ébauché pour s'interrompre aussitôt.

  - Modifiées

    un **col-** *un* **ami** *également qui plonge souvent avec nous*

à moins **qu- de** *faire le Judo*

beaucoup de **cop-** *de* **camarades** *de mes petites filles*

Le locuteur ne complète pas ce qu'il avait commencé à dire mais « corrige » et poursuit par un autre élément.

Il convient de noter que dans un cas comme dans l'autre, il peut arriver que même si la place syntaxique est conservée, la catégorie morphologique ne soit plus la même : dans l'exemple suivant on passe de l'amorce d'un verbe (*pleut*) à un nom commun (*pluie*) :

en hiver au Portugal il **p-** *il* **p-** *il y a des moments de* **pluie**

- Amorce sur un emplacement syntaxique **différent** :
  - Inachevées

mais descends parce que **b-** *mais le problème c'est que le cerceau*

comme ça + on **tou-** *+ j'aurais pas assez de place pour le faire*

c'est pas un **sty-** *il avait pas la bonne mentalité*

Ce qui suit l'amorce occupe une autre place syntaxique ; on n'est donc pas en présence d'un piétinement sur la même place syntaxique. Ces amorces sont des lapsus interrompus et non corrigés par la suite.

Sur un corpus de plus d'un million de mots [Henry et Pallaud, 2004] recensent près de 60% d'amorces de type complétés, ainsi que 22% d'amorces inachevées et 18% de modifiées. Ces chiffres tendent ainsi à montrer que dans les trois quart des cas (amorces complétées + inachevées) l'amorce témoigne plus d'une disfluence que d'une erreur de langage (plus communément appelée lapsus). Toutefois, nous ne reprendrons pas cette distinction disfluence/erreur de langage dans la mesure où nous appréhendons les amorces uniquement en terme de disfluences.

2.3.3 Répétitions

Nous ne pouvons présenter ce phénomène sans nous appuyer sur les travaux de [Henry, 2002a] ; [Henry *et al.*, 2004]), concernant la distribution syntaxique et prosodique des répétitions en français parlé. Elle montre notamment l'intérêt de leur étude d'un point de vue linguistique et phonétique, car elles peuvent apporter des indices précieux pour la compréhension des structures et des modes de fonctionnement du langage. En effet, tout ne se répète pas, et surtout ne se répète pas n'importe comment.

D'un point de vue général, nous pourrions définir ce phénomène comme étant la répétition d'un ou plusieurs mots ou comme la reprise " à l'identique, (...) d'une syllabe, d'un mot ou d'une amorce de mot, de plusieurs syllabes ou de plusieurs mots, sans aucune valeur sémantique " (cf. [Candéa, 2000b]). Il convient cependant de distinguer les répétitions qui relèvent de la langue, de celles qui correspondent à des phénomènes de performance propres à l'oral (*i.e* disfluentes). Cette distinction peut être facilement réalisée pour des énoncés qui contiennent des suites impliquant des mots grammaticaux telles que *je je*, *le le*, etc. Toutefois, lorsque la répétition fait intervenir des mots lexicaux (mots qui ont une charge lexicale pleine) ou encore certaines formes de mots grammaticaux (par exemple, répétitions en nous nous, vous vous, etc.), cette catégorisation devient complexe.

C'est souvent le manque d'indices qui rend cette distinction difficile, par exemple, dans les suites en nous nous comme dans :

> **nous nous** *sommes appelés aussi* ([Henry, 2002a]).

Dans ce cas là, il n'est pas évident de trancher et de déterminer si ces suites correspondent à des retouches ou alors à des phénomènes d'emphase où la première occurrence serait dans ce cas un élément disloqué par exemple. La difficulté est la même avec les exemple suivants :

et **vous vous** *êtes ouvert ces jours-là*

ça ça *peut être intéressant*

et voilà quoi ça se passe **très très** *bien*

De plus, la répétition n'est pas systématiquement une redondance. Il est par exemple possible qu'elle soit simplement dotée d'une fonction communicative. C'est notamment le cas lorsque le locuteur n'est pas certain que son message sera correctement perçu par son interlocuteur, du fait d'une mauvaise articulation, de bruits paraverbaux, etc. le message sera répété. Kurdi (2003) évoque la fonction pragmatique que peut prendre la répétition pour marquer une affirmation, une négation, une insistance, etc.

oui oui *ça c'est une bonne remarque*

non non non *c'est pas ce que je voulais dire*

A la différence des répétitions « faits de langue » (i.e non disfluentes telles que les reprises de pronom personnel dans le cas d'un verbe pronominal. Exemple : *nous nous somme recontrés*) que l'on retrouve aussi bien à l'oral qu'à l'écrit, les répétitions disfluentes n'apparaissent qu'à l'oral. Leur présence aboutit à la formation d'énoncés agrammaticaux dans le sens où il est évident qu'aucune grammaire ne retiendra comme acceptables les énoncés suivants où le déterminant défini *le*, les prépositions *pour* et *de* sont répétés :

essayez de faire comprendre que **le le le le** *structuralisme*

je leur donne les clefs **pour pour** *se sortir* **de de** *là tout de suite*

Même si elles échappent aux règles grammaticales, ces répétitions disfluentes semblent toutefois obéir à certaines contraintes posées par la syntaxe. En effet, les répétitions n'apparaissent pas au hasard, mais essentiellement à l'initiale de frontières syntaxiques majeures, révélant ainsi dans un premier temps une incomplétude syntaxique en laissant un vide lexical ; le remplissage lexical s'effectue dans un second

temps pour aboutir à un énoncé achevé. Cette remarque est toutefois à relativiser dans le cas de dialogues fortement interactifs, où les échanges entre locuteurs peuvent mener à des productions moins "structurées".

Pour le cas des répétitions disfluentes, celles-ci ont principalement lieu sur les débuts de syntagmes et qu'elles font intervenir majoritairement des mots grammaticaux (structurant la langue). Trois critères suffisent alors à circonscrire le phénomène :

- la longueur du répétable : **nombre d'éléments** contenus dans le répétable.
- l'empan : **nombre de répétés** contenus dans la répétition.
- la **succession** des termes répétés : présence ou non d'autres éléments dans la production de la répétition.

[Henry, 2002b] distingue alors plusieurs catégories de répétitions :

- Les répétitions que l'on peut appeler « simples » (un seul élément contenu dans le motif répété) par rapport aux répétitions « complexes » (plusieurs éléments contenus dans le motif répété).

  Répétitions simples

  le Maire a trouvé encore **un un un** *terrain*

  ça fait encore partie **de de** *ma profession*

  Répétitions complexes

  l'évolution **de cette de cette** *enquête*

  oui **c'était c'était c'était** *plus attardé qu'ici*

- Les répétitions de type « uniques » (éléments répétés une seule fois) et les répétitions « multiples » (unités répétées plusieurs fois).

  Répétitions uniques

  euh **la la** *propriété qui venait nous dire bonsoir*

ils montent à Pau p**our pour** *aller chercher* **les les** *boissons*

Répétitions multiples

sur la tête **d'un d'un d'un** *homme*

pour vous situer **la la la la** *ville*

– Les répétitions que l'on peut qualifier de « continues » (termes répétés produits en contiguïté) par opposition aux « répétitions discontinues » (qui acceptent l'insertion d'un ou plusieurs autre(s) élément(s) entre les termes répétés).

Répétitions continues

c'est tous les systèmes **de de de de** *codes*

enfin bon **tout un tout un** *style de vie*

Répétitions discontinues

c'était euh + **c'était** *à l'hôpital*

ma grosse difficulté c'était **de** + euh ben **de** *pas me rendre malade*

Dans l'optique de traitements ultérieurs, et dans un souci de simplification typologique, ces différents types de répétitions peuvent être résumés en deux catégories principales : les répétitions simples (regroupant répétitions simples, uniques et continues) et complexes (regroupant répétitions complexes, multiples et discontinues) et ne seront pas nécessairement distinguées dans les modules de traitement automatique.

Dans l'ensemble de ces cas, le phénomène de répétition se présente donc comme un décalage entre la disponibilité immédiate des tournures syntaxiques en mémoire du locuteur, et les disponibilités moins importantes du lexique qu'il connaît. Le cadre syntaxique est posé dans un premier temps et le remplissage lexical intervient ensuite.

2.3.4 Autocorrections

Comparativement à d'autres disfluences, le fonctionnement syntaxique des autocorrections n'a été que très peu décrit. Pourtant, ce phénomène est similaire à celui de répétition, à la différence près qu'il y a substitution d'un ou plusieurs mots par d'autres, et que la portion corrigée permet de modifier ou clarifier de façon plus ou moins forte le sens de l'énoncé plutôt que de simplement le répéter totalement ou en partie. Ceci permet par exemple de corriger le genre, le nombre, le choix d'un nom ou d'un adjectif, etc.

lui ai donné **un une** *petite bouteille*

ça c'est les spécialités **que j'ai + qui demandent** *de l'apprentissage*

Par ailleurs, il convient de retenir que les autocorrections témoignent du contrôle que nous exerçons sur notre langage tout en le produisant [Blanche-Benveniste, 2003]. De même que les amorces et les répétitions, elles sont généralement peu sensibles pour les interlocuteurs. Or, l'autocorrection n'est pas complètement aléatoire et porte souvent sur un segment qui peut compter un ou plusieurs syntagmes [Core et Schubert, 1999], c'est pourquoi elle est souvent couplée à une répétition partielle du segment corrigé.

j'ai **des** + j'ai **un** *petit domaine de création*

Dans cet énoncé, l'autocorrection se fait en répétant le segment *j'ai* et en remplaçant le mot *des* par le mot *un*. On note que les deux mots ont la même catégorie morphologique (article défini) et la même fonction syntaxique (déterminant).

[De Smedt et Kempen, 1987], cités par [Fornel et Marandin, 1996], distinguent trois niveaux d'autocorrections (qu'ils nomment « réparations ») à partir de la catégorie du segment corrigé ; le phénomène n'est alors pas désigné de la même manière :

– Lemma substitution : Correction au niveau **lexical**

j'ai pas **le la** *passion*

mais + **à sur** *Limoges dès la première année*

– Reformulation : Correction au niveau **syntagmatique**

je lui ai **écrit** *je lui ai* **dit**

j'avais + j'étais *pas à l'aise avec lui*

– Restart : Correction au niveau **phrastique**

moi c'est vrai que je travaille dans **le enfin** *je suis née + de parents de commerçants*

donc nous avons un service euh + qui **s'occupe** *euh* **ce** *que nous appelons le service investigations*

Dans le cadre de la présente étude, nous désignerons « autocorrection » les phénomènes qui interviennent aux deux premiers niveaux. Nous réservons plus loin le terme « inachèvement » pour ce que [De Smedt et Kempen, 1987] nomment « restarts ».

Cette distinction rappelle que, de la même manière que pour les répétitions, l'autocorrection peut porter sur des unités plus ou moins complexes. A cet égard, [Candéa, 2000b] considère que les autocorrections ne sont pas toutes semblables et en distingue deux types : les autocorrections « immédiates » (a) et les autocorrections « complexes » (b). La différence entre ces deux types d'autocorrections réside dans le fait que la réparation réalisée par le locuteur (c'est à dire le mot ou le syntagme conservé au final) ne se produit pas au même endroit dans l'énoncé.

a) **mon un** *vieux collègue de sciences naturelles*

mais tu peux pas **faire mettre** *de l'eau froide sur n'importe quel tissu*

b) *ben parce que le Charlemagne euh + paraît-il ne buvait que des* **rouges** *parce qu'il voulait pas il il se tâchait sa sa il ne buvait que des* **blancs** *pardon*

À la lumière des deux derniers exemples on s'aperçoit que l'autocorrection peut s'étendre sur plusieurs mots et occuper une grande place dans les transcriptions.

Il est à noter que dans l'autocorrection, le travail de dénomination occupe une place importante et s'accompagne le plus souvent de commentaires explicites. Le locuteur évalue explicitement la bonne adéquation des mots qu'il a choisis en les révoquant pour finalement les modifier. A cet égard, il convient de distinguer deux configurations principales d'autocorrections :

> Reprise avec enrichissement lexical, où chaque nouveau segment de l'entassement paradigmatique vient préciser la pensée du locuteur.
>
> *il faut peut-être se spécialiser dans* **l'audio** *enfin dans* **le cinéma l'audio-visuel**
>
> Correction à proprement parlé, où le locuteur tient à corriger (et non à préciser) une séquence d'origine qui – de son point de vue – est visiblement erronée.
>
> *bon ici on met un caniveau* **pas un caniveau un tuy- une buse** *en bas*

Les différentes configurations d'autocorrection ressemblent ainsi quelque peu à celles que l'on trouve pour les répétitions.

Comme nous l'avons vu pour les amorces, il se peut que la place syntaxique soit conservée mais que la catégorie morphologique ne soit plus la même. On trouve ainsi par exemple des verbes (*recrute*) corrigés par le substantif correspondant (*recrutement*) :

> *donc et* **recrute le recrutement** *s'effectue euh uniquement par euh par voie de concours*

Cet exemple permet d'identifier le mécanisme utilisé par le locuteur qui a préféré la formulation *le recrutement s'effectue uniquement par voir de concours* plutôt

que *recrute par voie de concours.*

Si, comme nous l'avons vu plus haut, les répétitions disfluentes s'avèrent délicates à distinguer des répétitions « faits de langue », il en va de même pour les autocorrections dont le fonctionnement se révèle également proche de celui des énumérations (phénomène que nous avons évoqué précédemment). Cette proximité s'avère délicate lorsqu'on cherche par exemple à identifier automatiquement une autocorrection, car la confusion avec une énumération est très fréquente. En effet, l'énumération ou « ensemble d'items » comme le désigne [Gala Pavia, 2003] correspond à une succession d'éléments ayant la même fonction syntaxique et dépendant de la même tête (cf. *infra*). Il suffit pour s'en convaincre d'examiner quelques exemples d'énumérations pour se rendre compte de la similitude entre les deux phénomènes :

> *on travaille* **pour la police pour la gendarmerie** *euh on travaille* **pour beaucoup de monde**
>
> *lorsque* **les premières perceptions les premiers affects les premières émotions**
>
> *il y a* **un truc un machin une chose** *qui se détache*

[Gala Pavia, 2003] rajoute également que les structures de listes d'éléments à l'écrit peuvent avoir plusieurs relations au niveau sémantique telles que hyperonymie / hyponymie (de type « est un »), ou encore d'holonymie / méronymie (de type « fait partie de »). Ce type de relations se retrouve dans les corpus oraux augmentant ainsi les confusions possibles entre ces deux procédés.

> *euh* **des plantes des fleurs** *euh extrêmement euh euh rares en haute montagne*

Dans la plupart des cas problématiques c'est l'écoute de l'enregistrement audio qui permet de désambiguïser (lorsque cela est possible).

2.3.5 Inachèvements

Nous avons vu que plusieurs phénomènes (tels que répétitions et autocorrections) introduisent une rupture syntaxique mineure de l'énoncé. D'autres phénomènes sont plus complexes, et posent des problèmes syntaxiques plus importants car ils contiennent des ruptures syntaxiques profondes : les inachèvements. [Strassel, 2003] présente ce phénomène comme un cas où le locuteur abandonne un énoncé ou un constituant, qu'il ne corrige, ni ne répète partiellement ou complètement, mais au lieu de cela restructure l'énonciation et recommence. Elle ajoute également que le plus souvent, un inachèvement n'ajoute pas d'information à l'ensemble du discours.

Or, il convient de préciser que contrairement à ce que peuvent laisser penser les exemples d'énoncés inachevés, ceux-ci ne sont pas dénués d'apport d'informations. Candéa (2000) rappelle que les travaux de [Blanche-Benveniste *et al.*, 1979] sur l'organisation syntaxique des « bribes » montrent comment les faux départs construisent la séquence maximale d'un énoncé.

Les inachèvements peuvent également être perçus comme des échecs de la part du locuteur qui l'obligent à changer de construction [Blanche-Benveniste, 2000]. Par exemple, un locuteur peut abandonner sa tentative de définition pour procéder par une suite d'illustrations. Une autre possibilité est que le locuteur lance une construction qu'il semble abandonner mais qu'en fait il reprend un peu plus loin.

Nous pouvons nous appuyer sur la distinction établie par [Assie, 2005] pour illustrer plus spécifiquement trois possibilités d'interprétations pour un énoncé inachevé.

– Un effet d'indicible où le locuteur cherche ses mots

> *euh certains* **je** comment dire *certaines chaînes de restauration rapide*
> *il y a des après* **il y a** comment ça s'appelle *c'est qui qui est Beauty success*
> *qui vous a maquillée et et ah et voilà*

– Un échec amenant le locuteur à changer de construction

> *moi c'est vrai que je travaille dans* **le enfin** *je suis née de parents de commerçants*
>
> *mais moi* **je c'était** *juste l'année en plus où la majorité venait de passer à dix-huit ans*
>
> *vous savez que* **on a c'est** *comme ça petit à petit en travaillant qu'on s'est aperçu euh que euh cinquante pour cent*

– Une construction semble être abandonnée mais est finalement reprise plus loin

> *et* **je pensais** *enfin je c'est pas je pensais mais* **j'avais prévu** *de dormir dans ma 4L*
>
> il **se tâchait sa sa** *il ne ne buvait que des Blancs pardon euh parce qu'il ne voulait pas* **se tâcher sa sa barbe**

Les exemples ci-dessus ne sont pas sans rappeler ceux des amorces inachevées, à la différence de ces dernières il ne s'agit pas d'un mot mais le plus souvent d'un syntagme tout entier dont la production est amorcée. Toutefois, alors qu'il ne manque pas de travaux sur les différents types d'amorces (cf. [Pallaud, 2002] ; [Liu, 2003]), les inachèvements apparaissent comme le phénomène le moins étudié parmi les disfluences. Curieusement il n'a pas été considéré par de nombreuses études principales sur le traitement des disfluences ([Heeman, 1997] ; [Shriberg, 1994] ; [Core et Schubert, 1999] ; etc.)

Ce phénomène est néanmoins évoqué par [Lickley, 1994] où il est question de « faux départs » pour catégoriser un ensemble de phénomènes :

– Changement de mot
– Changement de qualification
– Changement de prononciation
– Changement syntaxique (qui correspond à ce que nous qualifions « inachèvements »)

Ce manque de descriptions linguistiques peut être attribué à l'importante complexité que représentent les inachèvements par rapport aux autres disfluences. Cette problématique se retrouve chez [Candéa, 2000b] qui parle de « faux départ **complexe** » pour désigner ce phénomène, et dont elle ne propose pas de classement du fait de la difficulté à délimiter une séquence d'origine et une séquence de correction puisque les deux se chevauchent ; comme nous l'avons vu, le locuteur ne corrige pas un seul trait en gardant la place syntaxique créée auparavant, mais abandonne la structure syntaxique en cours soit pour la reprendre plus tard, soit pour insérer de nouveaux constituants ou de nouvelles propositions.

2.4 Interaction entre disfluences : les disfluences combinées

Nous avons vu que les disfluences englobent un ensemble d'évènements langagiers pouvant apparaître indépendamment de tout autre. Cependant, dans le discours spontané, les disfluences n'apparaissent pas obligatoirement seules puisqu'il arrive également que les locuteurs produisent une série de disfluences simultanément au sein d'un même énoncé. Comme nous le verrons ci-après, une multitude d'exemples montre comment elles interviennent en combinaison les unes avec les autres. [Strassel, 2003] avance l'idée que cette combinaison peut s'opérer en série, où les disfluences apparaissent les unes après les autres, ou de façon imbriquée, c'est à dire lorsqu'une partie de la disfluence contient elle même une autre disfluence.

Les pauses (remplies et silencieuses) et l'allongement étant des phénomènes d'une grande fréquence à l'oral, ils se retrouvent dans toutes les configurations possibles d'association (répétition, amorce, autocorrection, inachèvement). Sans négliger l'étude de l'interaction entre les pauses[5] et les autres phénomènes, nous

[5]Emploi générique pour les pauses silencieuses, pauses remplies et allongements.

avons choisi de décrire d'autres types de combinaisons.

Nous appellerons « disfluences combinées » tout ensemble de disfluences mettant en oeuvre au minimum deux phénomènes apparaissant simultanément dans un énoncé. Les phénomènes associés occupent par ailleurs le même emplacement syntaxique.

Cette notion peut s'illustrer à l'aide des exemples suivants :

a) *j'ai mon lundi on a notre* **rep-** nos **repos** *euh qui sont obligatoires*

b) ces les **les les les** *médias euh* **natio-** *euh* **nationaux** *hein*

La combinaison peut s'opérer de façon relativement simple comme dans a) où il y a mélange entre amorce (*rep- / repos*) et autocorrection (*notre / nos*). Les cas peuvent être plus complexes comme dans b) où on retrouve à la fois autocorrection (*ces / les*), répétition (*les / les / les / les*) et amorce (*natio- / nationaux*).

Nous tentons de dresser dans les lignes suivantes une liste des combinaisons possibles - illustrées d'exemples - qui ne prétend évidemment pas être exhaustive[6] étant donné la diversité de ce type de phénomène. A cet égard, [Piu, 2006] a observé sur deux corpus oraux (de nature différente) que les disfluences combinées représentent de 10 à 17% des occurrences observées.

2.4.1 Disfluences combinées à partir de répétition

Si l'on revient sur le phénomène de répétition par exemple, les différents types de répétitions (cf. *infra*) peuvent de plus être associés à d'autres éléments (autocorrections, amorces, etc.) rajoutant ainsi de la complexité dans la typologie des nombreux exemples déjà recensés.

[6]Les travaux de [Candéa, 2000b] montrent également différents types de combinaisons de disfluences.

– Répétition et autocorrection

La combinaison qui, toutes proportions gardées, semble apparaître le plus fréquemment au cours de nos observations touche la répétition associée à l'autocorrection (dans cet ordre[7]) :

> *vous dites que la rencontre* **de de ces** de cette *ethnie v- vous a complètement b- euh changé*
>
> *il y a une remise en question au niveau euh physique euh* **qui qui** *bon* que *j'ai remarquée en tout cas*

– Répétition et amorce

La combinaison particulière entre répétition et amorce permet d'observer plus clairement que d'autres type de combinaison l'élaboration progressive du discours par le locuteur. En effet, l'amorce d'un mot n'est pas toujours complétée ou modifiée immédiatement mais peut nécessiter le besoin de répéter en ajoutant de l'information sur la production du locuteur à chaque élément répété.
Par exemple :

> **de la de la** fl- **de la** flore
>
> **les les** en- **les les** enfants *d'aujourd'hui*

On voit clairement dans cet exemple trois phases de recherche de formulation du locuteur : le premier élément répété est allongé, signalant ainsi qu'il ne trouve pas le mot souhaité, puis l'amorce indique qu'il a trouvé le début du mot qu'il complète ensuite en confirmant le mot qu'il vient d'amorcer.

– Répétition et inachèvement

La répétition d'une unité ou d'un syntagme entier traduit manifestement la recherche de mots ou de structures syntaxiques adéquats. Or, il arrive parfois

[7]Nous verrons plus loin que la combinaison « autocorrection + répétition » constitue un autre type de combinaison où les exemples ne sont pas du même type que ce dont il est question ici.

que cette recherche soit abandonnée par le locuteur, donnant ainsi lieu à une répétition d'éléments générant ainsi un énoncé laissé inachevé.

> **c'est c'est c'est c'est** quand même une euh *ce poulet blanc euh qui est élevé*
>
> *la sauce avec euh* **un un** degré quand même euh **bon euh**

– Répétition d'autocorrection

Dans le même ordre d'idée que les exemples précédents, il arrive que les disfluences mise en combinaison soient mêlées, un phénomène prenant place dans la configuration d'un autre :

> *c'était* **un une un** une envie un besoin *qui est commun*

Dans l'exemple ci-dessus l'autocorrection un une fait l'objet d'une répétition qui précède ensuite une nouvelle autocorrection.

Il existe également d'autres configurations qui affectent le phénomène de répétition sans pour autant mettre en jeu d'autre disfluences, mais plutôt la répétition elle-même comme en témoigne la catégorie suivante.

– Répétitions avec précision

La répétition peut être enrichie d'autres éléments (adjectifs qualificatifs par exemple) qui viennent s'insérer au fur à mesure de l'énonciation du locuteur, qui se rend compte que son énoncé ne comporte pas suffisamment de détails (la différence avec le phénomène d'autocorrection est dans ce cas relativement faible) :

> *il y a euh des fiches sur* **la la** *faune* **la** grande *faune /*
>
> *le grand cru que* **qui est connu qui est** mondialement **connu**

La combinaison de phénomènes ne s'applique évidemment pas uniquement aux répétitions puisqu'il en est de même pour l'ensemble des disfluences présentées

jusqu'à présent.

2.4.2 Disfluences combinées à partir d'autocorrection

Il convient de rappeler que le fonctionnement de l'autocorrection ressemble quelque peu à celui de la répétition en ce sens que la distinction porte sur l'opération effectuée sur l'élément repris (répété dans un cas et modifié dans l'autre). Il n'est donc pas étonnant d'observer des combinaisons similaires aux exemples précédents.

– Autocorrection et amorce

notre rôle *est enfin les* n- nos **interventions**

– Autocorrection et inachèvement

qui *euh* **dans lequel** nous *il y a les agents qui sont concernés*

– Autocorrection et répétition

les *euh* **la** la *scolarité n'était pas mixte à l'époque*

qu'un qu'une *euh* qu'un *vin issu*

– Autocorrection, répétition et amorce

une *euh* **la** la la *euh* le cé- le cépage *de euh la ville de Beaune*

les gens *disent* **les parents** les l'entour-

2.4.3 Disfluences combinées à partir d'amorce

Avant d'être complétée ou modifiée, l'amorce peut être suivie d'une autre disfluence qui vient le plus souvent éloigner la complétion ou la modification du segment amorcé.

– Amorce et autocorrection

notre rep- **nos** repos *euh qui sont obligatoires*

– Amorce et répétition

c- c- c'est *de la bonne heure*

– Amorce et inachèvement

l- la il faut fabriquer le programme

– Amorce, répétition et autocorrection

il p- il p- il y a *des moments de pluie assez importants des fois*

– Amorce et incise

ça **ça nous a ét-** même moi qui le enfin je le dis parce que je le voyais *mais ça m'a* **ça nous avait aussi étonnées** *parce que c'était imprévu*

Les différentes catégories d'exemples de disfluences combinées reflètent sans doute encore mieux la mise en place du discours par les locuteurs du fait des multiples phénomènes mis en oeuvre.

2.4.4 Les exemples « inclassables »

Le point précédent nous a montré que les disfluences peuvent apparaître en combinaison les unes avec les autres. Les associations entre phénomènes peuvent parfois atteindre un tel degré de complexité que l'on assiste alors à des productions « chaotiques » qu'il peut être difficile de catégoriser.

En effet, certaines constructions rendent compte de combinaisons de disfluences encore plus déroutantes – au premier abord – que celles des phénomènes présentés ci-dessus. Nous laissons donc ces cas particuliers dans une catégorie provisoire dite d'exemples « inclassables » en attendant d'obtenir des données permettant de nous éclairer sur leur analyse, ou de trouver une explication permettant d'en rendre compte. Ce sera précisément l'objet de notre approche théorique plus loin

dans ce mémoire.

Nous trouvons par exemple des cas tels que :

> *de- que la : + euh d- euh cin- bon plus de cinquante pour cent*
>
> *disons que ça a quand même beaucoup beaucoup enfin pas ma- em- comment dire ça a énormément évolué quoi*
>
> *je dirai que c'est quand même des s- ce qu- ce qui est demandé ici c'est quand même euh + euh + c'est plus ouais ça s- c'est euh un comment dire*

2.5 Conclusion

Nous avons vu que le français parlé comporte de nombreuses spécificités qui le distingue clairement du français écrit. Nous avons plus précisément vu à l'aide d'une typologie détaillée que les disfluences sont nombreuses et touchent des niveaux variés de l'énoncé.

Nous verrons dans la suite de notre étude que des énoncés comportant ce type de phénomènes vont s'avérer problématiques pour effectuer une analyse syntaxique automatique robuste. En effet, la prise en compte de telles particularités reste encore un défi majeur pour la majorité des étapes d'analyses en TAL. La question sera alors de savoir quelle est l'importance de ces cas en terme de fréquence dans les corpus oraux, et de savoir si cette fréquence dépend d'un contexte syntaxique spécifique.

Chapitre 3

Mise en relief de l'observable : les corpus oraux

3.1 Introduction

Ces dernières années ont vu les linguistiques « de corpus » s'affirmer à part entière dans la recherche fondamentale en linguistique qui semble vouloir renoncer à l'« héritage » laissé par Noam Chomsky et renouer avec une tradition anglo-saxonne fondée sur l'empirisme. [Halliday, 1985] précisait déjà que l'invention du magnétophone portatif devrait être considérée comme un point de départ déterminant pour le développement de la linguistique. Cet outil offre la possibilité pour la première fois d'étudier des échantillons de sa propre parole, et de les conserver de façon aussi stables que pour les échantillons de données écrites. Or, le magnétophone n'a pas été utilisé d'emblée dans cette optique ; l'invention datant de 1930, elle servit dans un premier temps à étudier les langues dépourvues d'écritures (patois, dialectes, etc.). Il n'en était pas encore question pour les langues de culture plus « importante » (comme le français par exemple). L'un des premiers obstacles à l'étude de l'oral venait donc du manque de corpus oraux[1] à étudier.

Généralement, la tâche que se donnent les linguistes de corpus consiste à se détacher

[1] On entend généralement par « corpus oraux » les annotations (sous forme de transcriptions orthographiques) ainsi que les enregistrements (fichiers sons et/ou vidéo).

de la notion d'intuition du locuteur natif, de manière à privilégier l'observation de données réelles extraites de corpus informatisés (écrits ou oraux). On assiste ainsi (depuis une vingtaine d'années notamment) au développement de grands corpus de langue écrite mais surtout de langue parlée. Cette évolution a considérablement modifié l'approche du langage et des sciences qui s'y intéressent. Cela a permis par exemple de voir émerger une nouvelle linguistique moderne.

On observe alors de plus en plus d'approches linguistiques qui basent leurs observations sur de vastes corpus oraux numérisés, et des outils d'analyse automatique toujours plus performants se multiplient. Les nouvelles technologies en matière de stockage, de diffusion mais aussi d'exploitation des enregistrements sonores, couplées aux outils de traitement automatique du langage, sont à l'origine de nouvelles perspectives de traitement automatique. Néanmoins cette situation ne va pas sans poser de nombreuses questions juridiques et éthiques mais aussi techniques, méthodologiques et théoriques. Pour mettre en avant notre objet d'étude il nous a semblé nécessaire de faire un point sur les données disponibles en traitement automatique de l'oral. Nous opposerons dans un premier temps les corpus oraux aux corpus écrits afin de montrer que ces deux types de données ne fournissent pas les mêmes informations. Une deuxième partie sera ensuite consacrée aux principaux corpus d'oral spontané disponibles en français mais aussi à l'étranger.

3.2 Corpus écrits VS corpus oraux

Alors que l'on trouve à l'heure actuelle des quantités importantes de textes écrits (cf. la base *Frantext* et ses 210 millions d'occurrences), on dispose à l'inverse de très peu de corpus oraux transcrits. Pourtant, [Cappeau et Seijido, 2005] rappellent que les travaux d'autres pays (Grande-Bretagne, Portugal et Italie, par exemple) ont montré la nécessité de disposer de corpus oraux d'importance : constitution de nouveaux outils de description de la langue en vue de l'enseignement (en tant que

langue maternelle et en tant que langue étrangère) ; constitution de banques de données pour servir de comparaison dans le domaine de l'acquisition du langage, des troubles du langage, de l'évolution de la prononciation et du lexique, etc.

De plus, malgré les évolutions technologiques de ces dernières années, l'analyse automatisée de l'oral reste encore marginale et ce, entre autre, du fait du manque de corpus oraux. Plusieurs raisons peuvent expliquer cet état de fait. D'une part, la langue écrite a été longtemps l'objet essentiel des études linguistiques reléguant l'oral au rang de langue fautive, inorganisée, une sorte de « brouillon » de l'écrit. Ce n'est que depuis quelques années que l'étude de la langue parlée a suscité un intérêt grandissant auprès des linguistes qui ont écarté ces préjugés normatifs et puristes pour étudier l'oral. D'autre part, s'il est relativement aisé d'accumuler de larges quantités d'écrit sous forme électronique (et l'augmentation du nombre de pages Web sur Internet repousse chaque jour ces possibilités), la constitution de corpus oraux est davantage problématique car elle représente un investissement en temps bien plus prohibitif et se révèle extrêmement fastidieuse.

Ce n'est qu'avec l'apparition des magnétophones portables et plus récemment des lecteurs enregistreurs *mini-disc* qu'il a été possible d'effectuer des enregistrements quasi-illimités et de grande qualité. Mais avec le perfectionnement des moyens techniques pour saisir la parole, les différences entre oral et écrit sont apparues de façon plus marquées ; les problèmes posés par la transcription de grands volumes de données orales ont obligé à faire un certain nombre de choix théoriques et à définir des conventions de transcriptions variables selon les équipes de recherches et les objectifs des études ([Blanche-Benveniste, 2000]).

Une des premières difficultés au moment de transcrire des corpus oraux consiste à se détacher de ses habitudes de lecture largement influencées par la pratique écrite. En effet, celle-ci accrédite l'idée selon laquelle nous parlons avec des phrases, des mots bien distincts et complets, des signes de ponctuation, etc. alors qu'il s'agit

foncièrement de notions purement graphiques. En ce qui concerne la ponctuation par exemple, il s'agit d'un procédé propre à l'écrit et lorsque l'on étudie des productions orales il est difficile d'établir une correspondance entre les signes conventionnels de l'écrit et les indices de l'oralité qu'ils sont censés représenter. [Blanche-Benveniste et Jeanjean, 1987] expliquent les désavantages liés à l'introduction de la ponctuation dans les transcriptions :

> " *La ponctuation, si on la met trop tôt, préjuge de l'analyse syntaxique et impose un découpage sur lequel il est difficile de revenir.[. . .] Les conventions graphiques mettent trop d'ordre dans un domaine où la langue parlée a des mécanismes complexes et mal connus.* "

Cette idée apparaît de nouveau plus récemment lorsque [Blanche-Benveniste, 2000] revient sur le fait que les différentes marques de ponctuation comme le point d'exclamation, la virgule, la majuscule ou les guillemets fournissent des équivalents approximatifs de plusieurs types de productions orales. De plus, ces équivalents sont trop peu nombreux pour pouvoir refléter la grande diversité des effets de l'oralité, tel que l'accent d'insistance par exemple, l'allongement, la montée de la voix ou encore le changement de débit, et tout ce que l'écriture est incapable de représenter, comme le ton ironique ou les différentes forces illocutoires.

[Véronis, 2004] signale en outre que le manque de corpus oraux entraîne un double paradoxe ; premièrement, la linguistique se limite essentiellement à l'observation des données écrites alors qu'elle affirme fermement (depuis les conceptions Saussuriennes notamment) la primauté de l'oral, dont l'écrit ne serait que le système de codage à l'aide de signes visibles. Le second paradoxe vient du fait que la suprématie des corpus écrits est indéniable alors que ce sont les technologies de la parole qui ont participé fortement au regain d'intérêt pour les corpus à partir des progrès enregistrés au moyen de méthodes empiriques.

Il en résulte que les études sur corpus sont massivement basées sur l'écrit induisant

un biais important en linguistique descriptive : [Véronis, 2004] rappelle qu'il est hasardeux d'extrapoler sur la langue en général à partir des observations réalisées sur l'écrit au risque de mette au point *in fine* la grammaire du journal *le Monde* ou du *New York Times*.

Cette remarque s'applique également aux technologies de la parole ; il semble effectivement inapproprié d'entraîner des modèles de langage sur des données écrites, même en quantité volumineuse. En effet, celles-ci sont considérablement éloignées des productions que l'on trouve à l'oral et dégrade inévitablement les performances des systèmes concernés. [Baude *et al.*, 2005] signalent par ailleurs que les grands corpus oraux actuels sont « rentables » à plus d'un titre. Plusieurs organismes se sont spécialisés dans la diffusion des corpus disponibles et ce pour des domaines aussi variés que le dialogue homme-machine, la synthèse ou la reconnaissance de la parole ou encore les communications téléphoniques.

3.3 Principaux corpus oraux transcrits existants

Ces dernières années, l'étude de l'oral était réservé essentiellement aux domaines où il s'exerçait habituellement : phonétique, phonologie, prosodie, ou encore études des langues sans tradition écrite (parlers régionaux et autres langues plus rares). Il existe aujourd'hui de nombreux types de corpus oraux dans différentes disciplines conçus pour répondre à des besoins bien définis (renseignements touristiques, etc.). Il s'agit le plus souvent d'enregistrements de données sonores (éventuellement enrichies de données vidéos), presque toujours accompagnées de transcriptions et de traitement informatisés.

Lorsqu'il s'agit de se consacrer au langage lui même, le choix d'un type de transcription dépend des finalités de l'étude. [Baude *et al.*, 2005] rappellent à ce propos qu'une transcription engage toujours une théorie. Dans le cadre d'études acous-

tiques par exemple (prononciation, acquisition du langage, etc.), le choix se portera de préférence sur des transcriptions phonétiques ou phonologiques. Dès lors que l'on s'intéresse à la description du français parlé, il convient d'utiliser des transcriptions d'enregistrements en orthographe standard, pour en rendre – entre autres – la lecture plus accessible. La nature des corpus diffère donc selon les aspects du langage que l'on souhaite étudier ; nous traiterons ici uniquement de corpus oraux qui ont été transcrits orthographiquement.

Nous ne présentons pas ici une liste exhaustive des différents corpus oraux transcrits disponibles[2] ; nous exposerons brièvement quelques uns des grands corpus non-francophones ainsi que le descriptif des ressources les plus conséquentes pour le français.

3.3.1 Corpus non francophones

Bien avant les études menées sur le traitement de l'oral pour le français, des transcriptions de corpus étaient déjà disponibles pour d'autres langues.

C'est par exemple le cas pour les corpus anglophones. En Angleterre, le British National Corpus[3] (BNC) représente une collection de 100 millions de mots constituée à partir d'échantillons d'écrit et de langue parlée provenant de multiples sources. Il a été conçu afin de représenter le plus largement possible l'anglais britannique contemporain tant parlé qu'écrit. La section parlée du British National Corpus représente sans conteste le plus volumineux des corpus de données orales à l'heure actuelle. Celle-ci comprend 10 millions d'occurrences (alors que la partie écrite de ce même corpus représente 90 millions de mots). Aux Etats-Unis, le Santa Barbara Corpus of Spoken American English[4] a été constitué par les chercheurs du Départe-

[2]Un inventaire des corpus de français parlé plus détaillé est disponible dans l'ouvrage de [Blanche-Benveniste et Jeanjean, 1987], et plus récemment dans le document de [Baude *et al.*, 2005].

[3]http ://www.natcorp.ox.ac.uk/corpus/

[4]http ://www.linguistics.ucsb.edu/research/sbcorpus.html

ment de Linguistique de L'Université de Californie. Il est basé sur un recueil de 60 enregistrements de parole spontanée dans les interactions naturelles effectués partout aux Etats-Unis. Ce corpus représente une grande variété de populations aux origines régionales, âge, environnements sociaux et ethniques différents. Ce corpus est de plus une composante de l'International Corpus of English (ICE) dont il est la principale ressource des données parlées spontanées.

En plus de ces ressources, d'autres corpus sont en cours de constitution ; par exemple en Hollande (Corpus Gesproken Nederlands), en Israël (Corpus of Spoken Israeli Hebrew[5]) ou encore au Portugal (Corpus de Portugais Parlé).

3.3.2 Corpus francophones

ESLO : Corpus d'Orléans

Le corpus ESLO[6] (Enquête Socio-Linguistique d'Orléans) constitue, par son ampleur (350 heures d'enregistrements représentant 4 500 000 mots transcrits) et sa cohérence, le plus important témoignage sur le français parlé avant 1980. Constitué par des universitaires britanniques à des fins didactiques (enseignement du français langue étrangère dans le système public d'éducation anglais), il représente à la fois une précieuse masse de documents et une collection de bandes magnétiques vieillissante.

Partant de l'expérience acquise, le CORAL (Centre Orléanais de Recherche en Anthropologie et Linguistique) a mis en chantier une nouvelle enquête dénommée ESLO2. L'objectif est d'évaluer, à une quarantaine d'années de distance, la dynamique sociale du français (et des usages de la langue comme des jugements sur son emploi) en prenant en compte la diversité des changements en fonction des paramètres sociaux. Engagé depuis 2005, ce programme a pris la forme d'une préservation d'ESLO1 (en transférant les données sonores contenues dans l'enregis-

[5]http ://www.tau.ac.il/humanities/semitic/cosih.html
[6]http ://crdo.vjf.cnrs.fr :8080/exist/crdo/projets.htm

trement magnétique et en assurant son indexation) et d'un traitement numérique des données, avec un recensement des travaux effectués sur le corpus.

CLAPI

La banque de données CLAPI[7] (Corpus de Langue Parlée en Interaction) est constituée actuellement à Lyon (laboratoire ICAR) en vue de réunir des corpus de « parole en interaction » les plus diversifiés possibles, dans des situations non provoquées par les chercheurs : conversation à table, concertations entre notaires, appels à des centre d'aide sociale d'urgence, etc. Cette banque de données comporte à l'heure actuelle près de 600 heures d'enregistrements audio et vidéo, des transcriptions et des « métadonnées » décrivant les caractéristiques des locuteurs.

Le projet Phonologie du Français Contemporain

Le projet international PFC[8] (Phonologie du Français Contemporain) propose une banque de données regroupant un ensemble de corpora de français oral contemporain dans l'espace francophone. Ce projet s'adresse à la fois aux chercheurs, enseignants/apprenants du français, ou encore au grand public.

Il offre une base de données qui a pour objectif, à terme, de constituer la plus grosse base de données orales portant sur le français, puis toutes langues confondues. Cette base peut être utilisée dans le cadre de la recherche (phonétique, phonologie, syntaxe, pragmatique, sociolinguistique, analyse conversationnelle, etc.), de lŠenseignement / apprentissage du français (langue étrangère, maternelle ou seconde) et de la diffusion des savoirs (conservation du patrimoine linguistique francophone et présentation générale du français oral contemporain pour les non-spécialistes).

Le projet couvre 33 zones géographiques (Afrique, Antilles, Belgique, Canada, France, etc.) au coeur desquelles sont parallèlement menées 75 enquêtes auprès de

[7] http ://corpus.univ-lyon2.fr/
[8] Description détaillée : http ://www.projet-pfc.net/

450 locuteurs. Dans l'état actuel de la base, 27 enquêtes sont consultables en ligne, comprenant 297 locuteurs.

C-ORAL-ROM

La ressource C-ORAL-ROM est un corpus multilingue (dont une partie française) de parole spontanée pour les principales langues romanes, composé d'environ 1 200 000 mots.

Le corpus est composé de quatre collections d'enregistrements comparables de sessions de parole spontanée pour l'italien, le français, le portugais et l'espagnol (environ 300 000 mots par langue). Les collections ont été fournies par les organismes suivants :

- Università di Firenze (Dipartimento di Italianistica, LABLITA) ;
- Université de Provence, Aix-en-Provence (Lanoratoire Description Linguistique Informatisée sur Corpus) ;
- Fundação da Universidade de Lisboa/Centro de Linguística da Universidade de Lisboa ;
- Universidad Autónoma de Madrid (Departamento de Linguística, Lenguas Modernas, Lógica y F. de la Ciencia, Laboratorio de Lingüística Informática).

La ressource a pour but de représenter la variété des actes de parole de la langue de tous les jours et de faciliter l'induction de structures prosodiques et syntaxiques dans les quatre langues romanes traitées, d'un point de vue quantitatif et qualitatif. La ressource a été conçue pour la modélisation prosodique, les procédures de test en TAL et des études de la parole spontanée basées sur les corpus.

Le design de la ressource C-ORAL-ROM vise à assurer un maximum de possibilités d'occurrences pour une grande variété de types d'actes de parole et de contours prosodiques naturels, qui sont les traits linguistiques les plus particuliers que l'on puisse trouver en parole spontanée. A ces fins, les paramètres de variation princi-

paux du domaine de la parole (variation de canaux, structure du dialogue, domaine sociologique d'usage, et domaine sémantique d'application) sont représentés dans un schéma de design de corpus, couvrant une grande variété de domaines sémantiques et pragmatiques d'application.

VALIBEL (Belgique)

Le corpus Valibel, placé sous la responsabilité de Michel Francard, a été constitué dans une optique d'exploitation pluridisciplinaire (phonologie, syntaxe, sociolinguistique, etc.). Il comporte plus de 500 entrevues dans divers formats (son, vidéo, textes) représentant 373 heures d'enregistrements. Ceux-ci ont été intégralement transcrits à l'aide des conventions des transcriptions du centre de recherche Valibel, et représente un volume de données avoisinant les 4 millions d'occurrences. (Une description plus détaillée est présentée dans l'article de Francard (1990)).

OTTAWA-HULL (Canada)

Pour finir de passer en revue les principales ressources orales disponibles, il existe au Canada une banque de données d'environ 3,5 millions de mots tirés des conversations informelles d'un échantillon représentatif de 120 francophones natifs d'Ottawa-Hull, classés selon l'âge, le sexe et le statut minoritaire/majoritaire du français dans leur quartier de la région de la capitale nationale. À travers 283 enregistrements audio (qui ont été transcrits), le corpus comprend un éventail de traits vernaculaires et du discours bilingue naturel.

3.4 Corpus de travail : Le Corpus de Référence du Français Parlé

Les études sur l'analyse de la syntaxe du discours oral ont pour la majeure partie été menées sur de vastes corpus anglais (cf. 3.3.2). En effet, lorsqu'on s'intéresse aux productions spontanées (et, de fait, aux disfluences), les ressources en français

ne sont pas toujours faciles à obtenir.

Dans ce travail nous avons pourtant eu l'opportunité d'utiliser le Corpus de Référence du Français Parlé (CRFP). Celui-ci constitue une référence unique en la matière et nous a permis de travailler sur des données spontanées, non préparées, à l'inverse des corpus les plus répandus souvent planifiés à l'avance par les locuteurs, et restreints à un domaine précis.

Le Corpus de Référence du Français Parlé[9] (CRFP) répond à une requête de la Délégation à la langue française (Ministère de la Culture) qui a financé ce projet. La réalisation de ce projet avait été confiée, en 1998, à l'équipe du GARS, repris à partir de 2000 par l'équipe DELIC.

L'objectif de ce corpus est de mettre à la disposition de la communauté des linguistes, chercheurs et enseignants, 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 couranté », ce qui a amené à effectuer certains choix touchant aussi bien aux caractéristiques des locuteurs qu'aux situations de parole.

Trois types d'enregistrements ont été réalisés renvoyant à différentes situations de parole :

- *La parole privée* : entretien sollicité dans le cadre de l'enquête. Cette situation de parole renvoie à deux types de production : le récit de vie (récit d'un voyage, d'une expérience, souvenirs d'enfance, etc.) ou la présentation d'un « savoir-faire » professionnel ou autre.
- *La parole professionnelle* : entretiens dans lesquels les locuteurs ont été enregistrés dans l'exercice de leur fonction ou quand ils évoquent leur profession sur leur lieu de travail.

[9]Pour une présentation plus détaillée du CRFP voir : [DELIC, 2004]

– *La parole publique* : cette situation se distingue des deux autres par le fait que les intervenants s'expriment toujours en présence d'un public ; elle comporte une partie d'entretiens sollicités, le reste étant constitué d'émissions radiophoniques, de cours et conférences, de réunions politique ou associative et de quelques situations plus spécifiques (visite de musée, dégustation de vins, etc.).

Les locuteurs sont catégorisés selon trois tranches d'âge (18-30 ans, 30-65 ans et plus de 65 ans) et selon leur niveau scolaire (collège, bac + nombre d'années à l'université). Ils devaient être natifs de la région où était réalisé l'enregistrement. Le corpus total se compose d'environ 440 000 mots ce qui représente 134 enregistrements dont la partie transcrite correspond à une durée moyenne de 16 minutes et 48 secondes (soit 36 heures et 50 minutes de parole).

Dans le projet initial, il était prévu que le corpus comporte 160 enregistrements mais les problèmes de coordination et d'harmonisation, dus à la multiplicité des régions géographiques, des enquêteurs et des transcripteurs, avaient été sous-estimés. Ce projet représente une occasion unique de fournir à la communauté de recherche francophone un outil scientifique important, qui lui fait défaut, alors que des corpus oraux importants existent ou sont en cours de réalisation pour la plupart des langues européennes.

Exemple extrait du CRFP :

> oui nous avons des cibles qui sont fixées euh + par les disons les par un plan de vol + et c'est pas toujours évident il faut qu'on choisisse un point de départ pour pouvoir rejoindre ces cibles + et c'est pas évident parce que + en f- c'est là qu'il faut être écout- être très attentif à la au c- à la météo + et et pu- bien connaître un petit peu la région + c'est surtout ce qui compte c'est connaître bien la région

Nous avons plus particulièrement travaillé sur une sous-partie de ce corpus spécia-

lement choisie pour son caractère monologique et son hétérogénéité situationnelle. Cette sous-partie se compose de dix enregistrements faisant intervenir cinq hommes et cinq femmes. La durée totale du corpus est de 52 minutes et 78 secondes (en moyenne 5 minutes 27 secondes par extrait). La transcription orthographique a été effectuée entièrement à la main par des experts avec un cycle de réécoute/validation extrêmement strict. Les conventions de transcription adoptées ne contiennent aucun trucage orthographique (du type *p'tit*, *y'a*, etc.) ni aucune ponctuation, suivant la tradition de l'équipe. Par ailleurs, un certain nombre de phénomènes de production à l'oral ont été notés[10] : les répétitions, les amorces, les euh d'hésitation, les allongements, les pauses, les accents ainsi que les mouvements intonatifs majeurs.

Exemple extrait du sous-corpus :

> il y a : il y a une remise en question qui qui bon : + que j'ai remarquée en tout cas \+ et donc / là-bas c'était quoi c'était euh la jaunisse enfin / toujours est-il que / + il pleuvait pareil / + euh parce que en hiver au Portugal il p- il p- il y a des moments de pluie assez importants des fois / + et j'étais au fin fond de ma 4L euh complètement épuisée / ++ même je délirais / un petit peu / + enfin bon voilà euh le début de mon voyage / ça a été ça \++

Nous travaillons également sur une version enrichie de ce corpus où les phénomènes de disfluences ont été syntaxiquement annotés (voir [Piu, 2006]). Cette version nous permettra de dégager des pistes pour l'analyse syntaxique ultérieure de notre travail, dans la mesure où les patrons syntaxiques des segments disfluents ont été méticuleusement examinés.

Exemple extrait du corpus annoté :

> <dis type="rep">
> |il y a :

[10] Les phénomènes ont été notés dans les transcriptions mais n'ont pas fait l'objet d'un balisage.

|il y a une re*mise en question|
</dis> au niveau euh + physique / euh
<dis type="dc">
|qui
|qui <md>bon :</md> +
|que j'ai remarquée <md>en tout cas</md> \|
</dis>
+ et <md>donc</md> / là-bas c'était quoi c'était euh la jaunisse <md>enfin</md>
/ toujours est-il que / + il pleuvait pareil / + euh parce que en hiver au Portugal
<dis type="dc">
|il p-
|il p-
|il y a des moments de pluie assez importants des fois /|
</dis>
+ et j'étais au fin* fond de ma 4L euh complètement épuisée / ++ même je délirais / un petit peu / + <md>enfin bon voilà</md> euh le dé*but de mon voyage / ça a été ça \++

Exemples de patrons syntaxiques dégagés pour la séquence d'origine de l'autocorrection (*i.e* séquence qui va être modifiée) :

Ces informations permettent d'examiner ensuite la proportion globale de disfluences dans le corpus, ainsi que la répartition des différents types de disfluences. De plus, ceci constitue un atout précieux dans l'optique d'implémentation de règles syntaxiques. En effet, les patrons syntaxiques permettent de dégager les catégories qu'il est préférable de traiter en priorité pour chaque disfluence. Par exemple, pour la réalisation de règles de grammaires automatiques appliquées au cas des autocorrections, il sera nécessaire de se consacrer dans un premier temps au traitement des séquences [pronom personnel + verbe conjugué] et [préposition + article défini].

Séquences d'origine formées de plusieurs unités	
Patrons	**Effectif**
Pronom personnel + verbe "avoir" conjugué + (article indéfini)	5
Préposition + article défini	3
Pronom personnel + verbe conjugué	3
Article indéfini + (adjectif) + nom	2
Forme renforcée + pronom personnel + verbe conjugué	2
Pronom neutre "ce" + verbe "être" conjugué	2
Verbe conjugué + clitique objet indirect	1
Forme renforcée + pronom personnel	1
Pronom personnel + verbe "avoir" conjugué + clitique réfléchi + verbe infinitif + adjectif possessif	1
Pronom personnel + verbe "avoir" conjugué + forclusif	1
Pronom personnel + clitique réfléchi + verbe conjugué + article indéfini + nom	1
Pronom personnel + "ne" négatif + "avoir" conjugué + forclusif	1
Pronom relatif + verbe conjugué	1
Total	**24**

FIG. 3.1 – Exemple de patrons syntaxiques (extrait de [Piu, 2006]).

3.5 Conclusion

Les corpus de langue orale spontanée sont d'une importance fondamentale pour l'étude linguistique, comme pour la mise au point de nouvelles technologies vocales. Nous l'avons vu, de grands corpus d'enregistrements et de transcriptions ont été constitués et systématiquement exploités par des moyens informatiques. On ne se contente plus de l'intuition pour rendre compte de sa propre langue parlée.

Ce grand intérêt pour les corpus de langue parlée s'explique par plusieurs raisons. Les moyens techniques permettant de stocker et de rendre accessibles de grandes quantités de données, de langue parlée comme de langue écrite, ont changé les types d'analyse. Une certaine évolution des opinions a sans doute fait reculer les préjugés classiques accumulés contre la langue parlée réputée fautive, populaire et vulgaire pour devenir un véritable objet d'étude scientifique.

À ce titre, le CRFP constitue un corpus idéal pour l'étude des disfluences qui

apparaissent dans un contexte non préparé. Ce contexte représente en effet une production fidèle à celui des conversations que l'on peut entendre dans différentes situations de la vie quotidienne, ainsi plus proche de la réalité.

Tout en gardant à l'esprit les caractéristiques fondamentales du français parlé, la typologie des disfluences et des corpus oraux disponibles que nous avons dégagées dans ces chapitres, nous nous proposons à présent d'étudier les représentations formelles théoriques des phénomènes oraux développées au cours de ces dernières années.

Chapitre 4

Modélisation des phénomènes de production

4.1 Introduction

Nous avons vu dans le second chapitre que les disfluences sont sources de nombreuses difficultés à prendre en compte dans une perspective plus large de traitement de l'oral. La tâche s'annonce effectivement délicate car, au premier abord, les disfluences ne montrent pas de cohérence syntaxique particulière. Les énoncés disfluents semblent présenter une « rupture » de la représentation syntaxique qu'on ne trouve pas à l'écrit. L'aspect « désordonné » de certaines tournures syntaxiques à l'oral semble aller dans ce sens :

> *j'ai pris enfin j'ai j'ai je me s- je je leur ai dit c- enfin je leur ai dit ce que j'en pensais*

Pourtant, de nombreuses réflexions théoriques (que nous exposons ci-après) tendent plutôt à défendre l'idée selon laquelle les disfluences s'organisent de diverses manières et notamment syntaxiquement, au même titre que des énoncés ne présentant aucune déviance syntaxique.

De plus, de nombreux travaux ont permis de mettre en évidence l'organisation de certains phénomènes tels que les pauses remplies ([Campione et Véronis, 2004]),

les répétitions ([Candéa, 2000b] ; [Henry, 2002b]), etc.

Nous l'avons évoqué précédemment (cf. 2), les disfluences sont souvent considérées comme des éléments agrammaticaux. Nous pensons au contraire que ces éléments peuvent tout à fait s'insérer dans un cadre d'analyse grammatical de la même manière que l'écrit. Les disfluences ne sont pas des unités à la fréquence aléatoire, à tel point que de nombreuses études descriptives et applicatives dans des disciplines diverses (psycholinguistique, linguistique ou traitement automatique des langues) se sont développées afin de formaliser leur structure sous-jacente et notamment les contraintes syntaxiques auxquelles elles sont éventuellement soumises.

Nous donnerons plusieurs exemples de représentations qui ont été proposées dans cet ordre d'idée. Nous présenterons celle que nous avons choisi d'adopter pour notre étude dans la partie suivante.

4.2 Modèles psycholinguistiques

L'analyse de l'oral a été abordée dans un premier temps par les psycholinguistes qui ont envisagé les disfluences comme un moyen privilégié pour délimiter les étapes de la production langagière. La majeure partie de ces études porte sur la production des disfluences par les êtres humains selon plusieurs points de vue tels que les différences individuelles de production ([Maclay et Osgood, 1959]), l'interaction entre la structure des disfluences et leurs processus de génération ([Levelt, 1983]), etc. D'autres études, moins nombreuses, ont porté sur la perception des disfluences ([Lickley, 1994]).

4.2.1 Modèles de production du discours

En psycholinguistique, les « modèles de production du discours » ou « modèles de langage » ont fait l'objet de multiples études. En effet, la question récurrente des diverses recherches dans ce domaine a été de connaître comment le locuteur

produit son discours - et par voie de conséquence les disfluences - et de savoir si il en est conscient.

A cet égard, [Eklund, 2004] donne une vision d'ensemble des différents modèles de production qui ont été proposés, et ce dans des domaines aussi divers que la linguistique, la philosophie, la psychiatrie, la neurobiologie, etc. avec différents degrés de spécificités (voir par exemple [Garrett, 1980] ; [Bock, 1982] ; [De Smedt et Kempen, 1987] ; etc.).

Toutefois, malgré la richesse des différentes approches qui ont vu le jour, le premier véritable modèle de production de discours reste généralement attribué à [Levelt, 1989]. Par exemple, tandis que [Bock, 1982] basait essentiellement son modèle sur les phénomènes linguistiques propres à la syntaxe, la sémantique, etc., la proposition faite par Levelt s'est largement concentrée sur des erreurs de discours pour mettre en lumière les processus cognitifs mis en œuvre par le locuteur lors de sa production. Nous choisissons de présenter ce cadre d'analyse en premier lieu dans la mesure où une grande partie des modèles présentés ensuite découlent (jusqu'à un certain point) des travaux de Levelt. Ce modèle consiste en un jeu de divers modules tels que représentés sur le schéma suivant :

Dans un premier temps, le processus de conceptualisateur produit un message. Ce module a accès à un certain type de modèle de discours, un modèle de situation, etc. en fonction de la situation d'énonciation dans laquelle se trouve le locuteur. Par exemple, le locuteur ne mettra pas en œuvre le même modèle de situation pour produire son message selon qu'il s'adresse à une personne qui lui est familière ou à un supérieur hiérarchique sur son lieu de travail.

A partir du message préverbal (ou pré-linguistique) qui résulte de cette première phase de génération, le message est transféré vers le module suivant, le formulateur, qui rassemble les unités linguistiques appropriées (mots, sons, intonation, etc.), et

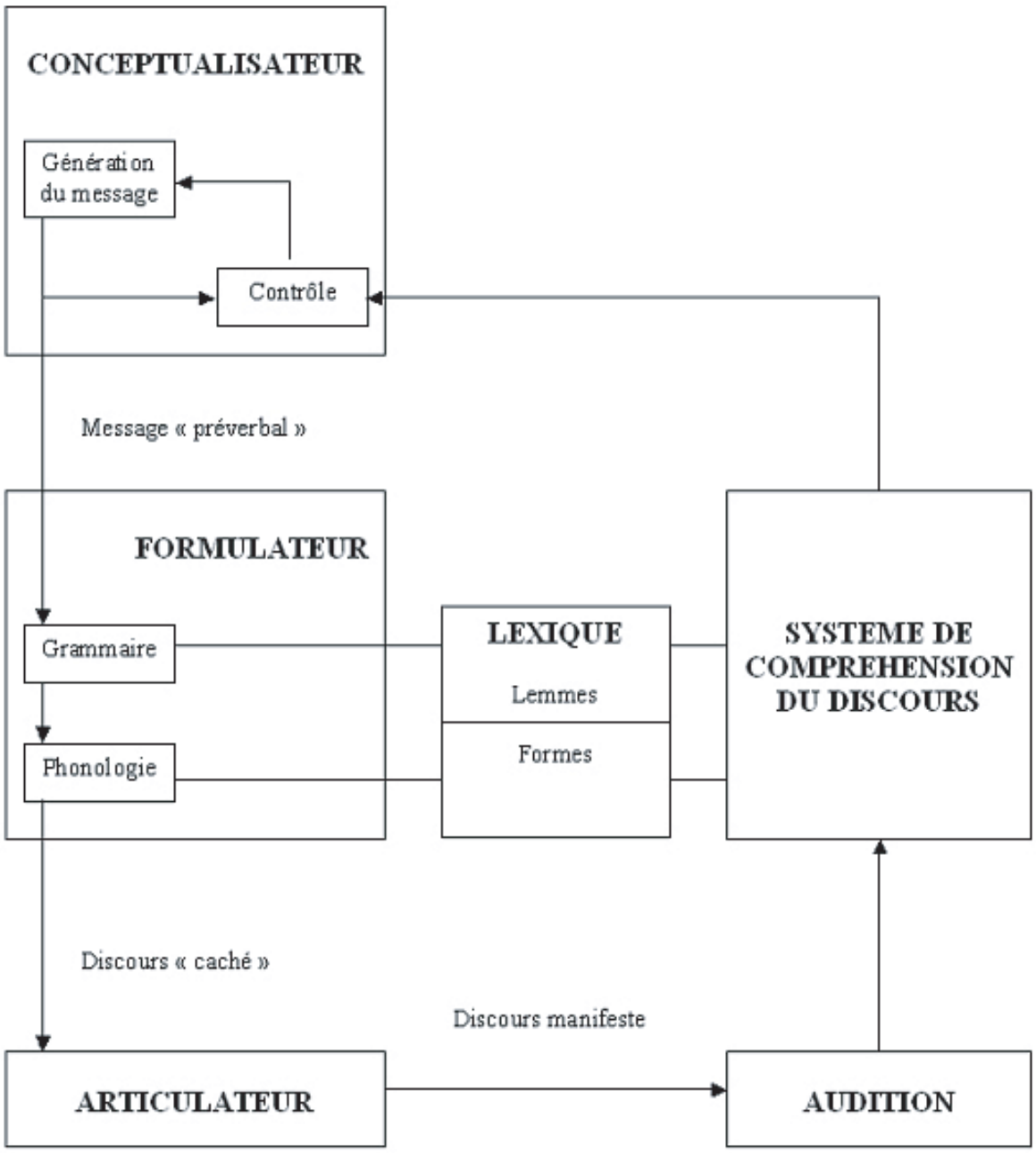


FIG. 4.1 – Version simplifiée du modèle de discours de [Levelt, 1989] (Emprunté puis traduit de [Eklund, 2004]).

qui fournit le message avec la forme linguistique requise. Ce module permet donc

de mettre en « forme orale » le concept précédemment généré.

Le formulateur contient deux sous-modules :

- Un pour le codage grammatical ;
- Un autre pour le codage phonologique, afin de mettre en place une structure linguistique de surface du message qui doit être prononcé.

Cette forme linguistique est alors incorporée au module suivant – l'articulateur – qui met en oeuvre de façon appropriée l'ensemble des articulateurs (les muscles) nécessaires à la production (physique) du message.

Ces premières étapes sont relativement simples. Ce qui est plus crucial et qui permet d'intégrer les disfluences dans ce schéma d'analyse, est que le modèle inclut aussi une fonction de contrôle (ou auto-contrôle). Celle-ci peut être considérée comme un processus propre au locuteur qui surveille la phase de production de discours et qui détecte puis répare, quand il le peut, ses éventuelles erreurs.

Dans le modèle de Levelt, le contrôle exercé par le locuteur sur son discours possède deux traits importants :

- Il est situé au sein même du conceptualisateur. Cela signifie que la génération et le contrôle de celle-ci s'opère simultanément dans le même module.
- Il se sert du système de compréhension de discours, c'est-à-dire que le système que nous mettons en œuvre pour comprendre les autres est également utilisé pour interpréter notre propre discours.

Le contrôle n'exige pas d'entrée acoustique, mais peut s'effectuer à un moment donné antérieur, au niveau cognitif. Il fonctionne alors au moyen de deux boucles différentes :

- Une boucle extérieure qui se sert du signal acoustique ;
- Une boucle intérieure qui suit la trace du message tout au long du processus de production.

La boucle intérieure permet de détecter des erreurs avant que quoi que ce soit n'ait été transmis au niveau du formulateur, c'est-à-dire que les messages peuvent déjà être modifiés au niveau du conceptualisateur. Cette remarque renvoie à la notion de « rédaction » présentée par [Hockett, 1973] (terme original « editing ») pouvant être manifeste lorsqu'elle est énoncée et ensuite corrigée en cas d'erreurs, ou « cachée » lorsque les erreurs sont corrigées avant même d'être énoncées.

En ce qui concerne la mise au point d'un modèle pour l'analyse des phénomènes disfluents, les premières réflexions notables une fois de plus issues des travaux de Levelt (1983) considèrent déjà les disfluences non pas comme des phénomènes sans organisation précise, mais possédant une structure syntaxique intrinsèque.

L'auteur s'est notamment intéressé à ce qu'il nomme « mécanisme d'autocorrection » (qui regroupe les cas de répétitions, amorces et autocorrections de notre typologie). À ce propos, il explique que, syntaxiquement parlant, un énoncé et sa correction (ou « réparation ») se rapprochent du mécanisme de coordination, et en suit les mêmes règles. La fonction de ce mécanisme consiste ici à conserver ou restaurer les aspects formels pertinents de l'énoncé pour que celui-ci soit syntaxiquement correct.

Il propose par ailleurs de représenter ce phénomène en étudiant la nature de la relation entre l'énoncé original produit et la partie « corrigée » (correspondant à la région disfluente). Il part du postulat d'une notion de bonne formation d'une réparation qui est, de plus, déterminée par le contexte :

> " *Just as we have intuitions about the well-formedness of sentences, we can have rather strong feeling about whether a repair « fits » or does'nt « fit »* "

Les exemples suivants illustrent le contraste entre un énoncé donnant lieu à une réparation « mal formée » (a) et une réparation « bien formée » (b) (le point de

départ de la réparation est noté #) :

a) *l'homme avec les lunettes a poussé le clown # avec les moustaches a poussé le clown*

b) *l'homme a donné un coup de poing au # une gifle au clown*

Il a ainsi proposé une convention pour représenter la bonne formation des corrections qui est, comme évoqué précédemment, déterminée par le contexte :

> " *An original utterrance plus repair <OR> is well formed if and only if there is a string C such as the string <OC or R> is well-formed, where C is a completion of the constituent directly dominating the last element of O [...]* "

En d'autres termes, cette règle indique que l'énoncé et la réparation qui lui est associée doivent suivre la règle de coordination syntaxique.

Ce concept de bonne formation aboutit à la mise en place d'un schéma détaillé pour représenter la structure des énoncés disfluents. Le fonctionnement de l'autocorrection est présenté comme une organisation en trois étapes :

- Le locuteur contrôle sa propre parole et l'interrompt lorsqu'il rencontre un problème.

  c'était l'époque où il y avait pas de + *comment dire* + *de de charges sociales*

- Le locuteur produit des hésitations, des pauses, et/ou des « commentaires rédactionnels » (*comment dire*, *je veux dire*, *disons*, etc.)

  c'était l'époque où il y avait pas de + **comment dire** + *de de charges sociales*

- Le locuteur met en place la correction elle même.

  c'était l'époque où il y avait pas de + *comment dire* + **de de charges sociales**

[Levelt, 1983] identifie donc trois zones principales dans ce type d'énoncé : l'énoncé original (OU pour « original utterance »), la phase d'édition, et la correction (R pour « repair »). Ces « places » dans l'énoncé sont ensuite plus détaillées tel que l'illustre le schéma suivant (emprunté et traduit de [Lickley, 1994]) :

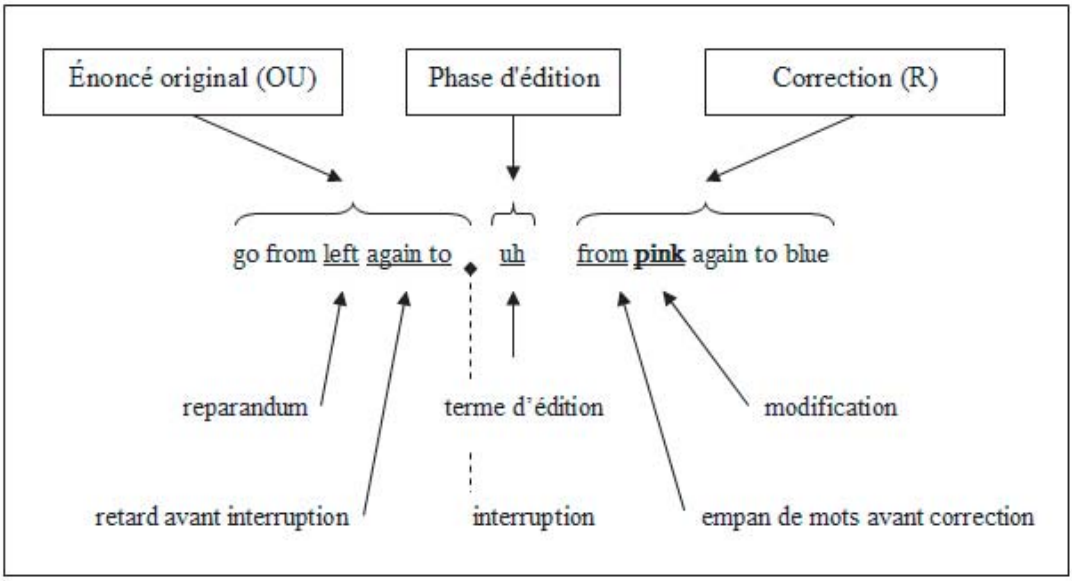


FIG. 4.2 – Schéma de la disfluence ([Levelt, 1983]).

Les résultats des différents travaux Levelt tendent à montrer que l'auto-contrôle exercé par le locuteur s'effectue au moment de la compréhension de son propre discours (discours en attente ou déjà produit) et qu'il n'a pas forcément besoin d'un feedback de la part de son interlocuteur. L'analyse de Levelt peut se résumer en disant que l'auto-réparation ressemble à la coordination en ce sens qu'elle est soumise à des principes de bonne formation similaires, et qu'elle est associée à une convention d'interprétation par les locuteurs qui leur est propre.

4.2.2 Modèles dérivés de l'approche de Levelt

Les réflexions menées par Levelt ont amené plusieurs auteurs à se pencher sur la problématique de la structure formelle des disfluences. Certains d'entre eux ont apporté plusieurs modifications importantes, voire remis en cause une partie du modèle initial, contribuant ainsi à renouveler la façon d'appréhender les phénomènes de production (par exemple Blackmer et Mitton (1991) [en ce qui concerne la catégorisation des disfluences] ou encore Van Wijk et Kempen (1987), Allwood et al. (1987), etc.)

D'autres, en revanche, ont conservé une approche nettement inspirée des conceptions originales présentées ci-dessus, pour une finalité similaire : dégager une structure propre aux énoncés disfluents.

Reparandum / Interruption point / Interregnum / Repair

Shriberg (1994) - reprenant les travaux de Levelt (1983) - propose un cadre d'analyse plus spécifiquement prévu pour l'analyse des disfluences. La structuration proposée consiste à « gommer » les phénomènes de production permettant ainsi de réduire les énoncés à un oral " propre ", proche de l'écrit.

Exemple :

> **they** they basically reviewed **oregon's plan or** the oregon plan toward **uh** nationalizing health care
>
> ⟶ they basically reviewed the oregon plan nationalizing health care
>
> ([Shriberg, 1994])

Dans l'exemple ci-dessus, la suppression des disfluences (répétition [*they they*], autocorrection [*oregon's plan or the oregon plan*] et pause remplie [*uh*]), rend ainsi l'énoncé analysable ultérieurement de façon analogue aux productions écrites.

Cette approche décrit l'organisation interne des disfluences en un ensemble d'es-

paces distincts délimitant les étapes de la production orale. Nous reprenons ici la terminologie proposée par l'auteur :

- Le *reparandum* (RM) : il s'agit de la partie qui sera abandonnée du profit de la réparation (repair).
- Le *point d'interruption* (IP) : celui-ci établit la frontière finale du reparandum et marque une rupture dans la fluidité du discours.
- L'*interregnum* (IM) : il désigne la région comprise entre la frontière finale du reparandum et la frontière initiale du repair. L'interregnum peut contenir un terme d'édition (éditing term) qui peut renvoyer à une pause remplie (euh) ou un commentaire épilinguistique (hein, tu vois, disons pour simplifier, je me rappelle plus du nom, je sais pas moi).
- Le *repair* (RR) représente la partie corrigée du reparandum et qui marque le retour à la " fluence " du discours.

Exemples :

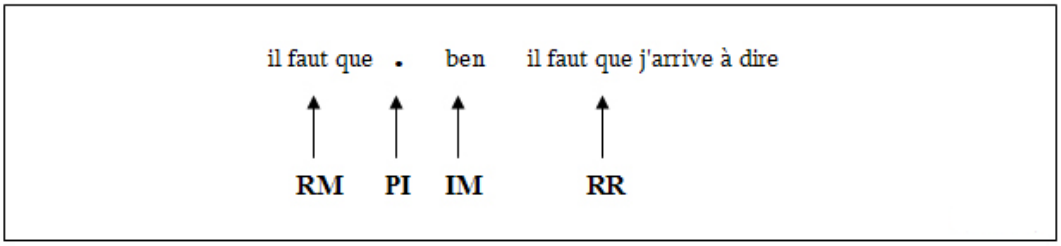


FIG. 4.3 – Structure de la disfluence selon [Shriberg, 1994].

Il est donc ici question de segmenter les disfluences en différentes régions. Cette segmentation est liée à des modifications des propriétés acoustiques et phonétiques de l'énoncé. Celles-ci peuvent par exemple correspondre à un allongement des syllabes avant le point d'interruption ou, à l'inverse, un raccourcissement des syllabes lorsque le locuteur détecte une erreur au cours de sa production.

Perception des disfluences

Tandis que [Shriberg, 1994] s'attache à décrire les régularités formelles des disfluences, [Lickley, 1994] propose une approche qui, bien qu'inscrite dans un champ d'étude similaire, vise à examiner les mécanismes mis en oeuvre par les locuteurs pour détecter les phénomènes de production. Plus précisément, l'auteur chercher à savoir comment les locuteurs sont capables de filtrer les discontinuités du signal de parole de façon si efficace, à tel point qu'ils passent la plupart du temps inaperçus pour celui-ci. Lickley explique que les modèles d'analyse informatique font le plus souvent l'hypothèse que le problème des disfluences se pose essentiellement au niveau de l'analyse syntaxique : tous les mots sont reconnus en amont de l'analyse syntaxique et les patrons de mots reconnus (et/ou les catégories syntaxiques) qui sont probablement des disfluences peuvent ainsi être facilement identifiés.

Cette idée constitue une des principales objections qu'il oppose à ces conceptions. Selon l'auteur, dans l'optique d'un modèle de traitement humain, on ne peut faire de telles suppositions ; lorsqu'un auditeur perçoit une disfluence au milieu d'un énoncé, il n'obtient que des informations partielles sur la syntaxe de ce qu'il entend et il est probable qu'il ne puisse pas « anticiper » tous les mots aux alentours de la zone interrompue de l'énoncé.

> a) *on savait jamais trop à quelle heure c'était et les* (là il était pas question de sortir)
> b) *le dernier jour il a* (il s'est baigné quand même)

Dans l'énoncé (a), on pourrait légitimement attendre un syntagme nominal (introduit par *les*) ; or, aucun élément ne peut permettre à l'auditeur de prévoir que le locuteur va laisser la première partie de son énoncé inachevée pour ensuite changer de structure. De même dans (b), le locuteur ne perçoit qu'une partie d'un syntagme verbal amorcé par *il a* qui est finalement modifié.

De plus, dans le cadre de ses expériences, l'auteur, de même que [Hirschberg et Nakatani, 1994],

simplifie davantage la structure des disfluences que ne le faisait déjà [Shriberg, 1994] à partir du schéma de [Levelt, 1983] :

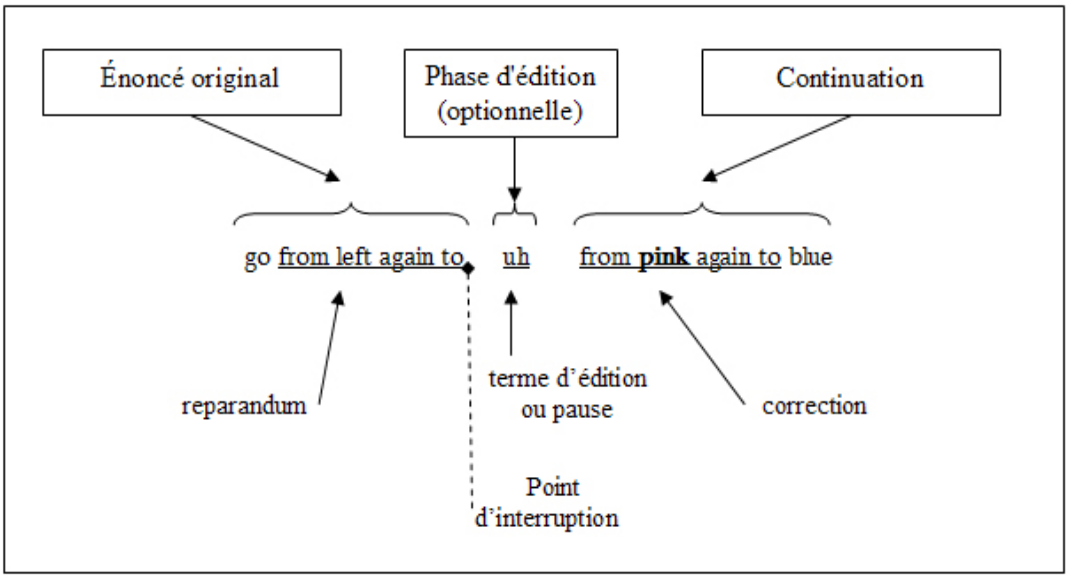


FIG. 4.4 – Structure de la disfluence simplifiée par [Lickley, 1994].

Il s'agit dans cette représentation de s'attacher plus précisément aux contextes droit et gauche du point d'interruption. En effet, selon [Lickley, 1994] il convient, pour traiter les disfluences, de pouvoir faire référence à chacune de leurs caractéristiques dans la structure de l'énoncé où elles apparaissent. Le point central à identifier portant sur l'interruption, il serait bon d'être capable de se baser sur les segments de discours avant (reparandum) et après (correction) l'interruption.

Les modèles de langage en psycholinguistique cherchent aussi bien à comprendre les mécanismes mis en IJuvre par le locuteur dans la production de son discours, qu'à prédire les différents phénomènes de disfluences par contraste avec les séquences de mots non disfluentes. Cependant, de même que les travaux antérieurs sur la détection des disfluences, ils ne donnent pas d'indications sur la manière dont les disfluences doivent être traitées dans une perspective d'analyse syntaxique (à

d'autres niveaux que celui des données en entrée du système) ou même sur un plan plus général d'analyse linguistique.

4.3 Modèles linguistiques

En plus des conceptions purement psycholinguistiques que nous venons de présenter, différentes études descriptives ont été menées afin de caractériser les principaux aspects linguistiques des disfluences, comme par exemple leurs propriétés acoustiques et prosodiques ([Hockett, 1973] ; [Hirschberg et Nakatani, 1994]). D'autres auteurs se sont plutôt centrés sur les caractéristiques syntaxiques de ces phénomènes.

4.3.1 Principe du bord droit

Dans une mouvance plus linguistique que les travaux précédents, d'autres auteurs ont tenté de rendre compte du caractère syntaxiquement régulé des disfluences. [Fornel et Marandin, 1996] suggèrent de traiter les auto-réparations[1] à l'aide de ce qu'ils appellent le principe « du bord droit » selon lequel toute séquence de réparation d'un énoncé (constituant R) doit pouvoir se rattacher comme constituant sur le bord droit de l'arbre syntaxique mis en place par la séquence d'origine (constituant O) :

> " *Un constituant R est une réparation licite pour O si et seulement si il est substituable dans le bord droit de l'arbre analysant O.* "

Cette contrainte serait la seule qui pèserait sur les autocorrections pour lesquelles il n'y aurait qu'un seul mécanisme de réparation. Elle devrait permettre à elle seule de les traiter, sans faire appel au sens ou à la référence et sans aucune distinction

[1] En première approximation, les auteurs précisent que les auto-réparations sont des énoncés dans lesquels le locuteur se reprend, soit parce qu'il hésite, soit parce qu'il se corrige, soit parce qu'il se présente comme hésitant ou se corrigeant.

en fonction de la catégorie de la séquence de réparation (mot outil, syntagme, proposition, etc.) à l'inverse de certaines études tels que [De Smedt et Kempen, 1987] qui distinguent différents mécanismes de réparations d'après la catégorie du réparateur : lexical, syntagmatique ou phrastique.

Les auteurs proposent ce cadre d'analyse en opposition aux modèles de [Blanche-Benveniste, 1990] qui ramènent les autocorrections au mécanisme de coordination ou celui de bonne formation contextuellement déterminée de [Levelt, 1983] (dont ils ne reprennent que le mode de représentation).

De Fornel et Marandin ont ainsi emprunté à [Levelt, 1989] un formalisme spécifique pour illustrer ce principe du " bord droit " :

- L'énoncé interrompu est appelé énoncé origine et désigné par O ;
- L'énoncé suivant O est appelé constituant réparateur : il est désigné par R ;
- L'interruption est représentée par le symbole #.

Dans cette approche, la suite **O # R** ne peut pas être analysée comme une structure de coordination telle qu'on la retrouve par exemple dans les conceptions de [Blanche-Benveniste, 1990] . Le constituant O est ici un segment analysable comme une configuration syntaxiquement bien formée même si quelques « sous-constituants » sont manquants. Le segment R forme une seule unité syntaxique pouvant être de niveau lexical, syntagmatique ou phrastique.

a) [$_O$ *on peut apprendre un certain nombre de*] # [$_R$ *de choses euh + dans la société*]

b) [$_O$ *le grand cru que : +*] # [$_R$ *qui est connu*] # [$_R$ *qui est mondialement connu /*]

Le principe du bord droit permet aux auteurs de justifier qu'on ne puisse pas observer des énoncés disfluents tel que :

**on peut apprendre un certain nombre de de choses euh + dans la société # d'informations*

dans la mesure où la réparation doit pouvoir être ici substituable au bord droit de la séquence d'origine. Afin d'illustrer ce principe, l'énoncé a) peut être représenté de la façon suivante[2] :

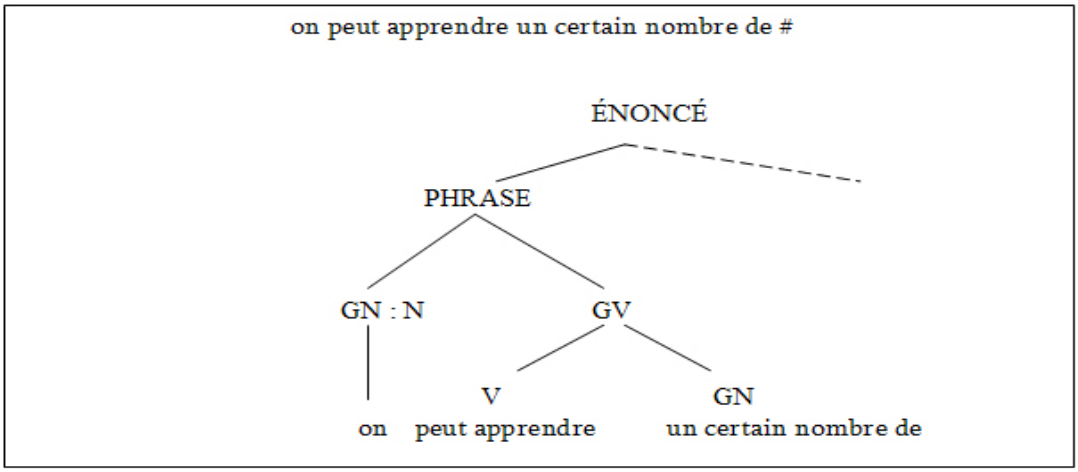


FIG. 4.5 – Illustration du principe du bord droit (1).

Le principe du bord droit délimite alors l'ensemble des catégories possibles pour R : S, GV, GN ou N. La figure suivante illustre quelques unes de ces possibilités (les segments dans la colonne de droite sont des réparateurs possibles pour l'exemple en question).

[2]Nous reprenons, pour illustrer, les catégories de « phrase » et d' « énoncé » admises par les auteurs.

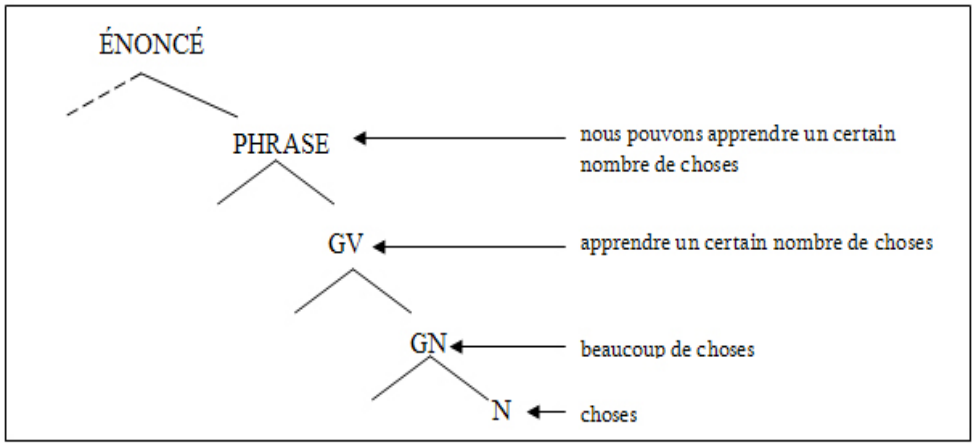


FIG. 4.6 – Illustration du principe du bord droit (2).

Les réparations ci-dessus donneraient respectivement des énoncés tels que :

- *on peut apprendre un certain nombre de* # **nous pouvons apprendre un certain nombre de choses**
- *on peut apprendre un certain nombre de* # **apprendre un certain nombre de choses**
- *on peut apprendre un certain nombre de* # **beaucoup de choses**
- *on peut apprendre un certain nombre de* # **choses**

qui correspondent ainsi à des réparations licites du point de vue de ce cadre d'analyse. Les auteurs précisent en outre que la configuration syntaxique de R peut, de même que O, être interrompue et donner lieu ainsi au phénomène de réparation en cascade.

Exemple :

> *il y eu # il y a pas eu de # comme on le sait il y a pas eu d'effusion de sang il y a pas* le : ++ # comment te dire il y a pas le truc du pressing quoi # ils sont pas spécialisés sur le pressing*

La relation de dépendance entre O et R implique dans ce modèle que ni la catégorie, ni la position de R ne sont libres.

Nous conservons de ce modèle l'idée selon laquelle il est possible d'assigner un statut grammatical aux énoncés disfluents. En effet, la démonstration proposée dans l'article de De Fornel et Maradin (1996) permet de comparer des structures repérées en corpus avec celles générées par le dispositif grammatical « classique » tel qu'on le retrouve à l'écrit.

Ainsi, partant de leurs réflexions théoriques, les auteurs soulignent le fait qu'en général les « auto-réparations » sont facilement décodables par le locuteur. Or, les résultats de [Fox Tree, 1995] tendent à montrer au contraire que ces « autocorrections » provoquent bel et bien un retard dans le décodage. Il est alors légitime de penser que ce retard dépend de la complexité de la structure abandonnée, et éventuellement du nombre d'autocorrections successives produites par le locuteur. Bien que, la plupart du temps, la présence d'autocorrections dans la parole ne pose pas de problèmes insurmontables à l'auditeur pour reconstruire le sens de ce qu'il perçoit, il peut toutefois arriver que leur multiplication rende impossible cette construction du sens.

4.3.2 Cadre d'analyse dédié : la répétition et l'autocorrection

Un cadre d'analyse n'est pas systématiquement prévu pour appréhender les disfluences dans leur globalité. L'exemple que nous présentons ici rend compte de la structuration des différents phénomènes de manière distincte les uns par rapport aux autres : dans sa thèse, [Candéa, 2000b] propose, entre autres, une définition de la répétition disfluente afin de présenter le mode de fonctionnement de ce phénomène :

" *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 dans 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* ".

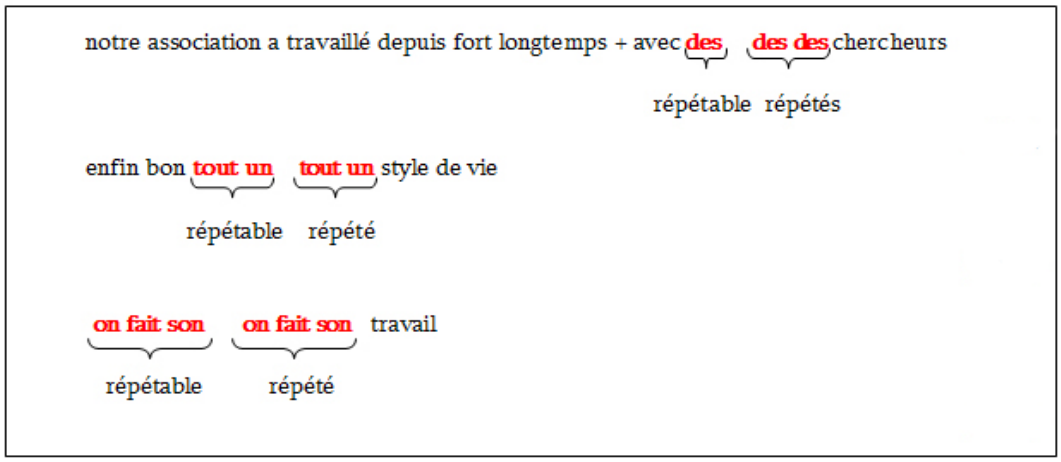


FIG. 4.7 – Exemple de répétable et répété(s).

La définition précédente peut être représentée à l'aide du schéma ci-dessous :

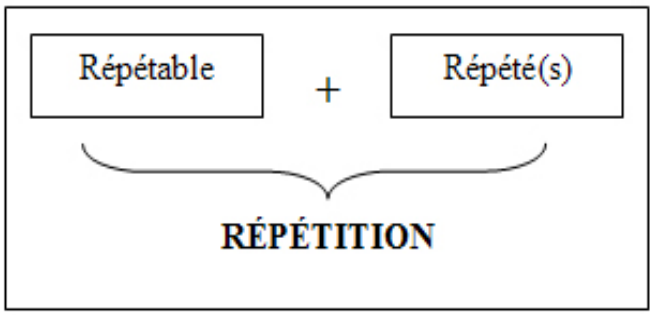


FIG. 4.8 – Schéma structurel de la répétition ([Candéa, 2000b]).

Dans le même ordre d'idée, l'auteur propose également de catégoriser le phénomène d'autocorrection immédiate (*i.e* portant uniquement sur des mots outils) en se basant sur la nature de la séquence corrigée.

Sont considérées comme autocorrections immédiates dans ce modèle, toute séquence XY où Y remplace X, et Y corrige un seul et unique trait phonétique ou morphologique de X, ou opère un changement de catégorie syntaxique sur X.

Exemples :

les boulangers vivent au-dessus **de du** magasin
X Y

il y a un petit panneau **sur un sur le** mur aussi
X Y

on sait pas quand euh **on a du mal à prévoir quand** on aura + euh des offres de céréales
X Y

Cette modélisation peut être illustrée de la façon suivante :

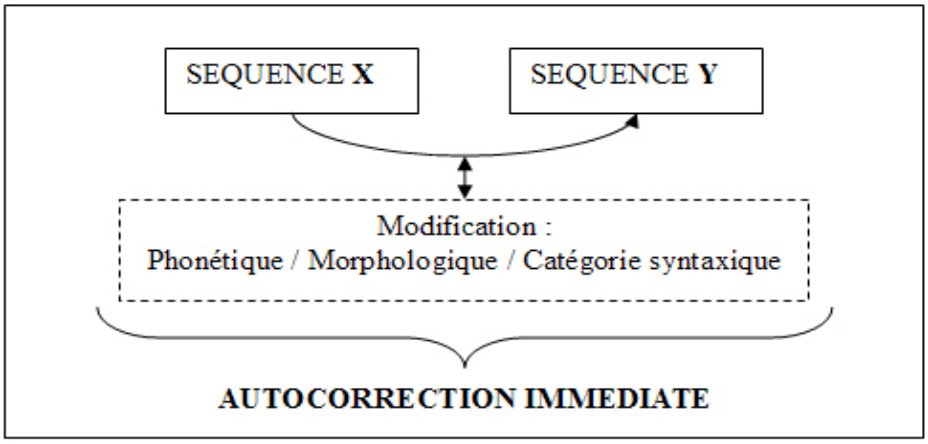


FIG. 4.9 – Schéma structurel de l'autocorrection immédiate.

L'auteur précise que ce type d'autocorrection présente ainsi un fonctionnement très semblable à celui des répétitions, d'autant qu'elles portent sur le même type d'unités que sont les mots outils.

Le manque principal de ce type d'approches provient du fait qu'elles ne bénéficient, à notre connaissance, d'aucune application pratique dans une architecture de traitemen automatique. L'objet de la section suivante est justement de présenter les approches développées dans cette optique.

4.4 Modèles théoriques en T.A.L

Au delà des conceptions psycholinguistiques ou linguistiques que nous venons d'aborder, il existe également divers cadres d'analyse appliqués au TAL (notamment dans le domaine de la syntaxe) qui ont tenté à leur tour de formaliser les phénomènes de disfluences en détaillant autant que faire se peut les mécanismes propres à leur structure.

Plusieurs travaux ont porté sur la détection et la correction des phénomènes de production[3]. Parmi les premiers travaux, nous présentons ici les travaux de [Carbonell et Hayes, 1983] ainsi que [Hindle, 1983]. Nous nous intéresserons plus loin à des études plus récentes, en tentant de les catégoriser selon différents types d'approches.

4.4.1 Patrons simples

Parmi les travaux précurseurs en TAL sur l'oral spontané, [Carbonell et Hayes, 1983] suggèrent l'utilisation de patrons simples pour traiter certains cas de disfluences. Les auteurs proposent trois méthodes (sous forme de règles syntaxiques) pour dé-

[3]Par ailleurs, des travaux récents ont porté sur la production des disfluences dans le cadre d'un moteur de génération incrémental ([Finkler, 1997]).

tecter la partie erronée des énoncés interrompus puis recommencés par le locuteur.

– (1) Lorsqu'une séquence de deux constituants avec un type syntaxique et sémantique identique est détectée où un seul est autorisé, il convient d'ignorer le premier.

  Exemple :

  je pense que ça c'est facile **à : +** **à** *se rappeler*

  ⟶ *je pense que ça c'est facile* **(à : +)** **à** *se rappeler*

  ⟶ *je pense que ça c'est facile* **à** *se rappeler*

– (2) Reconnaître les phrases correctives explicites[4] (telles que *je veux dire* par exemple), et si le constituant à droite de celles-ci est du même type syntaxique et sémantique que celui de gauche, alors il faut ignorer le constituant de gauche et la phrase corrective.

  Exemple :

  et cet élément **était encore euh je veux dire était un peu** *+ délimité*

  ⟶ *et cet élément* **(était encore euh je veux dire) était un peu** *+ délimité*

  ⟶ *et cet élément* **était un peu** *+ délimité*

– (3) En faisant de telles corrections (*i.e* 1 et 2), sélectionner le constituant minimal à gauche qui doit être ignoré.

  Exemple :

  bon voilà mais vous auriez pu **évoluer enfin** *avoir envie de faire autre chose*

  ⟶ *bon voilà mais vous auriez pu* **(évoluer enfin)** *avoir envie de faire autre chose*

[4]Pour reprendre la terminologie employée par les auteurs.

⟶ *bon voilà mais vous auriez pu* **avoir envie de faire autre chose**

Par ailleurs, ces règles peuvent intervenir séquentiellement pour traiter les exemples ou deux cas de figure ses présentent simultanément :

euh les les comment les compositions seront pas du tout pareilles en fait

Application de la règle (1) :

euh **les les** *comment les compositions seront pas du tout pareilles en fait*
⟶ euh **les** comment les compositions seront pas du tout pareilles en fait

Application de la règle (2) :

euh **les comment les** *compositions seront pas du tout pareilles en fait*
⟶ *euh* **les** *compositions seront pas du tout pareilles en fait*

La règle (1) peut être considérée comme un cas particulier de la règle (2) où il n'y aurait pas de phrase corrective.

Or, cette règle semble trop générale dans la mesure où elle peut s'appliquer (peut-être à tort) à des cas d'amplifications ou de précisions tels que :

c'est **difficile** *à expliquer* + **très difficile** *hein à expliquer*

tous ces **cépages** + *ces* **nouveaux cépages** *mûrissent très bien*

Un autre problème dans ce modèle vient de la nature « fragmentaire » de la plupart des corpus oraux spontanés ; dès lors, il n'est pas toujours facile dans un exemple donné de décider qu'un seul constituant est autorisé ou non.

De plus, les phrases correctives sont choisies à partir d'un petit ensemble de phrases socialement déterminées et arbitraires plutôt que d'après une règle générique pour l'ensemble de ces phrases. En l'absence de traits prosodiques, ces phrases correc-

tives peuvent induire des difficultés d'analyse syntaxique : comment savoir si *je veux dire* est une phrase corrective ou le début d'un nouvel énoncé ?

Enfin, les règles proposées par les auteurs ne traitent pas le cas des constructions inachevées. Leur étude laisse également très vague la notion d' « identité syntaxique et sémantique ». Par exemple, le terme « identique » ne peut pas impliquer une identité au niveau du nombre (singulier/pluriel) puisque des disfluences telles que

c'est **un** *c'est* **des** *fonctions qui existent*

le *euh* **les** *choses évoluent*

apparaissent de façon tout à fait triviale et récurrente dans les corpus et ne seraient alors pas prises en considération dans ce modèle.

4.4.2 Approche syntaxique déterministe

Dans un champ d'investigation similaire, [Hindle, 1983] propose une approche syntaxique déterministe (via un ensemble de règles) destinée à résoudre les autocorrections dans les corpus de parole. Il propose l'implémentation d'un algorithme déterministe pour des applications nécessitant une phase d'analyse grammaticale. A la différence de [Carbonell et Hayes, 1983], l'auteur met en avant le fait que les points d'interruptions dans les énoncés disfluents sont déjà détectés.

Il s'intéresse à la portion de discours qui est remplacée lorsque le locuteur corrige à la façon d'une « zone de suppression », car tout le matériau linguistique dans cette portion de discours doit être « exclue » de l'analyse syntaxique ultérieure. Hindle introduit l'idée selon laquelle les disfluences sont toujours signalées par une « marque de correction » discrète au niveau du point d'interruption.

Partant de ce principe, il avance donc un certain nombre de règles (que nous ne listons pas intégralement ici) dont :

(1) Règle de « reproduction de surface »

Il s'agit d'une règle pré-syntaxique chargée d'identifier les séquences de mots identiques de chaque côté de la marque de correction. Dans cette règle, les séquences identifiées ne doivent pas obligatoirement correspondre à des constituants complets.

Exemple :

et euh + et ensuite

⟶ **et** *euh* + **et** *ensuite* = séquence identique : « et / et »

ils ont ils ont créé le tout de A à Z

⟶ **ils ont ils ont** *créé le tout de A à Z* = séquence identique : « ils ont / ils ont »

(2) Règle de « reproduction de catégorie »

Soit X + signal de correction + X, où X est un constituant, supprimer le premier X ainsi que la marque de correction.

Exemple :

ils ont passé une heure de + comment dire de concert à la télé

⟶ *ils ont passé une heure* $\underline{\textbf{de}}_{X_1}$ **+ comment dire** $\underline{\textbf{de}}_{X_2}$ *concert à la télé*

⟶ ils ont passé une heure **de** concert à la télé

(3) Règle de « reproductions "empilées" »

Soit X inachevé + signal de correction + X, où X est un constituant, supprimer le premier X inachevé ainsi que la marque de correction.

Exemple :

on a des écoles euh en b- en breton

$\longrightarrow$ on a des écoles euh **en b-** **en breton**

X_1 X_2

$\longrightarrow$ on a des écoles euh **en breton**

Les deux premières règles sont probablement superflues tout comme la « règle de reproductions empilées » qui rapportera les mêmes résultats (cf. analyse de [McKelvie, 1998]). Les règles semblent en fait ambiguës, à tel point que pour une disfluence donnée, celle-ci peut syntaxiquement être analysée de diverses manières. La règle appliquée doit probablement dépendre de l'ordre d'application des règles au sein de l'analyseur.

Hindle soutient l'idée que l'étude des autocorrections en utilisant les susdites règles bâtit une structure qui permet, pour l'étude des disfluences, de renseigner sur la mise en oeuvre des caractéristiques syntaxiques de la langue. L'auteur avance ainsi l'hypothèse que, " *loin d'être un obstacle à l'apprentissage de la langue, les disfluences pourraient en fait faciliter l'acquisition de celle-ci en mettant en évidence les classes de constituants qui sont équivalentes* ".

Cette caractéristique de la parole disfluente est décrite dans cette conception comme une " *coupure brutale manifeste du signal de parole* " même si certains auteurs (voir notamment [Schegloff *et al.*, 1977]) ne considèrent cette « coupure » que comme l'un des nombreux initiateurs potentiels de la réparation. En effet, les aspects purement prosodiques des productions orales par exemple ne sont pas pris en compte, alors que même en l'absence de rupture syntaxique, ceux-ci peuvent renseigner sur la présence imminente d'une réparation

4.5 Conclusion

Bien que n'étant pas un domaine encore fortement prisé en traitement automatique des langues, l'oral, et plus spécifiquement les disfluences, ont fait l'objet de nombreux cadres d'analyse. Le domaine de la psycholinguistique notamment puis de la linguistique et de l'informatique ont permis de rendre compte d'un certain nombre de configurations de ces phénomènes à plusieurs niveaux (perception, production, syntaxe, compréhension, etc.).

Toutefois nous l'avons vu, les premières modélisations de ces phénomènes restent dans la plupart des cas à un niveau théorique. Néanmoins, ces réflexions ont mené plusieurs chercheurs à proposer des architectures en TAL pour résoudre en terme applicatif les problèmes posés par l'oral (principalement dans le domaine de la reconnaissance vocale, pour ne donner qu'un exemple) comme nous l'exposons dans le chapitre suivant.

Nous avons vu que les différentes approches ont des avantages et des inconvénients opposés pour le traitement de l'oral. En effet, les stratégies adoptées dépendent à la fois de l'angle sous lequel les recherches menées (du point de vue de la production ou de la perception des disfluences), du domaine de recherche, mais aussi des théories sous-jacentes qui guident ces approches.

Dans le cadre précis de notre étude, la première question est alors de savoir comment peut-on analyser ces données linguistiques d'un point de vue syntaxique. A ce propos, la problématique du traitement automatique des disfluences a été appréhendée de multiples manières. C'est ce que nous tenterons de voir dans la seconde partie de ce travail où nous étudierons les différentes études effectuées en terme de traitement automatique des disfluences avant de présenter notre conception théorique du fonctionnement de celles-ci.

Deuxième partie

Approche théorique et étude empirique

Chapitre 5

Traitement des disfluences : état de la technique

5.1 Introduction

Le traitement automatique de la parole s'est considérablement développé au cours des trente dernières années, avec des objectifs qui ont évolué avec le temps ([Carre *et al.*, 1991]). Dans les années 1960, le problème majeur était d'augmenter le nombre de communications transmises simultanément sur une seule ligne téléphonique. Face au manque d'infrastructures téléphoniques de l'époque, s'est ajouté le problème du dialogue homme-machine. Pour résoudre cette difficulté, il fallait disposer de nombreuses données orales dans différentes langues. Ce domaine a ainsi atteint durant la dernière décennie un degré de maturité qui s'est traduit par le développement de nombreuses applications commercialisées (dictée vocale, synthèse de la parole, etc.). Les progrès constatés doivent beaucoup à la généralisation de méthodes empiriques basées sur les données. Ce constat est particulièrement vrai en reconnaissance de la parole, où des corpus de grande taille sont utilisés pour l'estimation de modèles stochastiques de langage.

Une des différences pour un système entre traiter un message écrit ou un message oral tient notamment dans la présence d'accents et d'intonations différentes, ou des phénomènes propres à l'oralité (liaisons, modification d'un phonème par les

phonèmes qui l'entourent, etc.) qui rendent les outils statistiques particulièrement utiles pour le traitement de la voix. Malgré ces points de divergence, les concepts développés en TAL (pour l'écrit) restent pertinents en traitement du langage parlé, notamment dans les applications de dialogue naturel. En effet, dans ce cas, les messages parlés sont convertis en données écrites, et ce sont les textes ainsi produits qui sont utilisés par le système pour répondre à l'utilisateur. De même, la réponse produite l'est d'abord sous forme textuelle, avant qu'un synthétiseur vocal ne la transforme en voix.

TAL et traitement du langage parlé sont donc étroitement liés, ce qui explique que de nombreuses applications font intervenir les deux domaines. Il peut s'agir par exemple de la vocalisation de textes écrits pour les handicapés, pour l'apprentissage des langues étrangères ou encore pour des systèmes de traduction de l'oral d'une langue à une autre. D'autres solutions font intervenir des techniques complémentaires telles que la reconnaissance optique de caractères. Ainsi, on peut utiliser un stylo optique que l'on passe sur un journal et qui, soit vocalise les mots reconnus, soit les traduit, soit en donne une définition. Inversement, des logiciels ont été développés pour passer de l'oral à l'écrit, comme les solutions de dictée vocale. C'est par exemple le cas de la société IBM dans les années 90 qui a lancé sur le marché l'un des tout premiers logiciels de dictée vocale en continu.

Pour de telles applications – qui manipulent l'oral – il semble judicieux d'avoir recours, en complément des méthodes actuelles, à une analyse fine des différents niveaux de l'analyse linguistique (lexique, syntaxe, sémantique, etc.). Toutefois, dans le cas du français, peut être du fait du manque de données observables, le langage oral apparaît comme le « parent pauvre » des descriptions syntaxiques.

5.2 Problématique

Les progrès technologiques de l'informatique ont modifié l'étude des langues parlées, en permettant d'établir de grands corpus (reliés à des corpus de langue écrite), dont la taille se mesure en millions de mots, disponibles pour différentes langues. Ces données ont des applications pratiques immédiates : élaboration de grands lexiques sur des données à la fois écrites et orales (par exemple *The Birmingham Collection of English Text*) ; publication de grammaires ; établissement de corpus multilingues parallèles (pour faciliter l'étude contrastive, la traduction et l'enseignement des langues étrangères).

De la même manière, en terme d'applications, les analyses distributionnelles, menées pendant des années en interrogeant des informateurs (en consultant son intuition ou en consultant manuellement des corpus), peuvent désormais se faire par l'établissement de grands concordanciers. Les modestes corpus de français parlé de l'équipe de recherche (environ deux millions de mots) en donnent un aperçu. Néanmoins, l'usage de concordancier n'apparaît pas comme le seul moyen de traiter de tels corpus. Il est également possible d'avoir recours à d'autres types d'outils notamment si l'on s'intéresse à l'aspect syntaxique de ces données. C'est notamment le cas des systèmes de reconnaissance vocale par exemple, où malgré les études menées dans différentes langues sur le sujet (par exemple [Rose et Riccardi, 1999] pour l'anglais, [Hübener *et al.*, 1996] pour l'allemand, etc.), les dispositifs actuels butent généralement sur les données à traiter et n'atteignent des performances satisfaisantes que dans certains domaines restreints. En effet, le taux de reconnaissance varie considérablement selon plusieurs facteurs, tels que le débit de parole du locuteur, le bruit environnant, etc. Les erreurs peuvent consister en l'insertion, la suppression ou la substitution de certains mots. Hormis le débit ou la quantité de bruit, les phénomènes de disfluences jouent un rôle indéniable dans la dégradation

des performances des systèmes de reconnaissance.

Par exemple, lors d'une expérience[1] utilisant un système de reconnaissance vocale dans le domaine du renseignement ferroviaire, on a constaté que le dispositif n'intègrait pas la notion de répétitions ou d'amorces de mots inachevés. Le système fournissait donc de façon erronée une approximation lexicale au mieux de ses possibilités (cf. [Véronis, 2004]) :

non non non non je veux pas de Pa- de de Paris gare d'Austerlitz

a par exemple été reconnu comme

Nancy dans nonante jours à Le Havre de Paris gare d'Austerlitz

Ce cas d'erreur caractéristique peut entre autre s'expliquer par le mécanisme mis en IJuvre par un système de reconnaissance de la parole. L'objectif d'un tel dispositif est de transcrire automatiquement un signal sonore en texte ([Gravier *et al.*, 2006]). Il cherche donc dans un premier temps à reconnaître des mots en utilisant uniquement des critères d'ordre acoustique, sans essayer d'interpréter le contenu véhiculé par l'ensemble de ces mots. Dans un deuxième temps il est question de choisir la meilleure hypothèse en reconsidérant le signal sonore comme une suite de mots porteurs d'informations. L'utilisation de la linguistique apparaît adéquate dans ce contexte puisque le recours à la morphologie, la syntaxe ou la sémantique semblent pouvoir guider la sélection de la meilleure hypothèse.

Les traitements « classiques » en terme d'analyse syntaxique (notamment) s'avèrent efficaces pour l'analyse de l'écrit (taux de précision compris entre 90% et 97% selon le phénomène et l'outil [Guiguet et Vergne, 1997]). Cependant, les traitements appliqués aux corpus oraux sont loin d'être aussi performants.

En effet, comme nous allons le voir ci-après, la majeure partie des études réalisées sur le traitement automatique des disfluences a été menée sur l'anglais ou

[1] http ://lablita.dit.unifi.it/coralrom/

d'autres langues, mais rarement sur le français. Nous pouvons toutefois citer les travaux sur l'aspect syntaxico-sémantique des disfluences, tels que ceux menés par [Goulian *et al.*, 2002] ou [Antoine *et al.*, 2003] dans un système de compréhension automatique de la parole. Les autres études telles que celles de [Adda-Decker *et al.*, 2004] par exemple se concentrent plus sur la problématique de l'annotation de ces phénomènes et on ne trouve guère de traitement syntaxique sur le français. Pourtant, le domaine de l'oral spontané présente un grand nombre de spécificités incontournable dans une perspective d'analyse syntaxique.

De plus, le développement des systèmes de reconnaissance de la parole a fait émerger la nécessité imminente de produire des outils permettant l'analyse linguistique de l'oral. Le traitement de l'oral constitue une réelle difficulté, car bien que de nombreuses études aient été (et sont toujours) faites sur l'oral, et plus précisément sur les différentes disfluences, il manque encore de nombreux détails sur les aspects fonctionnels de la syntaxe de ces phénomènes et leur fonctionnement intrinsèque. En effet, on ne connaît pas véritablement les contraintes des disfluences entre elles, mais aussi vis à vis de la syntaxe. Les précédentes études ont par exemple montré qu'on ne répète pas n'importe quel segment, que les pauses ne sont pas marquées sur n'importe quelle place syntaxique, etc.

En traitement automatique du langage naturel, la plupart des systèmes complexes utilisent parmi leurs fonctionnalités de base un analyseur syntaxique. Comme le rappelle [Villemonte de la Clergerie et Rajman, 2003], l'analyse syntaxique constitue une étape essentielle dans le traitement linguistique dès lors qu'est recherchée une connaissance précise des relations grammaticales au sein d'une phrase. Cela concerne, par exemple, la traduction, la correction grammaticale, voire même l'extraction d'information (fouille de texte) ou les applications de type questions-réponses.

Cette remarque doit également s'appliquer au cas des disfluences, pour lesquelles il

existe différents domaines d'appréhension. Hormis l'aspect syntaxique auquel nous nous intéressons dans cette étude, il convient de souligner que des travaux ont déjà été conduits à d'autres niveaux. Nous pouvons par exemple citer les études menées par [Hockett, 1973], [Hirschberg et Nakatani, 1994], [Shriberg *et al.*, 1997], [Quimbo *et al.*, 1998] ou plus récemment [Liu *et al.*, 2003], [Delais-Roussarie et Choi-Jonin, 2004] qui se sont axées sur le traitement prosodique et intonatif de phénomènes propres à l'oral.

Nous allons ici nous limiter aux approches relativement récentes dans le domaine du traitement automatique des disfluences au niveau de la syntaxe. L'étude qui va suivre devrait nous renseigner sur les possibilités de mise en oeuvre d'une analyse robuste du français parlé spontané.

5.3 Études pratiques pour le traitement des disfluences

La position généralement admise dans les travaux sur le traitement automatique du langage naturel est que les énoncés disfluents sont en tant que tels agrammaticaux et ne peuvent pas être reconnus par un algorithme d'analyse syntaxique conçu pour reconnaître les énoncés canoniques d'une langue ([Fornel et Marandin, 1996]). Ils doivent donc être traités par un module spécifique de pré-traitement de l'entrée syntaxique qui mobilise des règles ad hoc d'ajustement opérant avant celles du module de reconnaissance et/ou d'analyse syntaxique ([Hindle, 1983] notamment).

Il s'agit essentiellement de nettoyer les énoncés des phénomènes de production qu'ils contiennent. La proposition est basée sur l'hypothèse suivant laquelle un segment disfluent est de l'ordre de l'accidentel et qu'elle est régularisable de façon externe à la grammaire en réduisant l'écart entre la structure donnée en entrée et les structures canoniques utilisées comme références. L'analyse que nous propose-

rons plus loin dans notre étude rejoint une conception différente où les segments disfluents ne sont pas exclus d'emblée du domaine de l'analysable.

Nous pouvons dégager deux axes de recherche principaux développés en vue d'apporter des réponses à la problématique du traitement automatique des disfluences. Le premier s'appuie sur l'utilisation d'une grammaire standard en s'assurant, comme nous venons de l'évoquer, de détecter et de corriger ou ignorer les phénomènes de production de façon à soumettre ensuite à l'analyseur un texte proche de l'écrit. Le second à l'inverse, préconise l'utilisation de mécanismes permettant la gestion des variations de l'oral (pouvant baser sa stratégie sur le recours à un analyseur syntaxique de l'écrit) et en les représentant à l'aide d'un formalisme syntaxique préalablement défini au même titre que pour l'écrit.

A partir de telles réflexions théoriques, diverses stratégies d'analyse (différentes selon la tâche à accomplir et le type d'approche) peuvent être adoptées telles que nous les présentons ci-après.

5.3.1 Détection et correction des disfluences avant analyse

Une des premières approches rencontrée en traitement automatique consiste a détecter puis corriger ou ignorer les disfluences (*i.e.* à ne pas les prendre en compte lors de l'analyse [cf. [Heeman et Allen, 1994]) avant de procéder à l'analyse syntaxique des données à traiter.

Standford Research Institute (*SRI) International*

L'approche développée au sein du SRI International ([Bear *et al.*, 1992] notamment), est fondée sur plus de 600 énoncés disfluents extraits du corpus ATIS[2] (Air Travel Information Service), et constitue l'un des premiers travaux sur les

[2]Système d'interaction verbale pour des applications de demande d'information dans le domaine des transports aériens.

disfluences dans un cadre applicatif[3]. La première étape de ce travail consiste à proposer un schème de notation qui vise à allier la simplicité à la finesse nécessaire pour la représentation des différentes formes de disfluences :

- Le point d'interruption est représenté par une barre verticale (|).
- Correspondance identique : pour montrer que deux mots aux deux côtés d'une interruption sont identiques, ceux-ci sont marqués par **M** (pour « matching »).
- Le remplacement : indique le remplacement d'un mot avant le point d'interruption par un mot après. Les deux mots doivent être similaires morphologiquement. En général ils doivent être de la même catégorie ou d'une variante morphologique de celle-ci comme les cas d'amalgames : *I / I'd*.
- Mots neutres : tous les mots dans la zone d'une disfluence sont notés **X**.
- Un tiret (**-**) est ajouté aux signes précédents en cas d'amorce.

Exemples :

I	want	fl-		flights	to Boston
		M1-	**|**	**M1**	

What	what	are the fares	
M1	**|M1**		

Show	me flights		daily		flights
	M1	**|**	**X**		**M1**

Il s'agit ensuite de combiner analyse syntaxique et sémantique (afin de réduire la surgénération des patrons) avec la technique de « pattern matching » ou reconnaissance de patrons (pour détecter les phénomènes simples tels que la répétition d'une séquence de mots comme *I would like a book I would like a flight* ou des

[3] Voir également [Becker *et al.*, 1999] pour le suédois.

anomalies syntaxiques simples comme : *a the*, ou *to from*).

Le système tente donc dans un premier temps d'analyser syntaxiquement et sémantiquement les énoncés qui passent dans un second temps au reconnaisseur de patrons. Dans ce cas, deux décisions peuvent être prises :

- Les parties d'énoncés qui ont été correctement traitées par les modules d'analyse syntaxique et sémantique et qui sont signalées comme étant disfluentes par le reconnaisseur de patrons sont considérés comme des surgénérations.
- Les parties d'énoncés partiellement analysées par les modules linguistiques et qui sont signalées par le reconnaisseur de patrons comme étant disfluentes sont considérées réellement comme telles.

L'inconvénient principal de cette combinaison est qu'elle est incompatible avec les approches d'analyse partielle qui sont les plus adaptées au traitement de l'oral, et amène au dilemme suivant : d'une part, l'utilisation d'une méthode d'analyse partielle empêche de juger la grammaticalité d'un énoncé et par conséquent rend ce type de combinaison impossible. D'autre part, les méthodes d'analyse classiques sont bien adaptées pour juger de la grammaticalité (tous les énoncés analysés sont entièrement corrects grammaticalement) mais échouent souvent face au traitement de phénomènes syntaxiques propres ou fréquents à l'oral comme les problèmes d'accord, les ellipses, etc. Les résultats obtenus par [Bear *et al.*, 1992] pour la correction des disfluences sont de 43% pour le rappel et 50% de précision.

Approche stochastique à base de patrons

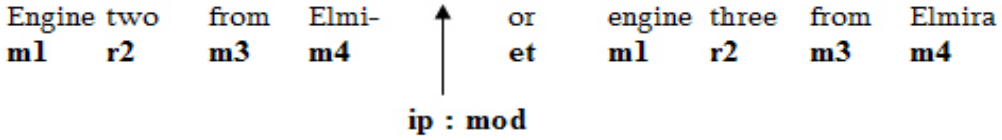
[Heeman *et al.*, 1996] proposent une approche stochastique dans le cadre du projet américain TRAINS[4] à l'université de Rochester. Dans un premier temps, Heeman a proposé une version modifiée du schème d'annotation du SRI :

- Le point d'interruption est marqué à l'aide du symbole « **ip** ».

[4][Allen et Schubert, 1991].

- Au niveau du point d'interruption, une série de suffixe est utilisée pour marquer le type de disfluence comme : « **mod** » (*modification*) pour les corrections, « **can** » (*cancels*) pour les faux-départs, et « **et** » (*editing terms*) pour les mots d'édition.
- Les cas ambigus sont marqués par un « + » à la fin.
- Chacun des mots est étiqueté « m_n » (où n est l'identifiant du mot), et « **r** » pour les mots qui font l'objet de la correction (mots de même catégorie syntaxique).

La différence principale entre le schème de Heeman et celui de [Bear *et al.*, 1992] est que celui de Heeman ne permet pas le partage de la zone remplacée dans le cas de disfluences imbriquées.



Suite à l'annotation des disfluences, Heeman obtient près de 1300 cas d'énoncés disfluents avec 160 structures différentes ([Heeman, 1997]). Afin d'éviter la surgénération de certains patrons, Heeman propose un ensemble de règles destinées à les contraindre ([Heeman et Allen, 1994]). Ces règles portent principalement sur la forme de la zone d'édition et sa localisation par rapport au point d'interruption d'une part, et le reste de la disfluence d'autre part.

Dans leur article [Heeman et Allen, 1994] présentent un cas de disfluence imbriquée et montrent très brièvement la façon dont leur système la traite sans donner aucune information sur le mécanisme de contrôle sous-jacent qui est pourtant le point clé dans ce type de configuration. À cet égard, [Kurdi, 2003] déduit que pour ces cas précis le système réinitialise le traitement à chaque détection et correction d'une

disfluence.

Les résultats de l'auteur sont présentés dans le tableau ci-dessous :

Phénomène	Actions	Rappel	Précision
Discontinuïtés	Détection	77,88%	82,51%
	Correction	75,65%	82,26%
Réparations	Détection	80,87%	83,37%
	Correction	77,95%	80,36%
Faux-départs	Détection	48,58%	69,21%
	Correction	36,21%	51,59%
Total	**Détection**	**76,79%**	**86,66%**
	Correction	**65,85%**	**74,32%**

TAB. 5.1 – Résultats obtenus par [Heeman, 1997] sur la détection et la correction des disfluences

Comparativement aux résultats de [Bear *et al.*, 1992], le travail de Heeman présente une avancée indéniable. Toutefois, cette avancée est somme toute relative étant donné que les deux approches n'ont pas été testées sur le même corpus de test, et ne définissent pas de la même manière les différents phénomènes de production (en particulier le faux-départ et l'autocorrection).

Des approches similaires ont été proposées par [Spilker *et al.*, 2000] (modèle de langage stochastique pour un système speech-to-speech), [Siu et Ostendorf, 1996] ou encore [Stolcke et Shriberg, 1996] dans l'objectif d'améliorer les résultats en reconnaissance de la parole. Les auteurs présentent un modèle de prédiction des disfluences probabiliste et utilise le contexte droit de la disfluence (du côté de la correction) pour prédire les mots suivants. Dans ce cadre d'analyse, les auteurs traitent uniquement les cas de pauses remplies, de répétitions et d'autocorrections. Ils suggèrent de supprimer la portion disfluente de l'énoncé (respectivement la pause remplie, la partie répétée ou corrigée) une fois celle-ci identifiée à l'aide des

calculs probabilistes de leur modèle (appelé « Clean-up Model » ou « modèle de nettoyage »). Exemple :

she uh got real lucky though devient : she got real lucky though
it's a it's a fairly large community devient : it's a fairly large community

L'une des hypothèses avancées porte sur le fait que les disfluences elles-mêmes peuvent être modélisées de façon semblable au mot, chacune ayant une probabilité d'apparition conditionnée par le contexte.

Un tel modèle réduit substantiellement les doutes sur un mot dans le voisinage des phénomènes de disfluences. Toutefois, les différences majeures sont faibles et n'ont aucun impact significatif sur la précision de la reconnaissance. En effet, en confrontant un modèle trigramme de disfluences à un modèle trigramme standard pour les répétitions et autocorrections, les auteurs obtiennent un taux d'erreur de reconnaissance des mots quasi-identique (50,21% pour le modèle standard et 50,23% pour le modèle de disfluences).

Les conclusions de ces expériences montrent que la modélisation des disfluences par ce type de modèle de langage n'améliore pas de façon significative les performances d'un moteur de reconnaissance vocale. La principale raison de cet état de fait est que les disfluences sont par nature des phénomènes locaux pouvant être parfaitement modélisée par des modèles N-grammes standards. Les conclusions de ces études confirment néanmoins que les disfluences possèdent une distribution systématique et non aléatoire.

Approche à base de méta-règles syntaxiques

Le travail de [Core et Schubert, 1998], s'inscrit dans le cadre général de l'analyse robuste des dialogues au sein du groupe de dialogue de l'université de Rochester. La particularité principale de ce travail est l'introduction d'informations linguistiques (notamment la syntaxe) dans le traitement des disfluences d'une manière originale (différente de celle de SRI). Dans cette approche, le traitement se fait en deux étapes :

1. Détection des disfluences : la détection des disfluences s'effectue à l'aide d'un modèle de langage statistique (celui de [Heeman, 1997]). La principale fonction de ce module est de détecter les disfluences et de proposer une première délimitation de chacune d'entre elles.

2. Analyse syntaxique : il s'agit ici de donner une interprétation couvrant la totalité des mots de l'énoncé d'entrée. Pour cela, les disfluences détectées par le module statistique sont analysées à l'aide de méta-règles dédiées spécialement à cette tâche. La différence principale entre le traitement dans cette étape et celui de Heeman, vient du fait que le système ne prend pas en compte les relations entre les mots (comme c'est le cas dans l'approche de Heeman) mais plutôt les relations entre les structures syntaxiques qui dominent les mots.

Deux types de méta-règles sont alors utilisées pour le traitement des disfluences :

– La méta-règle de la zone d'édition (ZE) : basée sur une liste de mots qui peuvent potentiellement constituer une zone d'édition ou une partie d'elle, la méta-règle détecte tous les segments susceptibles d'être une zone d'édition.

  XP peut correspondre à n'importe quel constituant d'un énoncé dont les sous-constituants (Y) peuvent être interrompus par une zone d'édition (ZE). La méta-règle permet d'analyser l'énoncé d'entrée sans considérer la zone d'édition (*i.e*

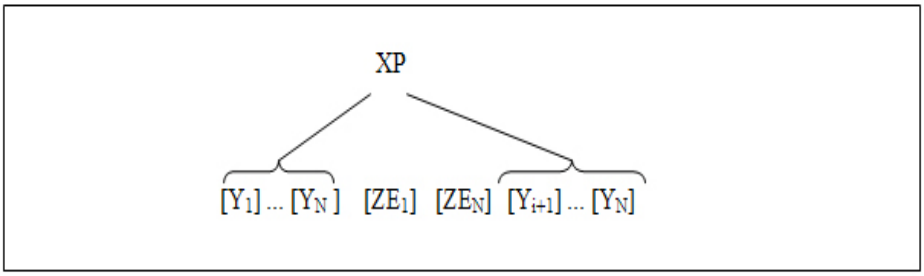


FIG. 5.1 – Méta-règle de la zone d'édition.

en ignorant cette zone).

– La méta-règle des autocorrections et faux-départs : sa fonction principale consiste à délimiter une disfluence amorcée (par le module précédent) en précisant le début et la fin des zones remplacées et remplaçantes puis, permet à l'algorithme d'ignorer la zone remplacée et de considérer uniquement la zone remplaçante.

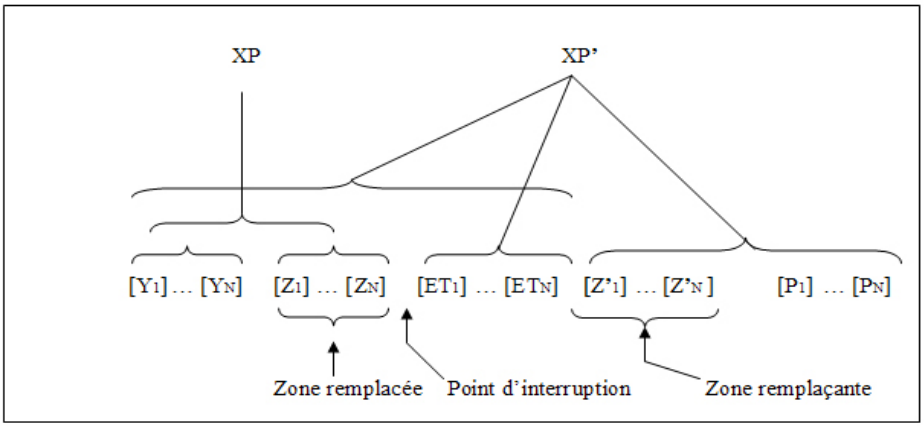


FIG. 5.2 – Méta-règle de traitement des autocorrections et faux départs.

Dans la figure ci-dessus, la nature des composantes XP et XP' n'est pas précisée, mais généralement, chaque composante est constituée d'un ensemble de syntagmes Z et Z' qui dépendent directement d'elle. Dans le cas d'autocorrection et faux-départs, les syntagmes Z et Z' tendent à être du même type. Ces méta-règles sont implantées selon le même principe que les méta-règles de la zone d'édition : les arcs qui se terminent avant la zone remplacée sont liés directement au début de la zone

remplaçante, permettant ainsi de traiter l'énoncé en ignorant la zone remplacée (ainsi que la zone d'édition qui peut la suivre).

En terme de calcul, l'ajout de ces méta-règles s'est révélé très coûteux ([Core et Schubert, 1999]). En effet, le temps de traitement d'un énoncé avec un analyseur simple est de 0.36 secondes alors qu'avec un analyseur enrichi par les méta-règles, le temps est de 0.91. En d'autres termes, l'ajout des méta-règles a multiplié par trois environ le temps de calcul. Les expériences ultérieures menées par [Core et Schubert, 1999] ont, entre autre, montré une différence importante en terme de précision avec le système de Heeman et al. (1996) : la version enrichie de méta règle affiche une précision inférieure de 40.33%.

Des approches similaires à celle de Core ont ensuite été proposées par différents chercheurs. Citons par exemple [McKelvie, 1998] qui propose également une approche à base de méta-règles, inspirée des conceptions évoquées précédemment. Outre les unités syntaxiques classiques, ses méta-règles considèrent deux catégories supplémentaires :

- Les syntagmes d'éditions (ED) qui sont les hésitations, bruits, exclamations, etc.
- Les marqueurs discursifs (AFF) qui correspondent à des mots comme *oui*, *bon*, etc. et qui, d'après l'auteur, marquent généralement le début et la fin d'un énoncé.

Exemple :

$$X \longrightarrow X, ED, AFF$$

Au sein de l'algorithme présenté par l'auteur, cette règle permet d'ignorer tous les syntagmes d'édition qui apparaissent après un constituant X.

Les études de ce type permettent généralement de réaliser une analyse syntaxique dite « robuste », mais n'attribuent aucun statut aux unités qui n'ont pas été considérées au moment du traitement (du fait de leur suppression). Bien que les

disfluences ne possèdent pas de fonction syntaxique intrinsèque, chaque segment qui occupe une place sur le plan syntagmatique remplit en lui-même la fonction syntaxique de cette place. Il semble alors délicat dans ce cas que seule une occurrence de chaque place du syntagme puisse être prise en compte dans l'analyse. Autrement dit, si l'analyse porte uniquement sur la « réparation », une interrogation reste en suspens concernant le statut des autres constituants de la disfluence (reparandum, etc.).

5.3.2 Prise en compte des disfluences avant et/ou pendant l'analyse

La seconde technique est celle qui consiste à rassembler les disfluences en un groupe. Par exemple, [Antoine *et al.*, 2003] proposent dans cette perspective des analyseurs qui forment des disfluences en rassemblant des chunks (chacun d'eux devant être une occurrence de la même place syntaxique) en vertu de « relations de dépendance sémantico-pragmatiques ». Dans un même ordre d'idées, [Godfrey *et al.*, 1992] ont mis en place une méthode d'annotation des disfluences qui consiste à réaliser un parenthésage de l'ensemble de l'accumulation paradigmatique. Le problème se pose alors pour déterminer la catégorie syntagmatique de cet ensemble qui peut être composé de diverses disfluences, et ne correspond pas systématiquement à un syntagme complet.

Feature System Parser : FEASPAR

Le travail mené par [Buoe et Waibel, 1996] traite de la conception d'un système capable d'apprendre l'analyse syntaxique de la parole en anglais : FEASPAR (*Feature Structure Parser*)

L'architecture d'un tel système consiste en un réseau de neurones et une méthode de recherche. Dans un premier temps le réseau segmente l'énoncé d'entrée en chunks qui sont étiquetés au niveau de leur fonction et des relations entretenues entre eux.

Le module de recherche s'assure ensuite de trouver la structure fonctionnelle la plus probable et cohérente.

Le système FeasPar est destiné à fonctionner dans le cadre d'un domaine précis. Le dispositif est en effet entraîné, testé et évalué sur le corpus *English Spontaneous Scheduling Task*[5] (ESST) et comparé avec les performances de l'analyseur syntaxique GLR* développé par [Lavie, 1995].

L'architecture de ce système est présentée ci-après :

[5]Application qui traite la situation d'une prise de rendez-vous entre deux interlocuteurs parlant des langues différentes.

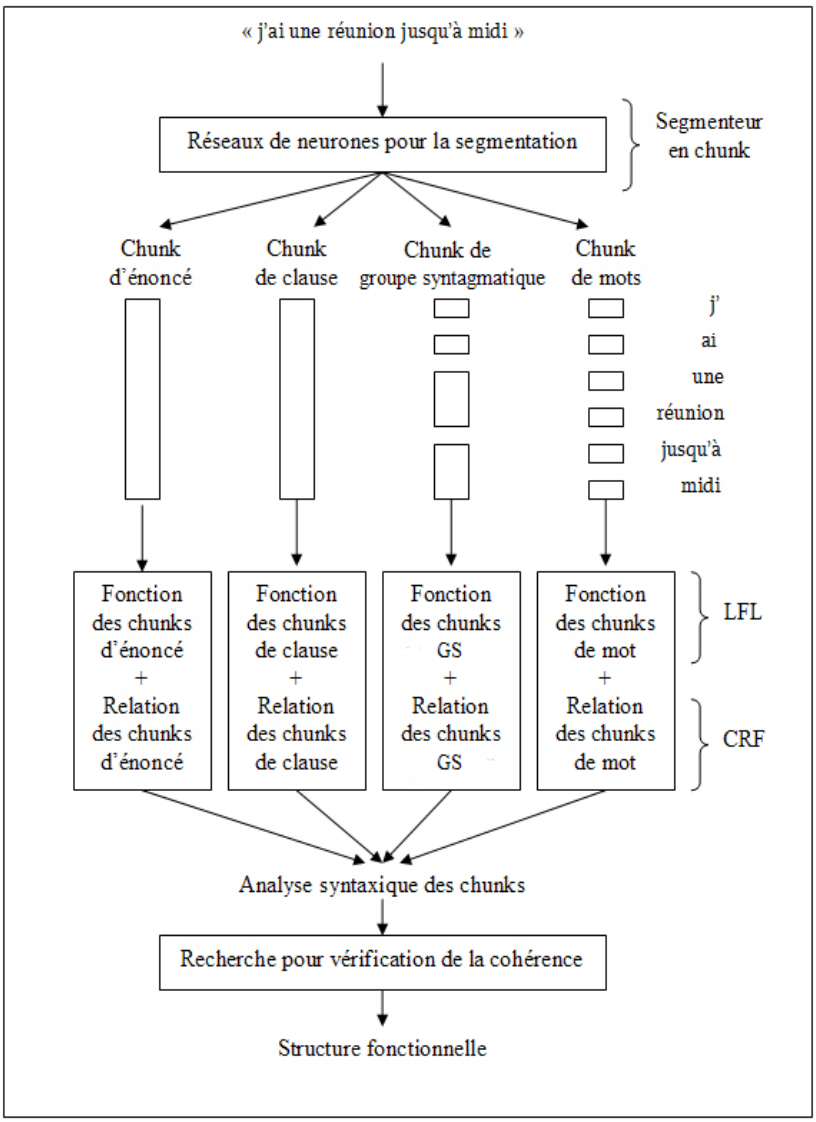


FIG. 5.3 – Architecture du système FEASPAR.

FeasPar utilise les réseaux de neurones pour apprendre à produire l'analyse syntaxique des différents chunks. Comme nous pouvons le voir dans le schéma précédent, le système est composé de trois modules principaux :

– Le segmenteur en chunks qui découpe l'entrée en chunks. Celui-ci comprend trois réseaux de neurones. Au total, il y a quatre niveaux de chunks : mot/nombres, groupe syntagmatique, clauses, énoncé. Notons que le niveau groupe syntagma-

tique est ambigu avec celui de clause et d'énoncé dans la mesure où l'auteur ne donne pas de critères de distinction entre ces différentes catégories.

- L'étiqueteur de fonctions linguistiques (LFL [*Linguistic Feature Labeler*]) attribue une fonction à chacun de ces chunks.
- L'identificateur de relations entre chunks (CRF [*Chunk Relation Finder*]) détermine comment un chunk est relié à son chunk « parent ». Il y a un réseau par niveau de chunk et par élément de relation entre ceux-ci.

Pour déterminer les capacités de FeasPar les auteurs ont comparé ses performances avec celles de l'analyseur GLR* également appliqué au corpus ESST. Les performances de FeasPar en terme de qualité d'analyse syntaxique (entre autres), pour les mêmes données, sont nettement supérieures à celles de GLR* (respectivement 71,8% de précision contre 51,6%).

Les résultats montrent un système *a priori* capable d'apprendre à traiter correctement les données orales. Néanmoins, les auteurs ne donnent aucun exemple d'énoncés soumis à l'analyseur et n'expliquent d'aucune manière ce qu'ils entendent par parole spontanée ou du moins ce qu'ils y intègrent. Il est donc difficile d'apprécier ces résultats dans la mesure où aucun moyen n'est donné pour transposer les performances de ce dispositif face à des énoncés disfluents.

Traitement des pauses remplies et des répétitions

Une grande partie des travaux consacrés aux disfluences, nous l'avons vu, relève du domaine de la reconnaissance de la parole en vue de son amélioration. Bien que les approches des divers laboratoires emploient sensiblement le même type de méthodes (approches stochastiques, modèles de Markov, réseaux de neurones artificiels, etc.), certains chercheurs proposent de nouvelles méthodologies pour le traitement des disfluences à mi-chemin entre suppression et prise en compte de ces phénomènes. C'est notamment le cas des recherches de [Stouten et Martens, 2004]

sur la reconnaissance de la parole spontanée appliquée au néerlandais, qui présentent une nouvelle méthodologie de traitement des disfluences. L'idée de base consiste à détecter les disfluences et déterminer la nature des phénomènes de production avant la reconnaissance, afin d'utiliser ces informations pour contrôler ou modifier la recherche. Cette étude vise ainsi à examiner les effets produits par les différentes techniques de traitement des disfluences pour la reconnaissance vocale.

Les auteurs considèrent le taux d'erreur de reconnaissance des mots intentionnels (*i.e.* non disfluents) comme critère d'évaluation principal de la performance du système. Dans l'état actuel de leur recherche, la méthodologie a été élaborée pour les pauses remplies et les répétitions de mots. Le principe de base de leur approche consiste à modifier la recherche sur les informations de base des disfluences qui sont extraites du flux de parole au moyen d'une interface acoustique. L'idée est de développer une interface capable de détecter les intervalles de temps qui correspondent généralement aux pauses remplies, ainsi qu'au deux parties de la répétitions (séquence initiale et séquence répétée). Ceci permettrait d'utiliser ces intervalles de temps pour modifier le comportement du moteur de recherche.

Stouten et Martens ont ainsi expérimenté différentes stratégies, allant de la non prise en compte des pauses remplies une fois détectées jusqu'à leur intégration dans un modèle de langage.

Pour ce qui est des pauses remplies, les résultats renvoyés par le système de reconnaissance (à paramétrage égal) montrent que l'adaptation du dispositif aux pauses remplies est une stratégie plus performante que celle consistant à éliminer les intervalles contenant une pause remplie (respectivement 47,19% contre 48,86%) et qu'il est pour cela préférable d'identifier les pauses remplies avant la reconnaissance

totale du signal de parole.

En revanche, le même type de méthode appliquée aux répétitions n'améliore pas (dans l'état actuel du système) de façon significative le taux d'erreur de reconnaissance.

Le système CORRECTOR

Comme le montrent les diverses approches présentées jusqu'ici, il y a globalement deux tendances qui s'opposent dans la littérature en terme de traitement automatique des disfluences. Dans le premier cas, il est possible d'utiliser les techniques dites de N-grammes et de patrons pour traiter ces phénomènes. Dans le second cas, la syntaxe est absolument nécessaire pour le traitement, et se généralise dans l'application à tous les phénomènes.

Les recherches de [Kurdi, 2003] tendent à mêler les deux tendances. En effet, il semble pour l'auteur que certains phénomènes comme les répétitions et autocorrections d'unités lexicales peuvent être traités par des approches à base de patrons et N-grammes, de façon plus simple qu'avec une grammaire « classique ».

En revanche, d'après les expérimentations de l'auteur, ces approches révèlent leur limite dès lors qu'il s'agit de prendre en compte suffisamment de contexte pour traiter certains cas. Kurdi a ensuite procédé à une modélisation syntaxique des amorces et inachèvements. Il défend également l'idée selon laquelle la prise en considération des dépendances entre les syntagmes constitue un facteur déterminant pour la détection de certains phénomènes.

Il préconise donc la combinaison entre les approches à base de patrons et celles d'analyse syntaxique pour une approche que nous pourrions qualifier d' « hybride » destinée à optimiser le coût du rapport traitement / efficacité dans l'analyse des disfluences. Ce type d'approche a été intégré dans le système CORRECTOR développé par l'auteur, qui se base à la fois sur des techniques de reconnaissance de

patrons et d'analyse syntaxique et sémantique superficielle.

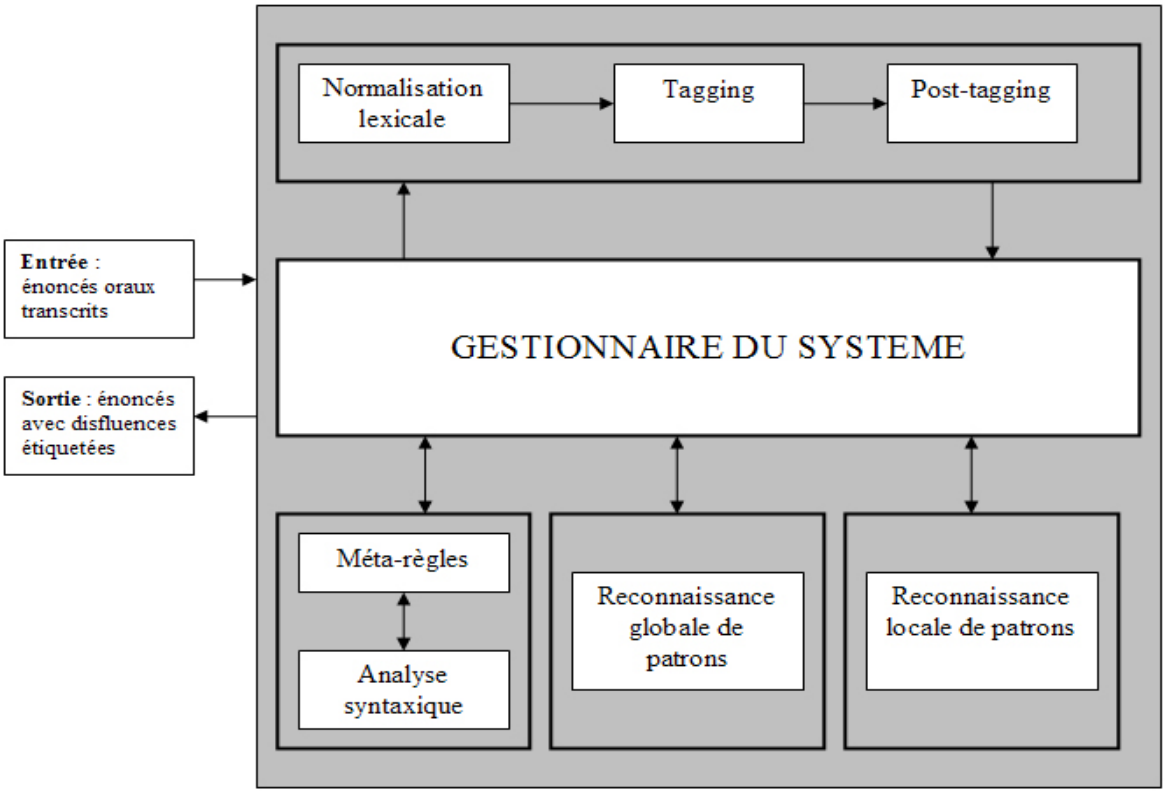


FIG. 5.4 – Architecture de CORRECTOR.

L'aspect principal de ce dispositif d'un point de vue logiciel est l'existence d'une unité centrale (le gestionnaire de système ou le hub) autour de laquelle communiquent les différents modules. Le système Corrector comprend sept modules répartis sur trois blocs qui couvrent des sources d'informations assez hétérogènes : lexique, patrons, méta-règles et règles syntaxiques. L'utilisation d'un gestionnaire de système (indépendant de ces sources d'informations) permet d'intégrer ces différentes sources d'informations au sein même du gestionnaire du système.

L'évaluation de ce système a ensuite été effectuée sur un corpus test (305 énoncés) dont les résultats – toutes disfluences confondues – sont présentés dans le tableau suivant :

Total disfluences	Détection	Rappel	89,67%
		Précision	92,76%
	Délimitation	Rappel	84,47%
		Précision	86,61%

TAB. 5.2 – Résultats du système CORRECTOR sur le corpus test anglais

Les techniques qui prennent en compte les disfluences, bien que se basant sur tout ou partie des descriptions données ci-dessus, diffèrent nettement dans leurs traitements, qui sont tous limités par la nature de leur représentation du phénomène. Toutefois, quelque soit l'approche adoptée, les résultats ne semblent pas encore suffisants pour prétendre au développement d'applications robustes sur le traitement de l'oral.

5.4 Conclusion

Les techniques qui prennent en compte les disfluences, bien que se basant sur tout ou partie des descriptions données ci-dessus, diffèrent nettement dans leurs traitements, qui sont tous limités par la nature de leur représentation du phénomène. Toutefois, quelque soit l'approche adoptée, les résultats ne semblent pas encore suffisants pour prétendre au développement d'applications robustes sur le traitement de l'oral.

Dans notre approche, nous allons tenter de voir comment peuvent être analysés de façon automatique les énoncés oraux disfluents, ainsi que les différents niveaux de difficulté qu'ils peuvent comporter. Nous travaillerons dans un premier temps sur un échantillon de nos données en ciblant notre analyse à l'aide d'une technique préalablement décrite, puis nous procèderons dans un second temps à une analyse linguistique détaillée des phénomènes de production sur corpus.

Chapitre 6

Cadre d'analyse et représentation choisis

6.1 Introduction

Dans le chapitre précédent, nous avons mis en avant la variété des modèles de langage ou cadres d'analyse existants pour l'étude des disfluences. Nombre de ces approches sont dérivées du modèle développé par Levelt et relèvent plus de la psycholinguistique et des sciences cognitives ; on s'interesse dans ce cas à comprendre la manière dont le locuteur prend conscience de sa production. De plus, lorsque le modèle proposé tient compte de la production des disfluences (par exemple [Shriberg, 1994]), l'analyse se place sur un plan purement syntagmatique, proposant ainsi une vision linéaire du fonctionnement des disfluences.

Notre approche ne se situe pas dans la problématique de la question du traitement des disfluences à l'aide d'un module générique (cf. [Shriberg, 1994]) mais plutôt d'un module spécifique pour chaque disfluence (cf. [Kurdi, 2003] par exemple), ainsi que sur la question de leur structure fonctionnelle (essentiellement paradigmatique dans notre cas). Nous pensons de plus qu'un cadre d'analyse unifié peut tout à fait convenir pour analyser ces phénomènes d'un point de vue théorique et avoir

ensuite recours à diverses stratégies adaptées à chaque cas.

Par ailleurs, lorsqu'il est question d'analyser ces phénomènes en termes de traitement automatique, la majorité des approches proposées tendent à vouloir les supprimer avant tout traitement ultérieur. Pourtant, les disfluences peuvent également s'organiser différemment, dans un cadre d'analyse qui ne nécessite pas de les supprimer. Il suffit pour s'en convaincre de les représenter schématiquement pour voir qu'elles peuvent parfaitement « se greffer » sur une représentation arborescente et répondre ainsi à une organisation syntaxique au même titre que d'autres éléments grammaticaux. C'est ce que nous montrerons plus loin en nous appuyant sur les théories de [Blanche-Benveniste, 1990]. Nous reprendrons notamment la notion de « mise en grille syntaxique » qui nous permet d'examiner plus particulièrement le fonctionnement des disfluences sur l'axe paradigmatique, et à partir duquel nous dérivons une représentation syntaxique de ces phénomènes (dite en « arbres marcottés », rappelant les structures en arbres de dépendance).

Notre étude s'insère donc dans le cadre d'un travail d'équipe sur la syntaxe du français parlé. En effet le DELIC a lancé un programme systématique d'étude des disfluences à l'aide des grands corpus informatisés dont elle dispose : étude des répétitions, des hésitations, des amorces, etc. Ces observations ont permis de dégager des régularités qui commencent à émerger : on observe que les répétitions frappent principalement les mots grammaticaux introducteurs de syntagmes (prépositions, articles), et que des interactions très fortes lient les différents types de disfluences, qui permettent d'envisager la mise au point de modèles prédictifs pour leur identification automatique.

6.2 La mise en grille

6.2.1 Principe

Une fois les enregistrements de français parlé transcrits, linéairement et sans ponctuation, une présentation reste à trouver pour les rendre lisibles. En effet, les divers phénomènes de disfluences sont pénibles à lire lorsqu'ils sont disposés de façon linéaire. Leur présence contrevient à nos habitudes de lecture.

[Blanche-Benveniste et Jeanjean, 1987] expliquent que le manque de lisibilité des transcriptions est en partie dû à l'appauvrissement que subit le texte oral quand on le prive des caractéristiques de prononciation, des intonations, et des gestes qui l'accompagnent.

Aussi, un modèle de représentation du phénomène, bien connu des syntacticiens mais encore peu exploité dans le domaine du TAL[1], est la représentation par « mise en grille syntaxique » développée par [Blanche-Benveniste *et al.*, 1979]. L'auteur propose pour la première fois de traiter les autocorrections (et plus globalement les phénomènes de production) avec des moyens syntaxiques canoniques en les rapprochant du mécanisme de la coordination : les phénomènes disfluents sont ramenés à une suite de « piétinements » sur une même place syntaxique.

En effet, le phénomène d'enrichissement lexical, très fréquent, est un exemple de caractéristique syntaxique fondamentale du français parlé. On assiste à l'instanciation de plusieurs syntagmes (de longueur variable) à la même position ou « place » syntaxique (il s'agit de la position syntaxique, c'est-à-dire la place dans l'ordre syntagmatique, et non pas de la fonction syntaxique).

C. Blanche-Benveniste et C. Jeanjean (1987) définissent la méthode de la mise en grille de la façon suivante :

[1] Voir cependant [Luzzati, 2004].

> " *Chaque ligne représente une suite concrète de mots telle qu'elle a été énoncée. Le mode de lecture consiste à lire chaque ligne l'une après l'autre, quelle que soit son étendue. Les colonnes verticales qui apparaissent indiquent le remplissage syntaxique d'un même emplacement syntaxique.* " (p.167)

Cette méthode a donc pour fonction de visualiser par écrit l'organisation du discours oral. Les éléments en rapport de dépendance sont placés sur l'axe syntagmatique. Dès que le rapport de dépendance est brisé, les constructions verbales occupent une ligne entière. Ainsi, les éléments qui occupent une même place syntaxique dans l'énoncé, sont alignés sur l'axe paradigmatique. Ce modèle permet donc de visualiser les configurations du discours grâce à une représentation qui suit deux axes simultanément : l'axe syntagmatique (ou horizontal) et l'axe paradigmatique (ou vertical).

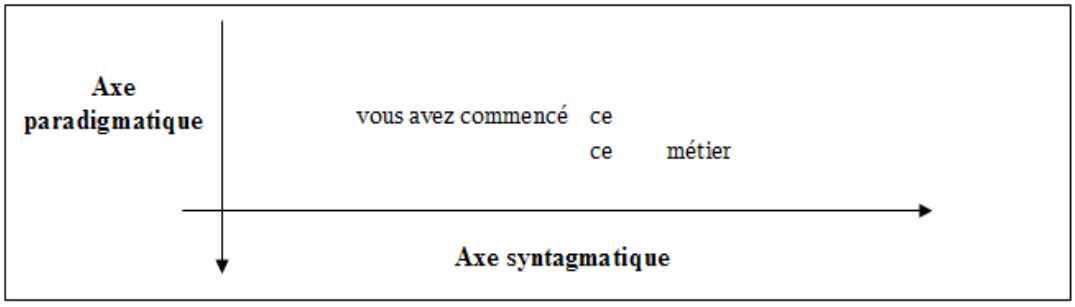


FIG. 6.1 – Représentation sur l'axe syntagmatique et paradigmatique.

L'axe syntagmatique est celui des unités qui sont à analyser dans leur successivité. Rappelons qu'au contraire sur l'axe paradigmatique s'établissent des rapports associatifs sur la base de relations diverses.

Ainsi, la rupture sur l'axe syntagmatique permet de poser le cadre syntaxique engagé par le locuteur (syntagme nominal, prépositionnel, etc.) puis l'axe paradigmatique rend compte du remplissage lexical. Voici quelques exemples :

je voulais voyager hein: toujours	**ce**
	cette envie de voyager /

FIG. 6.2 – Mise en grille d'une autocorrection.

il y a euh des fiches euh + euh **sur**	euh
sur	**la:**
	la
	la
	la
	la botanique bon \ il y a une mul*titude / + de sujets qui sont abordés hein /

FIG. 6.3 – Mise en grille de répétitions successives.

Le remplissage des emplacements syntaxiques s'agence alors en « listes de mots » aux valeurs multiples : mots, fragments de mots (amorces), syntagmes entiers. Ce cadre d'analyse permet donc d'appréhender aussi des phénomènes qui ne sont pas disfluents tels que les énumérations, les précisions, ou encore les répétitions « faits de langue » :

j'ai amené	**des affaires d'hiver** /
	des affaires euh d'été / +

FIG. 6.4 – Mise en grille d'une énumération.

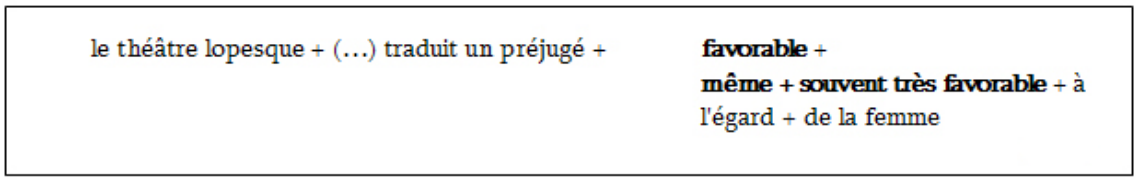


FIG. 6.5 – Mise en grille d'une répétition avec précision.

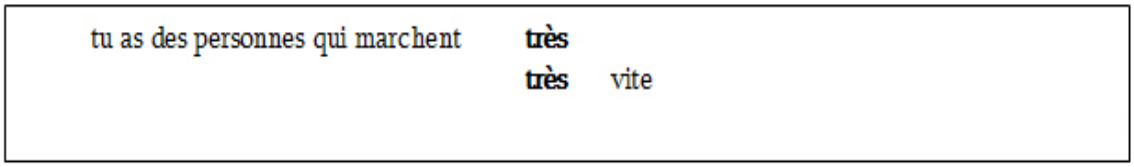


FIG. 6.6 – Mise en grille d'une répétition « faits de langue ».

Ce cadre d'analyse n'est pas conçu comme une analyse de discours, c'est la représentation graphique d'une analyse grammaticale qui permet d'intégrer les disfluences. A l'aide d'une telle représentation, les alignements paradigmatiques ressortent convenablement, et permettent de repérer où sont concentrées les disfluences dans l'énoncé.

6.2.2 Intérêts

L'analyse résultante de la transcription d'un corpus doit tenir compte de nombreux paramètres : les phénomènes de production de l'oral, les interférences éventuelles entre les locuteurs (chevauchements, etc). L'analyse ne peut pas se limiter à de courts passages analogues à ce que l'on serait tenté de qualifier (à tort) de « phrases ».

Le rôle du modèle de Blanche-Benveniste est de pouvoir s'appliquer à des passages assez étendus, où se dessine la structure grammaticale mise en oeuvre par les locuteurs. Il est ainsi question de rendre compte du fonctionnement de la production du discours par le locuteur. Aucune distinction n'est faite entre les différentes

configurations d'autocorrection et ne propose pas de classement en fonction de leur portée (mots outils, syntagmes, etc.).

Lorsqu'un locuteur s'y reprend à plusieurs fois pour produire un syntagme verbal du type sujet + verbe, en recommençant à partir du sujet :

je lui ai dit oui **je je je** *vais y penser*

il serait méthodologiquement incorrect de dire qu'il produit une suite syntagmatique faite de ces trois morceaux placés à la suite les uns des autres, comme s'ils devaient s'enchaîner. L'ensemble des trois reprises occupe le même emplacement syntagmatique, et sont donc listées à la verticale :

je lui ai dit oui **je**
je
je *vais y penser*

A l'inverse de l'écrit, l'oral ne filtre pas les productions disfluentes émises par le locuteur et celles-ci sont parfaitement prises en charge dans les grilles. La mise en grille syntaxique induit en effet une analyse structurelle et fonctionnelle des différents phénomènes de disfluence propres à l'oral, qu'une simple écoute ou une transcription linéaire ne mettrait pas aussi bien en évidence.

Pourtant, l'une des approches les plus répandues concernant la modélisation des disfluences est celle proposée par [Shriberg, 1994] (cf. 4.2.2). Rappelons que l'auteur décrit l'organisation interne des disfluences de façon linéaire en un ensemble d'espaces distincts délimitant les étapes de la production orale : le *reparandum* (RM), le *point d'interruption* (PI), l'*interregnum* (IM) et le *repair* (RR).

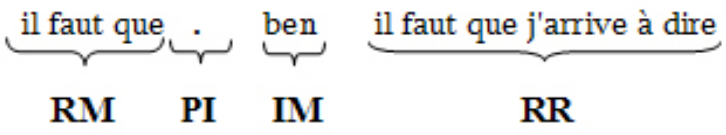


FIG. 6.7 – Structure [simplifiée] de la disfluence ([Shriberg, 1994]).

Cependant, ce modèle **RM/PI/IM/RR** révèle un certain nombre de limites. L'une d'elles concerne la non-récursivité de ce modèle (*i.e* l'impossibilité d'avoir un schéma **RM/PI/IM/RR** à l'intérieur d'un premier schéma **RM/PI/IM/RR**) empêche de rendre compte de certaines configurations syntaxiques telle que l'imbrication de disfluences.

Il est très fréquent d'observer de telles imbrications : une disfluence s'insère dans une autre avant que la première soit terminée créant ainsi une interdépendance entre les segments disfluents. Il s'agit en fait de plusieurs éléments sur l'axe paradigmatique qui se succèdent et qui se trouvent ainsi imbriqués les uns dans les autres. L'imbrication s'effectue au niveau de la syntaxe où l'on observe les unités syntaxiques fondées sur l'organisation des catégories grammaticales et de leur rection. Or, le formalisme décrit dans le modèle de [Shriberg, 1994] ne permet pas de rendre compte de la configuration suivante (où nous tentons de représenter, de façon erronée, les *reparandums* et *repairs*) :

> RM[*on va*] RR[*on va essayer de gérer un stock chaque semaine à peu près pareil*] RM[*pour avoir*] RR[*pour faire* RM[*des compositions*] RR[*des bouquets ronds*]] RR[*ou pour faire un peu de tout en fait*]

Il suffit pour s'en convaincre d'observer les travaux de [Piu, 2006] qui a mis en place un schéma d'annotation des disfluences dans deux corpus oraux. Les résultats montrent que ce modèle s'applique parfaitement à l'analyse des phénomènes de production pour lesquels un cadre d'analyse tel que celui de [Shriberg, 1994] s'avère

inadéquat (cf. [Bove et Piu, 2007]).

En effet, la solution adoptée dans cette étude privilégiant la technique de mise en grille pour l'annotation, permet de laisser le choix dans un processus d'analyse syntaxique ultérieur : gommer ou adapter les disfluences, ou les deux (supprimer les disfluences, analyser les énoncés, puis les restituer sur la représentation syntaxique pour garder l'information syntaxique du début de la disfluence).

Le fait de préserver l'emplacement du début de la disfluence permet d'indiquer ainsi les catégories syntaxiques les plus touchées par les disfluences. Cette information pourrait permettre entre autre de renseigner les moteurs de reconnaissance vocale pour détecter les portions disfluentes dans le signal de parole converti en texte, ainsi que dans l'analyse syntaxique du texte ainsi reconnu où des règles peuvent être implémentées pour prendre en compte les cas disfluents.

A l'aide de la mise en grille l'exemple précédent n'est plus problématique et peut être représenté comme suit :

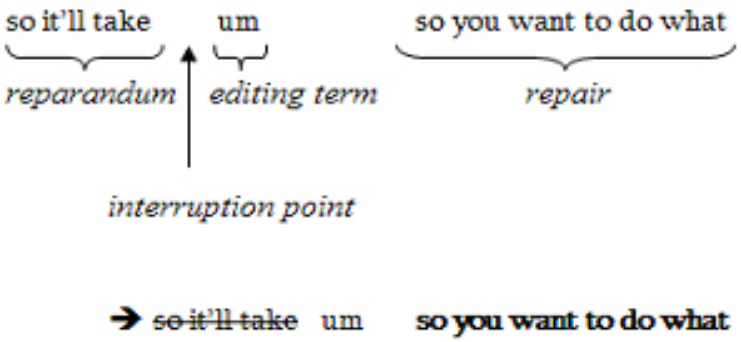
on va
on va *essayer de gérer un stock (...)* **pour** *avoir*
pour *faire* **des** *compositions*
des *bouquets ronds*
ou pour *faire un peu de tout en fait*

Dans l'exemple ci-dessus, l'entassement paradigmatique porte, entre autres, sur la « place syntaxique » de complément d'objet, qui peut être de différents types : disfluence (*pour avoir pour faire*), énumération (*des compositions des bouquets ronds*), coordination (*ou pour faire* [...]), etc. qu'il est parfois difficile de distinguer.

De la même manière, le modèle **RM/PI/IM/RR** permet dans de nombreuses

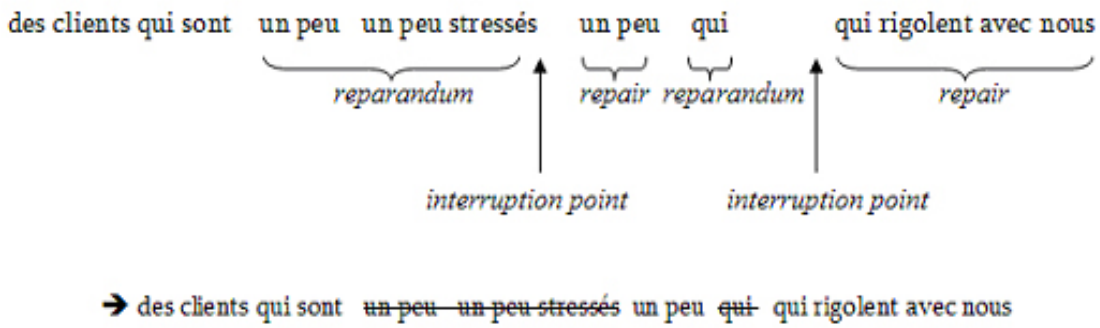
études (par exemple : [Bear *et al.*, 1992] ; [Heeman *et al.*, 1996] ; [Jorgensen, 2007]) de repérer le point d'interruption (PI) de façon à supprimer la partie gauche de celui-ci. En effet, dans cette stratégie la partie gauche du PI est considérée comme la partie de la disfluence à supprimer de façon à obtenir un énoncé proche de l'écrit (*i.e* correctement formé).

Par exemple :



En examinant notre corpus, nous relevons cependant des cas pour lesquels cette méthode s'avère inadéquate et génèrerait une analyse erronée des différentes parties du phénomène de production.

a)



b)

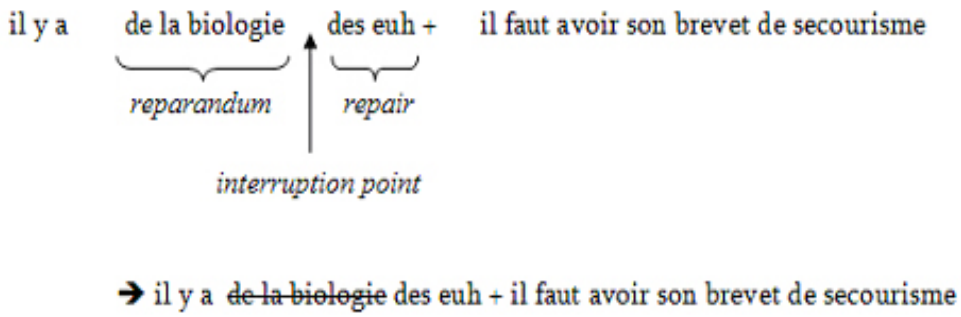


FIG. 6.8 – Exemples d'analyses erronées selon le modèle RM/PI/IM/RR.

La mise en grille permet de voir que la partie de la disfluence à conserver n'est pas obligatoirement la dernière produite par le locuteur :

a)

des clients **qui** *sont* **un peu**
un peu *stressés*
un peu
qui
qui *rigolent avec nous*

b)

il y a **de** *la biologie*
des *euh + il faut avoir son brevet de secourisme*

Il est alors nécessaire dans ce type de cas d'employer un autre mécanisme permettant de déterminer la partie « valide » de la disfluence.

Soulignons que ce type de représentation graphique ne va pas sans quelques inconvénients pratiques. La réalisation des grilles demande un temps prohibitif à partir du moment où le nombre de grilles à représenter est élevé. En effet, il n'existe pas encore (à notre connaissance) d'outil permettant d'automatiser cette tâche, et nous sommes donc contraints de réaliser les grilles manuellement. Il conviendra, dans le cadre de recherches ultérieures, de dégager des spécifications supplémentaires pour

faciliter la conception et la réalisation d'outils permettant l'automatisation de la mise en grille.

6.3 Conclusion

Nous avons choisi de nous baser sur le modèle de la mise en grille syntaxique dans notre perspective d'analyse linguistique. Dès lors, sans renier les apports théoriques des divers modèles présentés dans la partie précédente, nous suivons cette représentation dans la mesure où elle constitue une architecture qui suit un cadre d'analyse unifié et permet de traiter les disfluences avec une certaine neutralité (on ne gomme pas le phénomène disfluent). En effet, dans cette représentation, tous les essais de lexique sont conservés même s'ils ne font pas avancer le discours. De plus, ce type de représentation conforte l'idée selon laquelle les disfluences peuvent fonctionner de façon simultanée.

S'il a indéniablement été prouvé dans la littérature (et comme nous le verrons plus loin) que les disfluences sont soumises à certaines contraintes syntaxiques, il n'est pas pour autant clairement établi que ces phénomènes puissent être uniquement décrits par un modèle syntaxique simple et unique. Nous souhaitons formuler dans ce qui suit une représentation syntaxique précise des séquences de disfluences afin de rendre compte de leur structure fonctionnelle.

Chapitre 7

Arbres marcottés et aspects quantitatifs

7.1 Introduction

Le nombre d'entorses à la « norme » présents dans l'énoncé ci-dessous (cf. 2) pourrait laisser penser qu'une analyse linguistique détaillée du français parlé est difficilement envisageable :

j'ai pris enfin j'ai j'ai je me s- je je leur ai dit c- enfin je leur ai dit

Nous allons pourtant essayer dans ce chapitre de donner une description du formalisme adopté dans notre étude pour la représentation des disfluences (les arbres marcottés) et à son application (à travers une analyse empirique) sur ces mêmes données, en dressant une typologie des arbres ainsi créés.

Après avoir présenté la vision théorique adoptée sur le principe de fonctionnement des disfluences par le biais de la représentation en « arbres marcottés », nous procèderons à l'analyse linguistique détaillée des disfluences relevées dans le corpus précédemment décrit. Il s'agira d'examiner les patrons morpho-syntaxiques de chacun des phénomènes de production afin d'observer les fonctionnements structurels de celles-ci : quantification des phénomènes, principales catégories morpho-syntaxiques touchées, etc.

7.2 Représentation syntaxique en « arbres marcottés »

Le modèle de mise en grille est directement lié à une représentation initialement proposée par [Véronis, 2004] et inspirée du phénomène de liste (cf. 1.4.1) et permettant (par extension) de représenter les disfluences. L'idée consiste à décrire syntaxiquement les exemples d'énoncés disfluents à l'aide d'une visualisation « en trois dimensions ».

Nous introduisons cette représentation formelle des disfluences, à la suite de quoi nous proposerons quelques mécanismes de parsing permettant de les traiter (cf. III).

Rappelons que les listes correspondent à un ensemble de constituants qui occupent la même place syntaxique dans une unité maximale. Les listes peuvent s'imbriquer à plusieurs niveaux, de façon tout à fait surprenante. La figure 7.1 montre que le locuteur utilise à deux reprises une imbrication de liste à cinq et six niveaux, tout en incluant une parenthétique, pour un énoncé contenant en tout huit piétinements syntaxiques.

Nous avons dans nos corpus des exemples d'imbrications encore plus complexes, et il est frappant de constater que les locuteurs perdent rarement le fil de ce type d'énoncés et contrôlent même parfaitement l'organisation de ces imbrications. Ce type d'exemple met clairement en évidence la complexité structurelle fréquente des disfluences de notre corpus.

Ces listes développent les arbres dans une troisième dimension par un processus de « marcottage » qui engendre de nouveaux arbres à partir de certains noeuds.

La barre double verticale correspond aux éléments placés sous la même place syntaxique d'un point de vue paradigmatique. C'est précisément sur cet axe que se

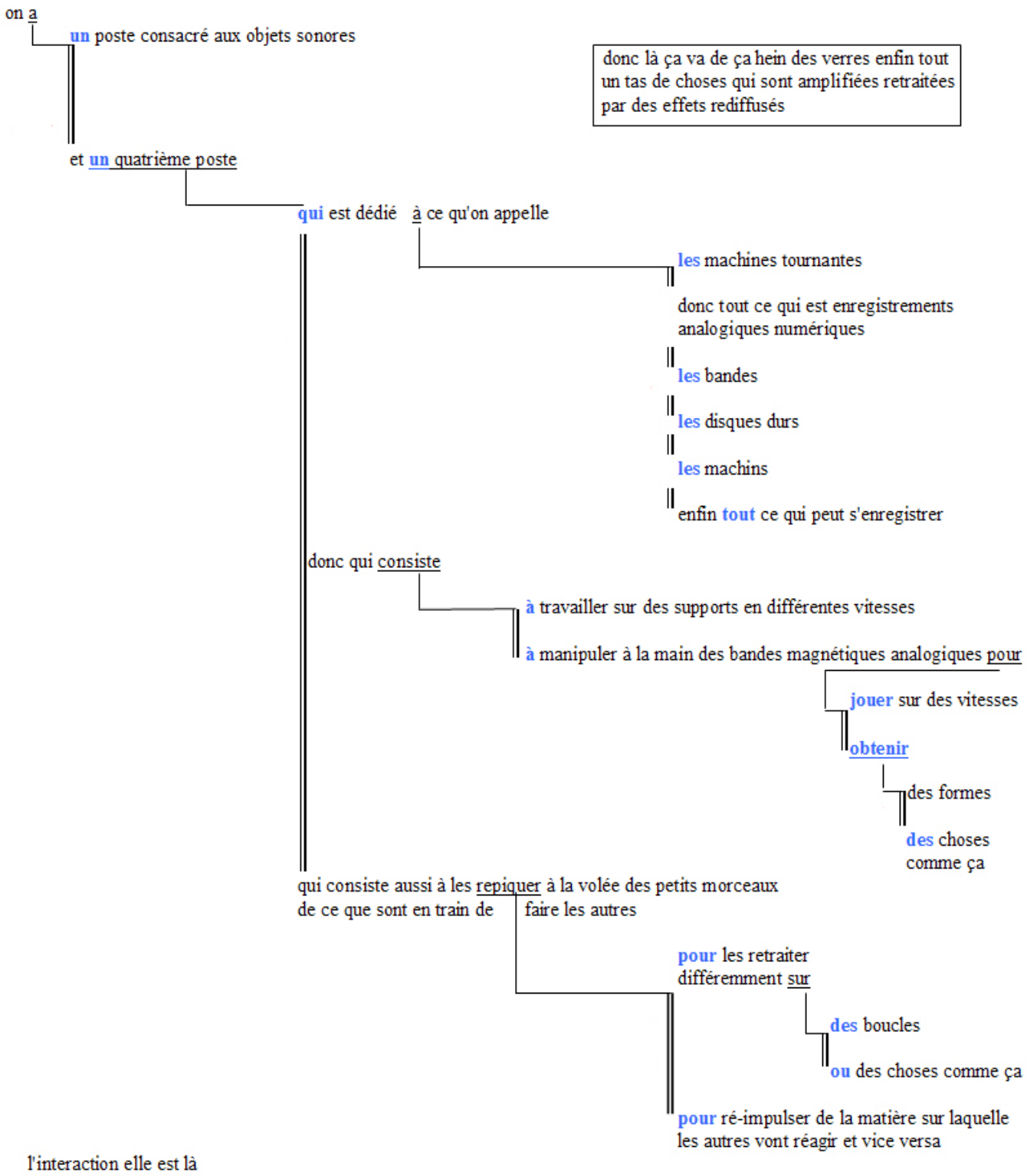
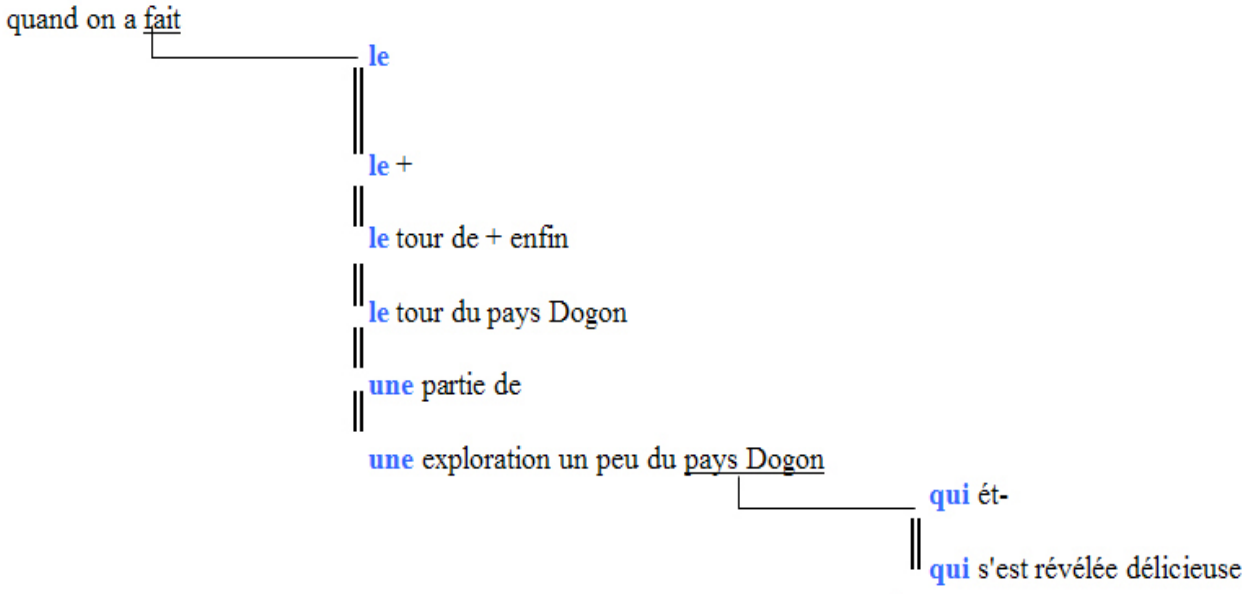


FIG. 7.1 – Exemple d'arbre marcotté.

dessinent à la fois les disfluences mais aussi les constructions juxtaposées, les énumérations ou encore les coordinations comme nous le verrons plus loin. A l'inverse, la barre simple horizontale correspond à la progression du discours sur l'axe syntagmatique.



La figure précédente, par exemple, représente des cas de répétition (*le le le tour de* [...]), autocorrection (*le tour de + enfin le tour du pays Dogon une partie de une exploration* [...]) et amorce (*qui ét- qui s'est* [...]).

Cette technique est qualifiée d' « arbres marcottés » dans la mesure où la schématisation des phénomènes rend compte de structures arborescentes sur lesquelles viennent se greffer les disfluences. C'est précisément cette représentation formelle que nous avons choisi afin d'exposer notre vision théorique et notre conception grammaticale du fonctionnement des disfluences. Cette schématisation rappelle, jusqu'à un certain point, la représentation à la manière des arbres de dépendance (voir [Tesniere, 1959]). Lorsqu'il y a production de disfluence, on assiste au développement des feuilles de l'arbre de dépendances représentées sur un plan différent

de l'axe syntagmatique ou paradigmatique.

Exemple :

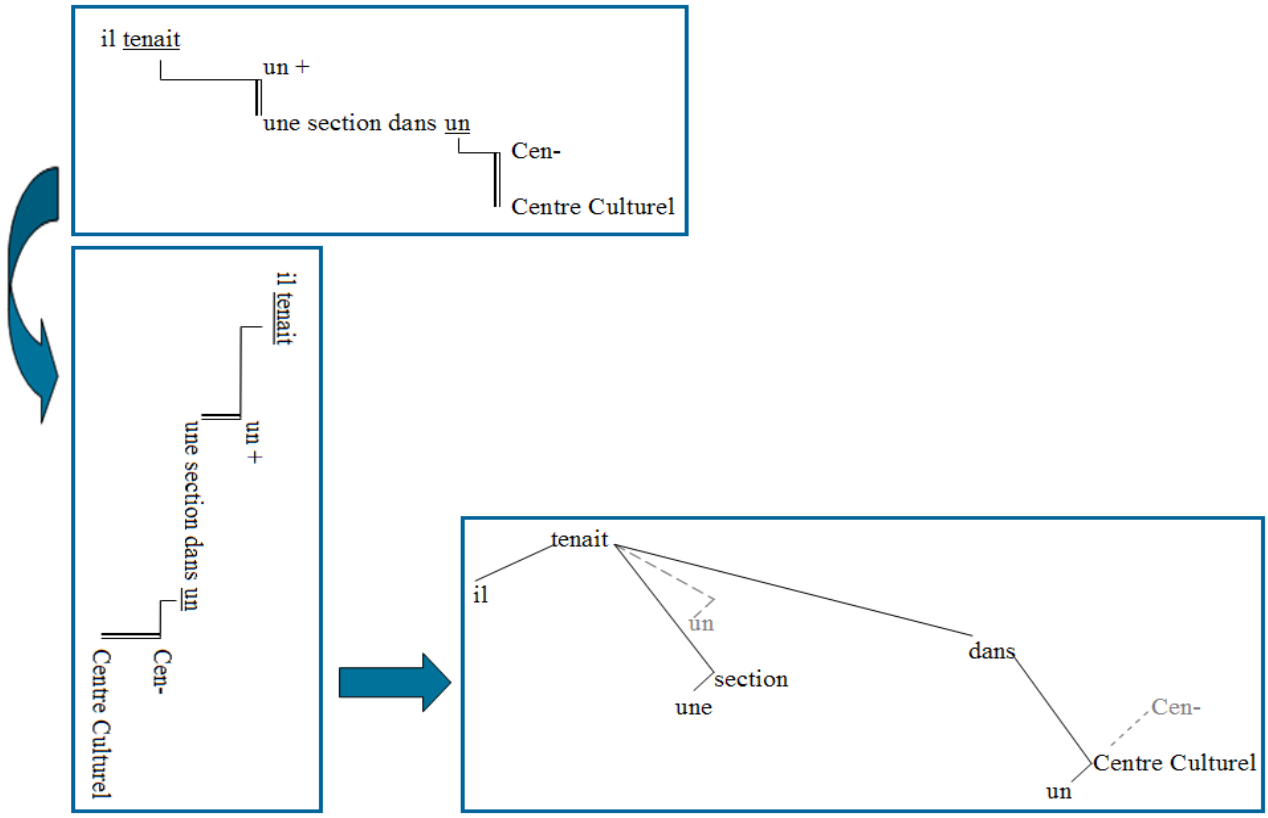


FIG. 7.2 – Passage d'un arbre marcotté à une structure en dépendance.

Dans la section suivante nous proposons entre autre une typologie non exhaustive d'arbres marcottés afin d'étudier et de rendre compte des niveaux syntaxiques sur lesquels se produisent les disfluences.

7.3 Typologie d'arbres et analyse quantitative

7.3.1 Typologie d'arbres marcottés

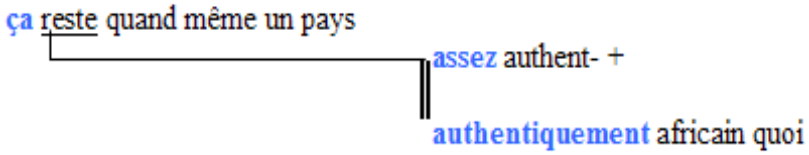
A partir de notre corpus de travail (sous-ensemble du CRFP) et du CRFP dans son intégralité, nous avons extrait plus d'une centaine d'énoncés (soit plus de 9000 mots) que nous avons représenté selon notre formalisme, ainsi que chacun des piétinements syntaxiques qui les composent. Ces énoncés ont été méticuleusement choisis pour leur caractère monologique, ainsi que pour la diversité et la richesse des disfluences qu'ils comportent. Nous proposons ici de les classer selon divers critères (type de disfluence, phénomène d'imbrications, etc.) pour lesquels nous donnerons systématiquement des exemples extraits du corpus.

Classement par phénomènes

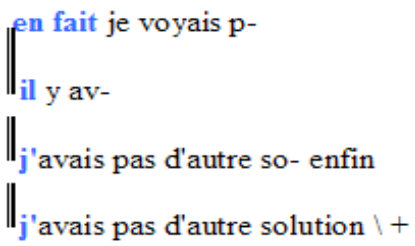
Le processus de « marcottage » permet de rendre compte de la majorité des configurations de disfluences. Il suffit pour s'en convaincre, de reprendre, en la simplifiant quelque peu, la typologie que nous avions présentée dans un chapitre précédent (cf. 2) illustrée d'exemples d'arbres.

Amorces :

Exemple :

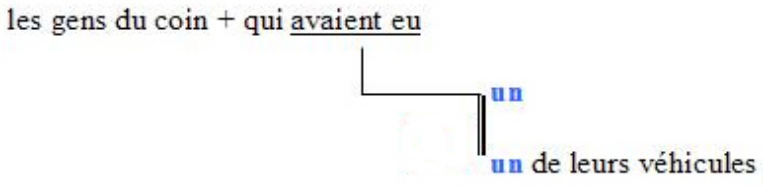


Exemple :

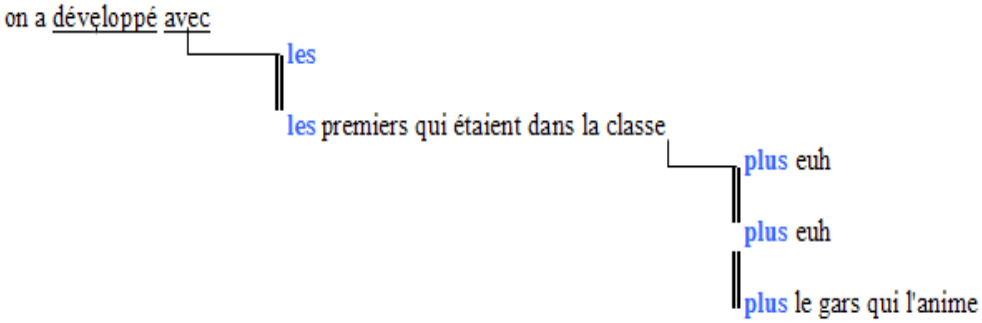


Répétitions :

Exemple :

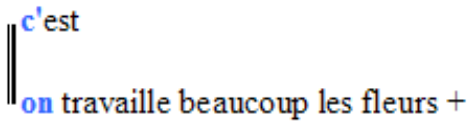


Exemple :

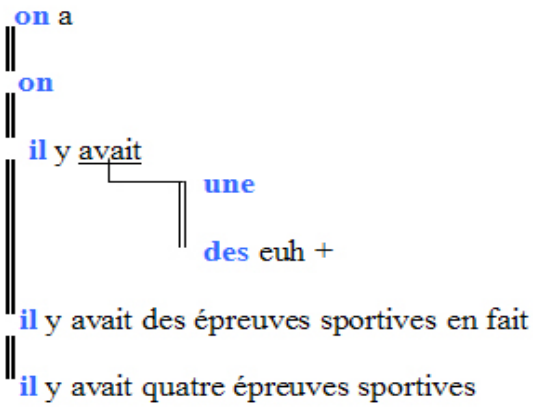


Autocorrections :

Exemple :

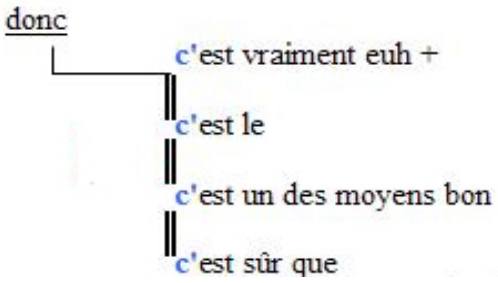


Exemple :

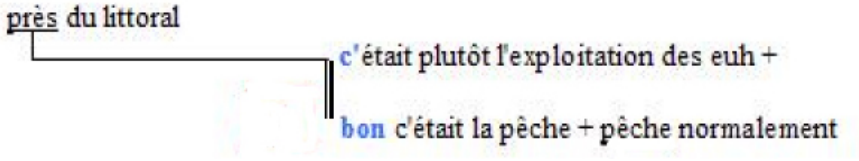


Inachèvements :

Exemple :



Exemple :



Quelques rares énoncés disfluents induisent néanmoins une ambiguïté quant au rattachement entre éléments recteurs et éléments régis. C'est par exemple le cas lorsqu'on observe une inversion de type complément / sujet (*quand je suis revenue / elles voyaient ça vraiment* [...]).

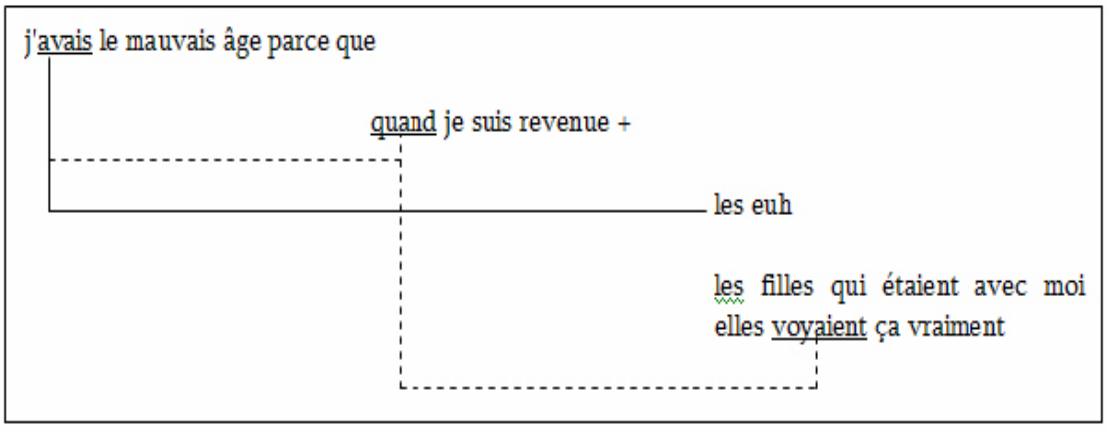


FIG. 7.3 – Cas d'ambiguïté recteur/régis.

Disfluences imbriquées et disfluences combinées

Comme nous l'avons évoqué, la représentation des imbrications de disfluences par mise en grille ne constitue pas *a priori* une difficulté.

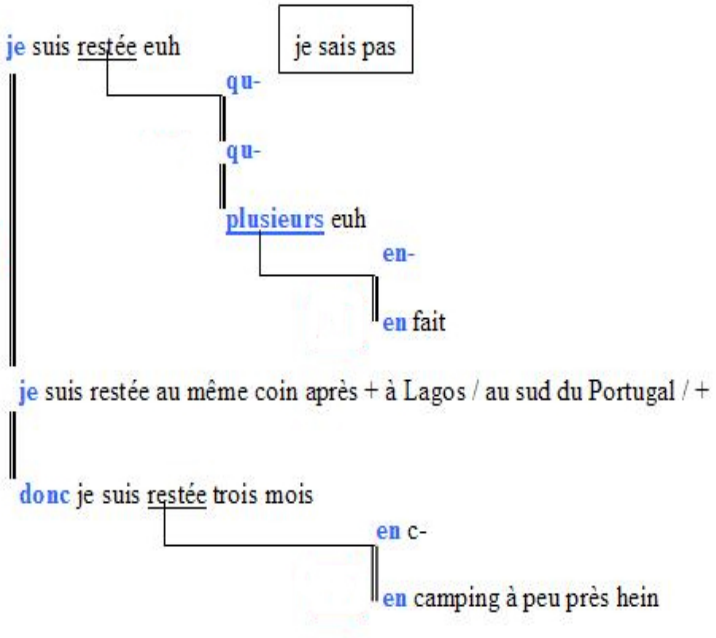


FIG. 7.4 – Exemple de disfluence imbriquée.

On observe ici une imbrication multiple où l'amorce *en- en fait* est imbriquée dans la répétition d'amorce *qu- qu- plusieurs euh* déjà imbriquée dans la répétition *je suis restée* (...) *je suis restée au même coin après + à Lagos*.

Comme le montre cet exemple, un énoncé ne contient pas toujours un seul type de disfluence. Plus de 16% des cas de notre corpus correspondent en fait à des disfluences combinées (cf. 2), dans lesquelles viennent éventuellement se greffer en plus des parenthétiques (segment encadré).

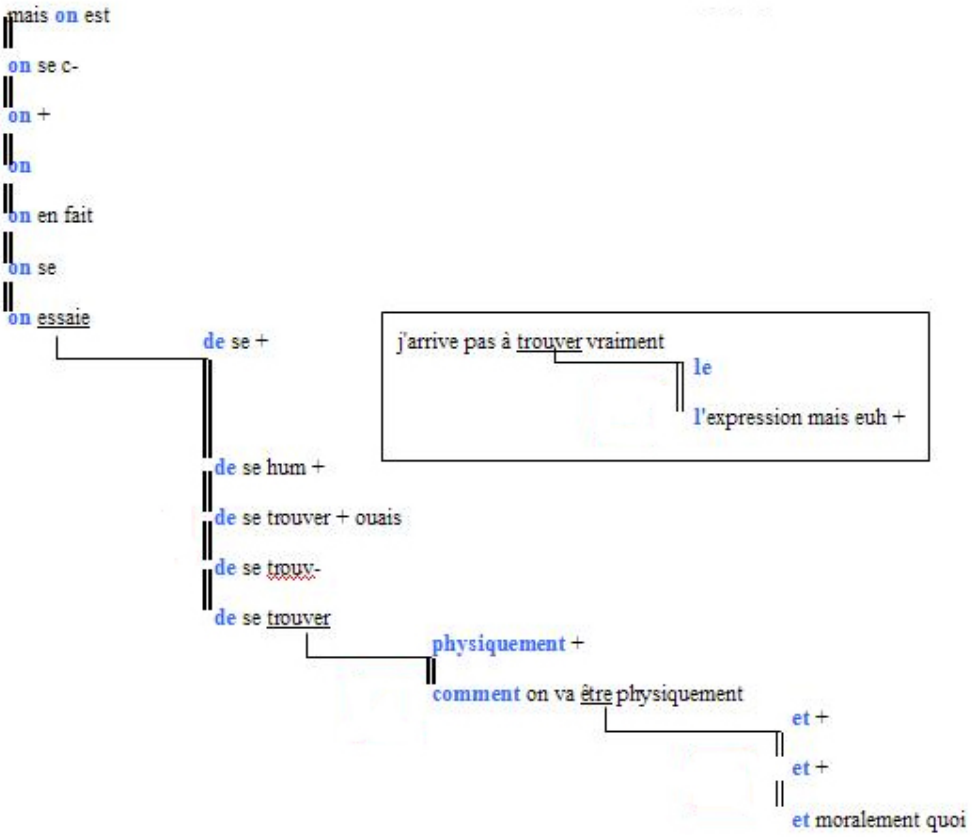


FIG. 7.5 – Exemple de disfluences combinées avec parenthétique.

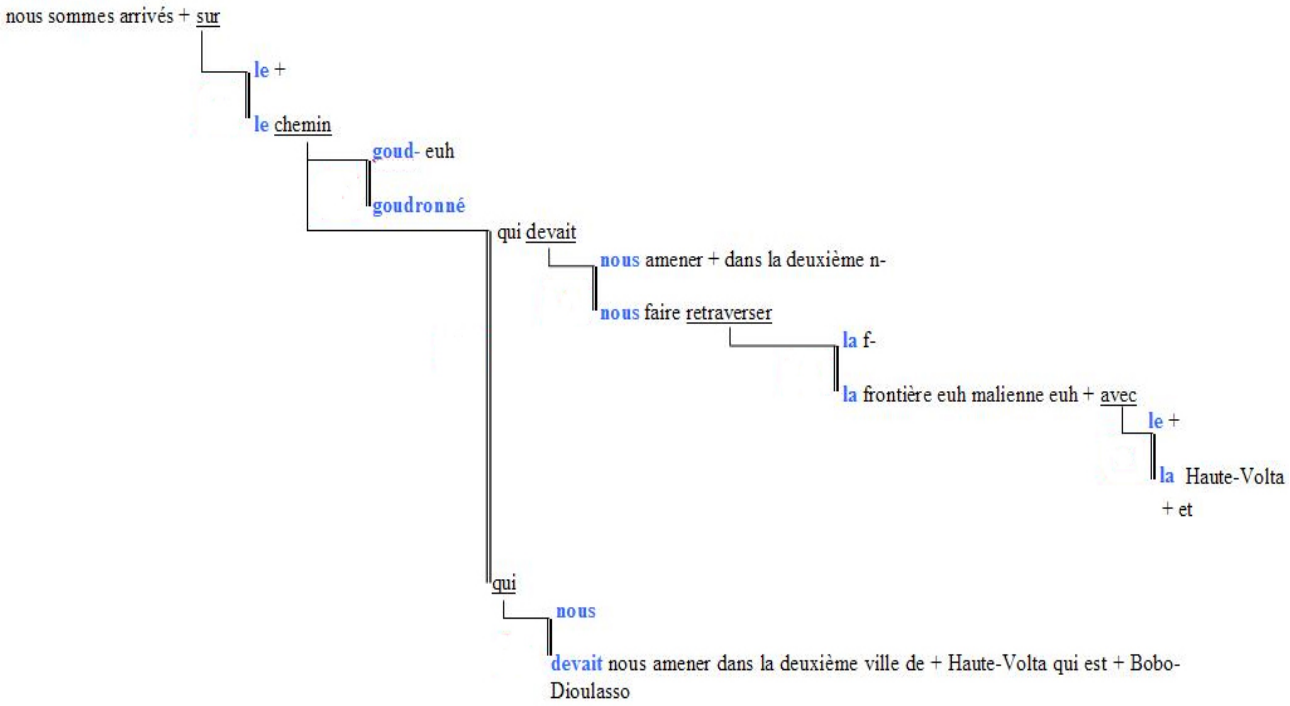
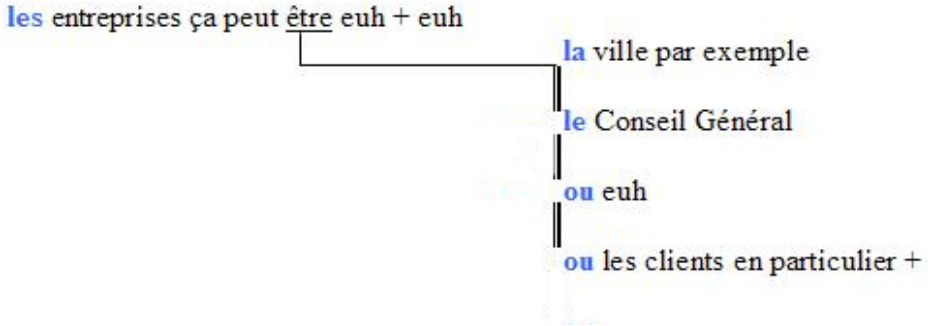


FIG. 7.6 – Exemple de disfluences combinées.

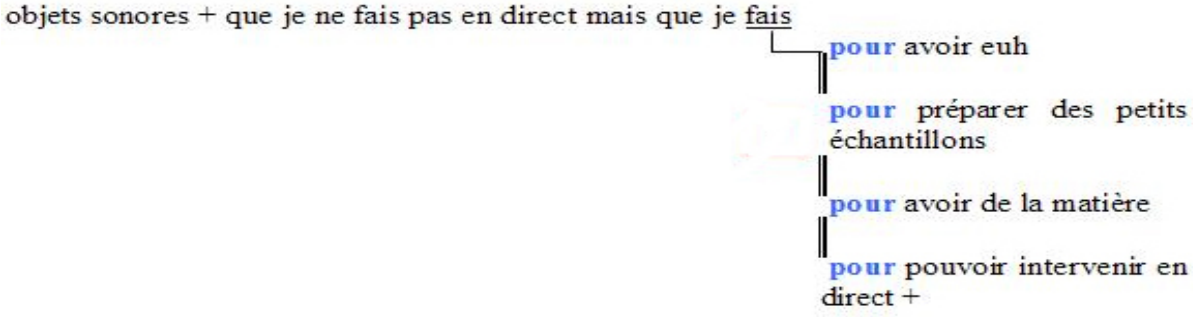
En plus de la combinaison de disfluences, on relève également de nombreux cas d'apparitions concomitantes de disfluences et d'énumérations, de disfluences et de coordinations, etc. Il convient de préciser que ce type d'exemples a parfois été délicat à déterminer, dans la mesure où – hormis les cas où les indices d'un type d'étiquette sont évidents – le choix entre deux types de phénomènes (par exemple disfluence et énumération) est subjectif. Nous sommes conscient qu'un tel choix pourrait être interprété différemment par un autre annotateur, et qu'il s'agirait alors de discuter ce choix afin de se mettre d'accord sur le statut des éventuels exemples concernés.

Exemples :

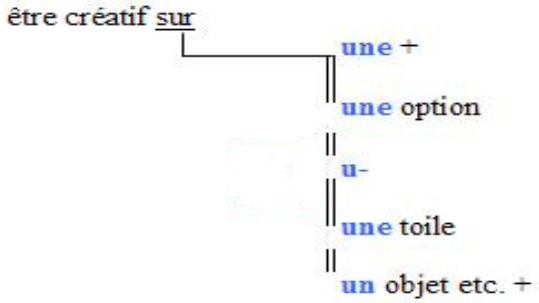
Répétition et énumération



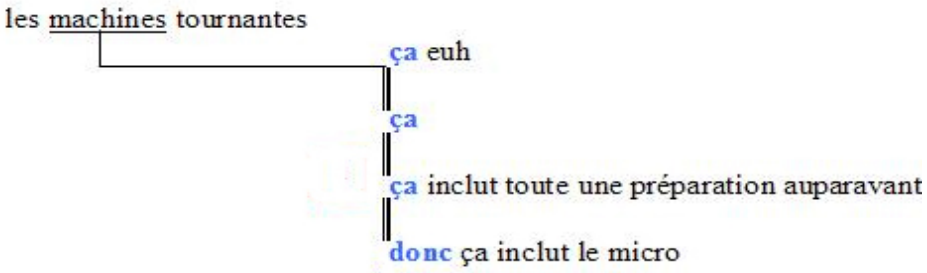
Autocorrection et énumération



Amorce et énumération



Répétition et coordination



A titre indicatif, sur les 604 piétinements syntaxiques annotés, 104 (21% du corpus total) se composent – entre autre – de cas d'énumérations. De la même façon, 537 énoncés comprennent au moins un cas de véritable disfluence. Au delà de cette dichotomie disfluences/énumération (que nous avons mis volontairement en avant du fait qu'elle pose un véritable problème d'ambiguïté en terme d'analyse automatique), nous observons d'autres caractéristiques comme nous les présentons ci-après.

Le rembobinage syntagmatique

En plus des combinaisons de disfluences, le mode de représentation arborescente révèle une autre caractéristique récurrente des énoncés étudiés : celui de « rembobinage syntagmatique ». Ce phénomène, peu évoqué dans les études sur les disfluences, (voir cependant [Martinie, 2000], [Henry, 2002b]), correspond au cas où le locuteur revient fréquemment au début du syntagme (SN, SV, SPrep, etc.) avant de compléter, corriger ou modifier son énoncé.

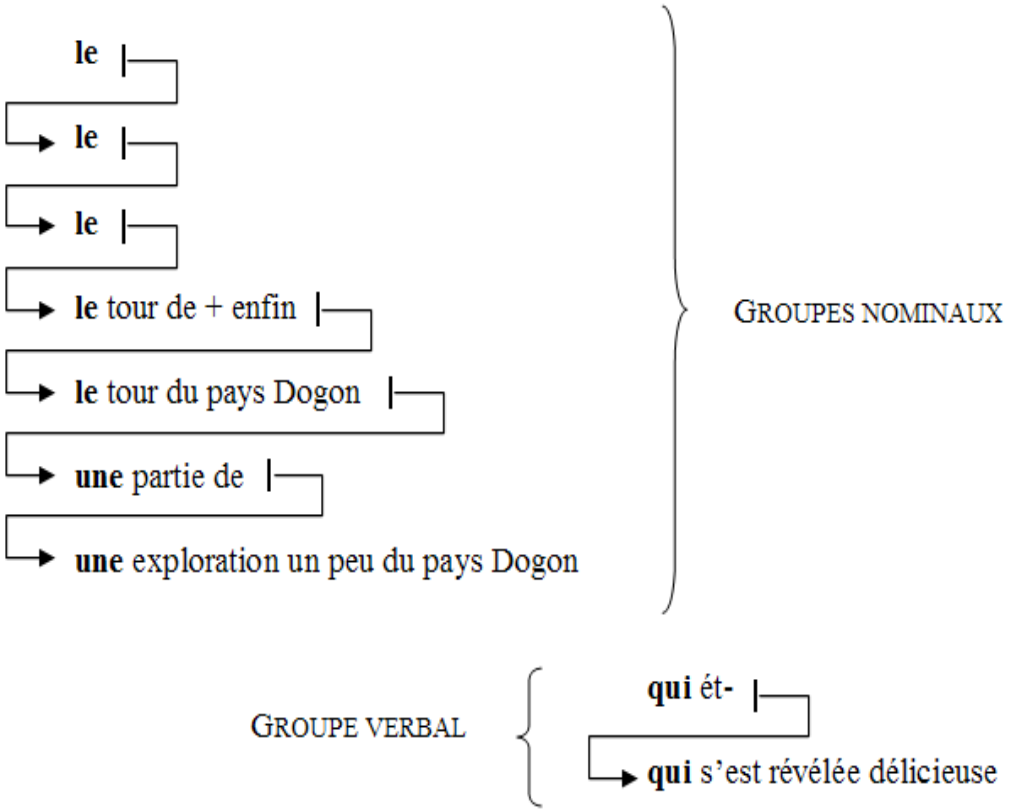


FIG. 7.7 – Exemple de « rembobinage syntagmatique ».

Les deux exemples suivants présentent respectivement des cas de disfluences sans, puis avec rembobinage syntagmatique :

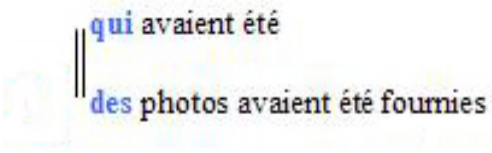


FIG. 7.8 – Exemple de disfluence **sans** rembobinage syntagmatique.

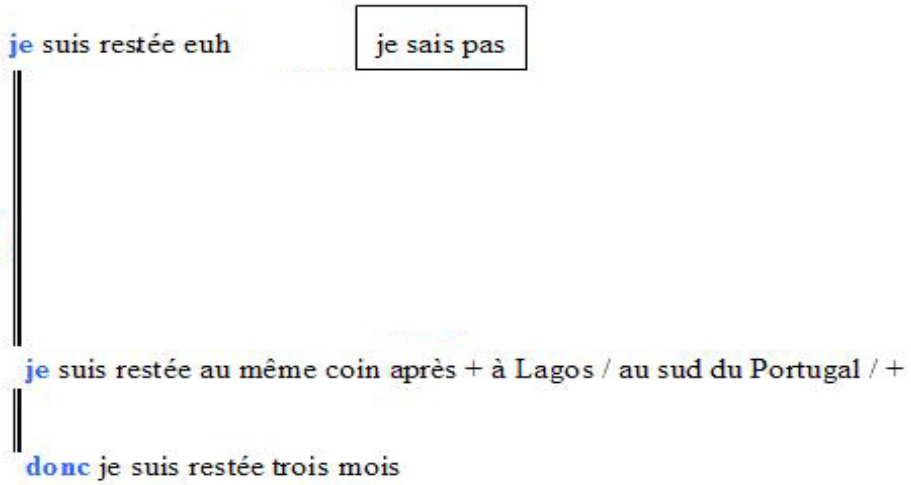


FIG. 7.9 – Exemple de disfluence **avec** rembobinage syntagmatique.

En terme quantitatif, nous relevons dans notre corpus 440 cas de piétinements syntaxiques avec rembobinage syntagmatique (soit plus de 72% du corpus total).

Cette remarque laisse alors entrevoir une piste d'implémentation exploitée par certains travaux en traitement automatique de l'oral ([Antoine *et al.*, 2003] notamment) et que nous souhaitons à notre tour approfondir, qu'est l'analyse syntaxique en constituants minimaux non-récursifs ou *chunks* ([Abney, 1991]). Un *chunk* est constitué d'un élément central, le plus souvent un nom (ou un pronom tonique) ou

un verbe (conjugué, infinitif, participe présent ou passé), entouré éventuellement à gauche de ses éléments périphériques ([Vergne, 1999]). Il s'agira dans notre cas d'observer la structure des chunks touchés par une disfluence afin d'en dégager les régularités éventuelles (cf. III).

7.3.2 Base de données d'abres marcottés

Tout comme le corpus de travail, l'ensemble des énoncés extraits pour la typologie d'arbres comporte plusieurs cas de disfluences, comprenant elles-même un certain nombre de « piétinements » syntaxiques. Afin d'organiser rigoureusement ces différents exemples, nous les avons classé au moyen d'une base de données relationnelle permettant d'afficher les différents arbres selon des caractéristiques récurrentes de chacun de ces piétinements.

Principe

Il s'agit de développer une ressource linguistique à disposition de la communauté en TAL, de façon à confronter nos données au modèle présenté, tout en soulignant ses intérêts et ses limites. La base de données proposée a pour objectif de décrire les données linguistiques et les processus de production des disfluences afin de dégager les caractéristiques de ces constructions. A notre sens, une description plus fine de ces phénomènes et leur intégration dans un modèle syntaxique formel peut permettre à terme la résolution de certains problèmes rencontrés en traitement automatique de l'oral.

Les connaissances extraites des corpus seront utilisées dans le cadre de l'analyse syntaxique automatique. Les modèles et outils développés pourront être éventuellement étendus aux constructions et phénomènes typiques de l'oral autres que les disfluences, et fournir un terrain d'expérimentation de choix pour les techniques d'analyse syntaxique robuste.

Après observations sur corpus, nous avons attribué différentes caractéristiques à

chacun des piétinements en fonction des cas suivants :

- Le piétinement contient une disfluence
- Le piétinement contient une énumération
- Le piétinement effectue un rembobinage syntagmatique
- Le piétinement est imbriqué dans un autre

Description

La base est destinée à représenter les différents attributs susmentionnés propres à un piétinement ; et dans l'optique d'améliorations futures, et ainsi pour une meilleure évolutivité, nous avons donc naturellement opté pour une base de donnée de type relationnel. Le système de gestion choisi est MySQL couplé au langage PHP pour effectuer les requêtes SQL dans la base. Le recours à ces deux choix est justifié par un avantage en terme de gain de temps, de gain de puissance et également de prix (MySQL existe en version gratuite chez tous les hébergeurs qui le supportent) vis-à-vis d'autres méthodes. Le modèle conceptuel de données correspondant à la base est présenté en annexe. Par ailleurs, cette méthode permettra de disposer de ressources de corpus oraux présentés autrement que par le biais de fichiers de transcription ou de fichiers audio parfois délicats à consulter. Les figures ci-dessous montrent respectivement l'interface graphique de la base[1] ainsi qu'un exemple de recherche et de résultats après requête.

[1] Base accessible à l'adresse suivante : http ://bove.remi.free.fr

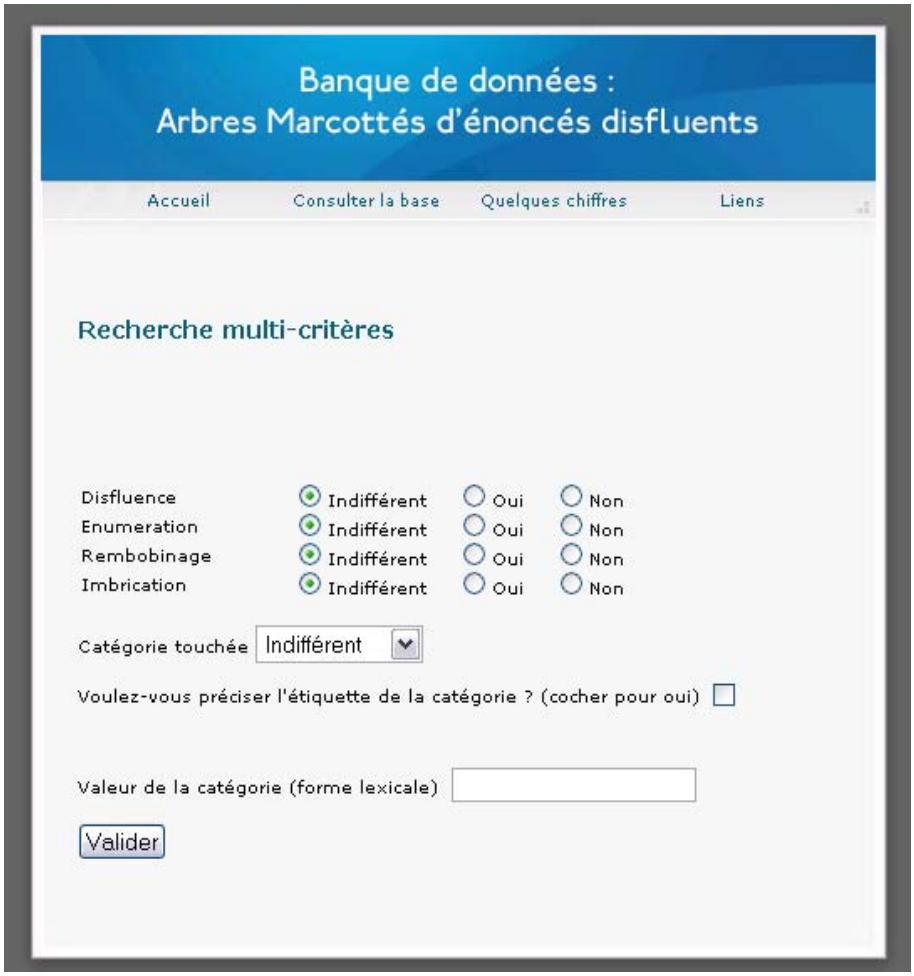


FIG. 7.10 – Interface de la base de données d'arbres marcottés.

L'interface permet d'opérer la recherche multi-critères. Il convient de noter qu'aucun critère n'est exclusif, et qu'il est par conséquent possible d'utiliser simultanément les différentes options. L'exemple ci-après correspond au cas où un utilisateur effectue la requête suivante :

- Recherche d'arbres marcottés où les piétinements contiennent une disfluence, et pas d'énumérations,
- avec présence du phénomène de rembobinage syntagmatique,
- où le critère d'imbrication est indifférent,
- où la catégorie morpho-syntaxique touchée est un déterminant.

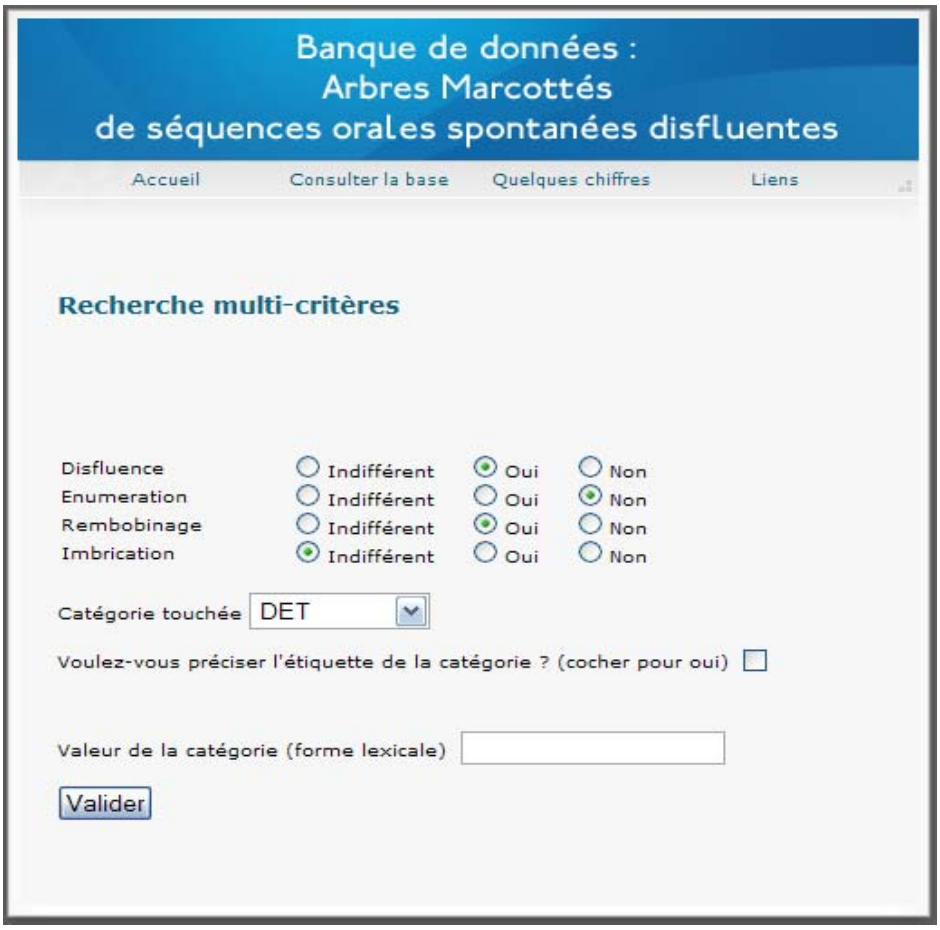


FIG. 7.11 – Exemple de requête.

Une fois la requête effectuée, l'utilisateur obtient une liste de piétinements répondant aux différentes options de la requête. Il peut ainsi – pour chacun d'entre eux – consulter l'énoncé correspondant dans le corpus transcrit, visualiser l'arbre marcotté dans son intégralité, ou visualiser chaque piétinement indépendamment du reste de l'arbre.

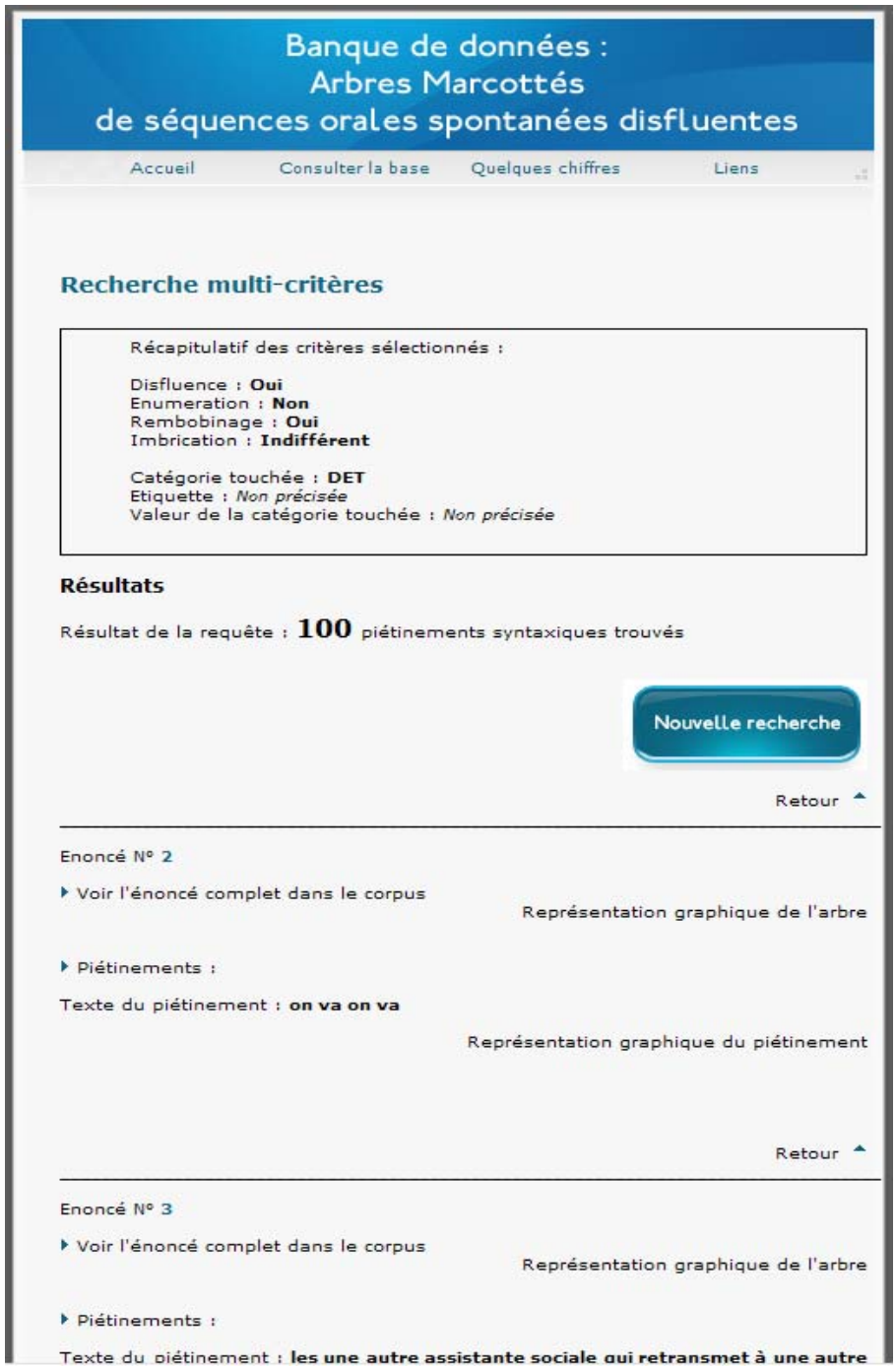


FIG. 7.12 – Exemple de résultat après requête.

À titre d'exemple, les deux figures suivantes présentent les visualisations obtenues lorsque l'utilisateur sélectionne respectivement l'arbre marcotté de l'exemple n ° 17 de la liste, puis le piétinement qui lui est associé.

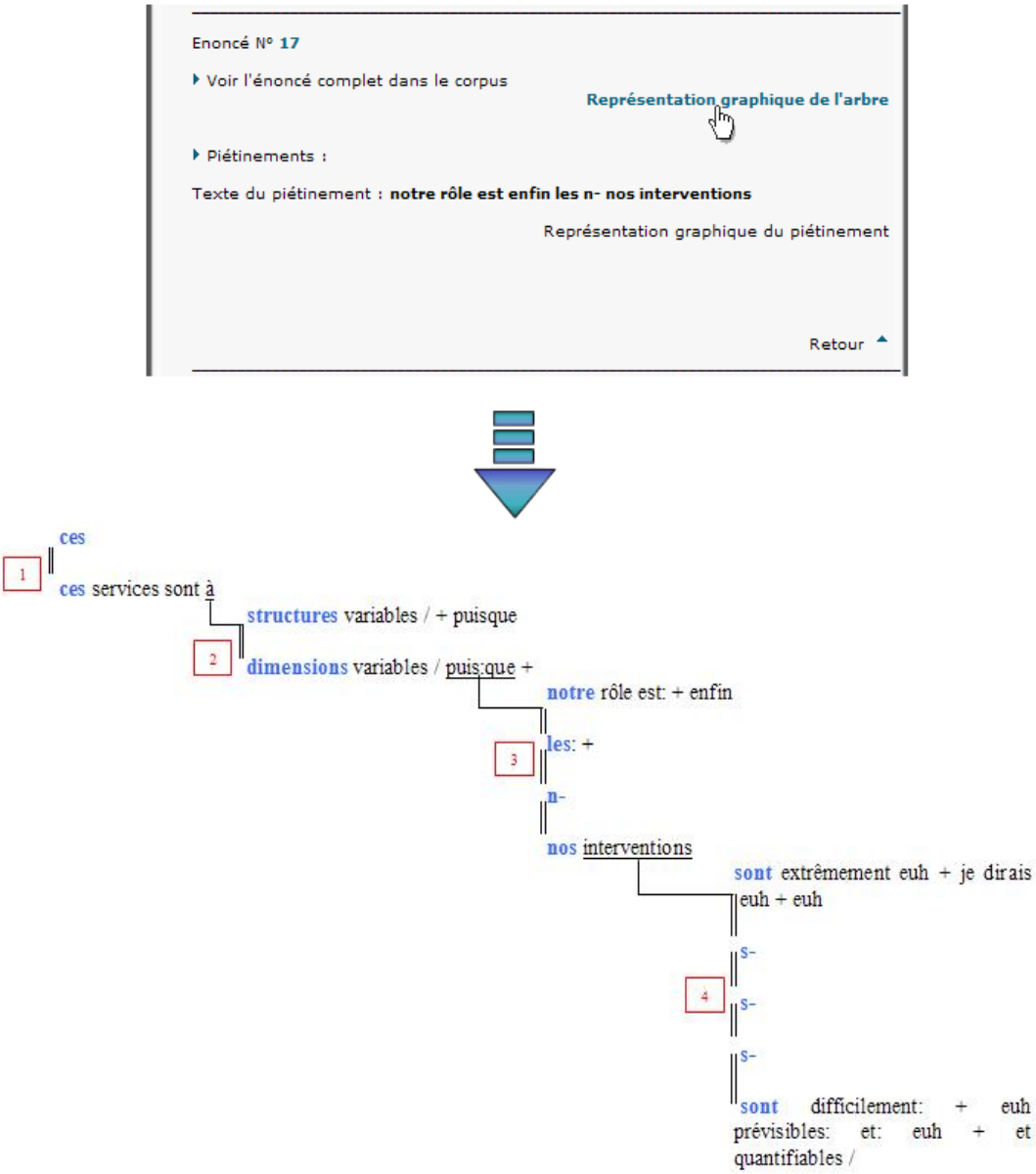


FIG. 7.13 – Exemple de résultat après requête : visualisation de l'arbre marcotté.

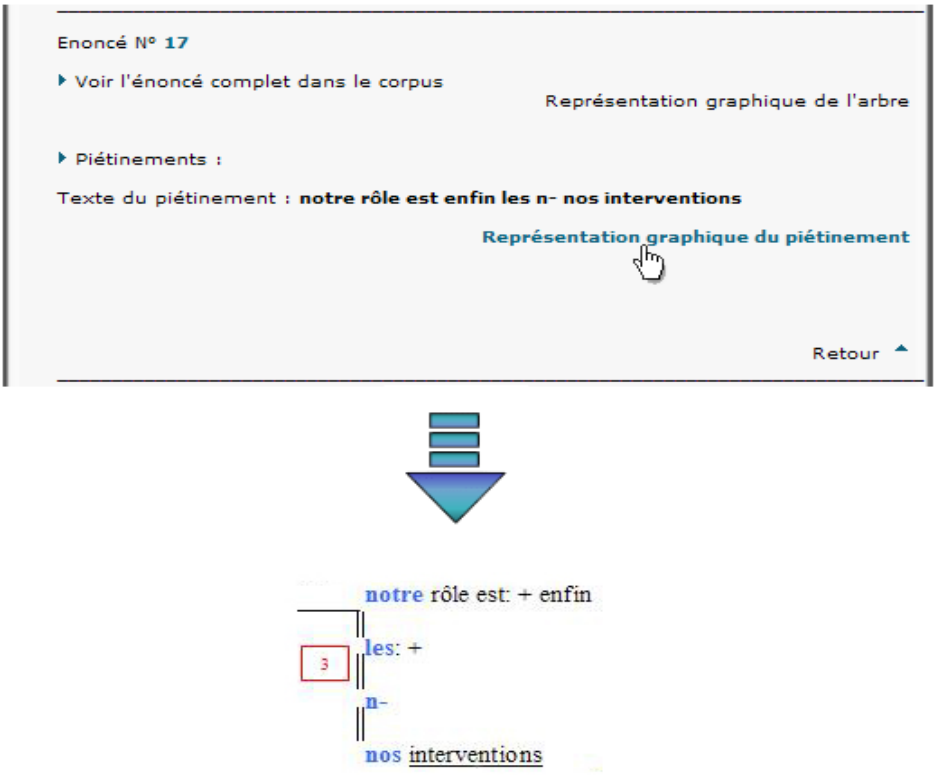


FIG. 7.14 – Exemple de résultat après requête : visualisation du piétinement.

Les critères de recherche restent limités dans l'état actuel de la base. Néanmoins, ce type de ressource peut déjà permettre d'acquérir des connaissances sur la tâche de description syntaxique et ainsi sur la syntaxe du français parlé. Pour encourager cette production de savoir, il serait intéressant à corut terme de rajouter la possibilité de rechercher un arbre par types de disfluences. À moyen terme, il s'agira d'enrichir cette base de données d'outils supplémentaires d'analyse de corpus : concordancier permettant d'effectuer des requêtes sur l'ensemble ou une partie du corpus, affiner le calcul de certaines distributions, mettre en place de nouveaux modules d'extraction permettant d'ajouter des propriétés syntaxiques sur les arbres, des interfaces de comparaison avec d'autres corpus, etc.

7.3.3 Approche linguistique des phénomènes de disfluences

Nous proposons d'étudier dans ce qui suit, les phénomènes de répétitions, d'autocorrections, d'amorces et d'inachèvements à partir des occurrences relevées dans notre corpus. Précisons que les résultats proposés dans cette partie ne s'appuient pas sur le corpus de la base de donnée (cf. 7.3.1) mais sur le corpus de développement (cf. 3.4). Il s'agit d'examiner grâce à une approche typologique d'une part les informations de structure (patrons morpho-syntaxiques) et d'autre part les marques associées à ces phénomènes.

Une stratégie d'analyse possible est d'examiner les segments disfluents avec leur contexte afin d'observer de façon empirique sur quelles catégories morpho-syntaxiques portent majoritairement chacun des phénomènes traités. L'utilisation d'un concordancier semble nécessaire dans la mesure où les concordances appuient le linguiste dans son analyse qualitative et quantitative et permet d'examiner le contexte du segment touché par la disfluence. C'est ce que nous avons réalisé à l'aide de scripts à base d'expressions régulières.

Après avoir annoté le corpus, [Piu, 2006] a réalisé une étude quantitative permet-

tant d'illustrer la typologie des phénomènes de disfluences. Une approche quantitative permet d'accéder plus facilement à la description des phénomènes qui présentent de l'intérêt et dont il aurait été difficile de cerner les contours *a priori*. A l'aide de scripts permettant d'automatiser les décomptes, l'auteur a donc quantifié les segments disfluents (quantification des types de disfluences) ainsi que les marqueurs discursifs.

Les résultats quantitatifs répertoriés dans le graphique ci-après illustrent les types de disfluences évoqués jusqu'à présent dans le corpus. Ils permettent ainsi d'avoir une idée de leur fréquence d'apparition.

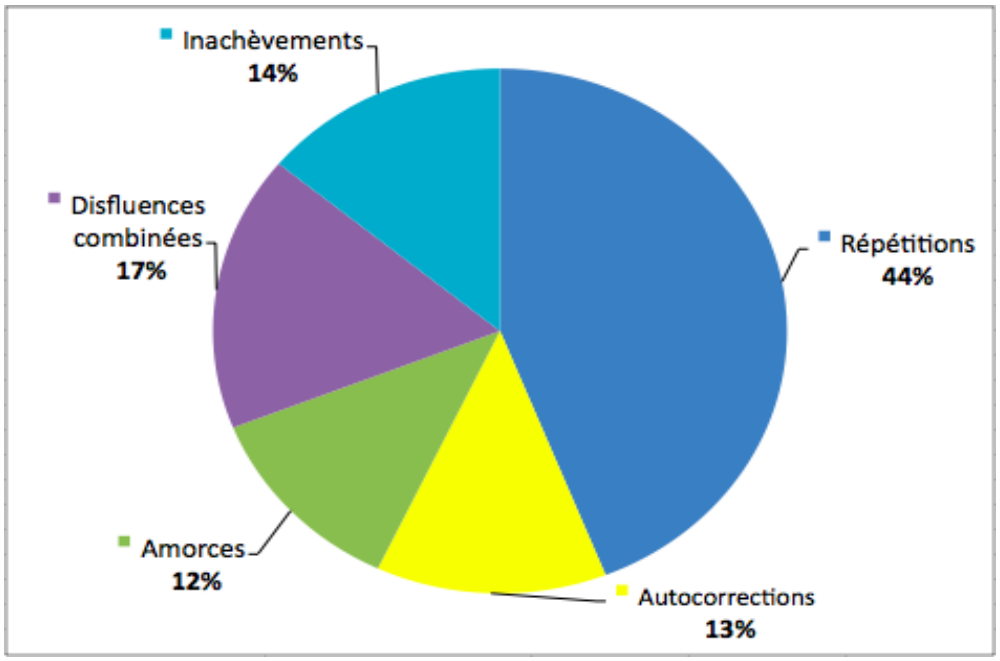


FIG. 7.15 – Répartition des types de disfluences du corpus de travail (d'après [Piu, 2006]).

En ce qui concerne le CRFP, [Piu, 2006] relève 293 disfluences : les répétitions constituent le type le plus largement représenté (44%). Les autres types sont repartis de manière plus homogène : leur pourcentage varie entre 17% pour les disfluences combinées et 12% pour les amorces. L'étude du corpus ainsi annoté soulève

d'autres questions et notamment celle de savoir quelles sont les catégories morpho-syntaxiques principalement touchées par chacun des phénomènes.

L'étude qui suit ne permet bien sûr aucune généralisation, ce n'est d'ailleurs pas notre but, mais simplement quelques remarques sur la fréquence des phénomènes attestés dans une situation de parole spontanée, ainsi que sur leurs typologies et les principales catégories morpho-syntaxiques touchées. Conscient que nous ne pouvons évidemment pas prétendre à l'exhaustivité, nous ne raisonnerons ici qu'en terme de représentativité.

Répétitions

Nous relevons 129 répétitions disfluentes, la répétition constituant le type de disfluence le plus représenté. Il convient donc d'étudier en priorité le fonctionnement intrinsèque et les contraintes distributionnelles des répétitions en intégrant la composante morpho-syntaxique.

Critères de classification de la répétition

D'après la définition proposée par [Candéa, 2000b], toute répétition se compose de deux éléments : un premier élément le « répétable » et un deuxième élément identique au premier le « répété ». Malgré ce trait commun qui permet de les définir, les répétitions ne sont pas toutes semblables et les exemples relevés dans le corpus en attestent largement. Comme nous l'avions vu (cf. 2.3.3) un certain nombre de critères permettent de trier et de classifier de manière plus fine les différents types de répétitions :

- La **longueur** du répétable c'est-à-dire le nombre d'éléments (Nel) contenus dans le motif répété (le répétable). Certains motifs ne présentent qu'une seule unité (ex. a et b) là où d'autres en comptent plusieurs (ex. c et d).

  (a) **et et** *sans doute pas beaucoup d' argent en poche d'ailleurs* (Nel = 1)

  (b) **le le** *nord euh de l' Espagne* (Nel = 1)

  (c) **ça serait ça serait** *néfaste pour toi* (Nel =2)

  (d) **il y a il y a** *une remise en question* (Nel= 3)

- L'**empan** de la répétition, c'est-à-dire le nombre de fois où les éléments constituants du répétable sont répétés (Nr) après le répétable.

  (a) **tu as tu as** *juste à brancher ta machine* (Nr = 1)

  (b) **de de de** *préparer* (Nr = 2)

  (c) **la la la la la** *botanique bon il y a une multitude* (Nr = 4)

– La **succession** des termes répétés c'est-à-dire la présence ou non d'autres éléments à l'intérieur de la répétition : certains termes sont produits en contiguïté alors que pour d'autres on relève la présence de divers éléments : allongement, pause silencieuse, pause remplie ou bien un autre mot ou groupe de mots.

En nous appuyant sur ces différents critères, il est possible d'opérer une classification des cas de répétitions rencontrés. Nous décrirons les cas relevés en nous intéressant aux caractéristiques du répétable en terme de longueur et de catégorie morpho-syntaxique, avant de décrire plus précisément les autres configurations rencontrées : nombre de répétés, association d'éléments à l'intérieur de la répétition.

Longueur du répétable et catégorie morpho-syntaxique

Après observations sur corpus, nous proposons un premier classement des répétitions en fonction de la longueur du répétable.

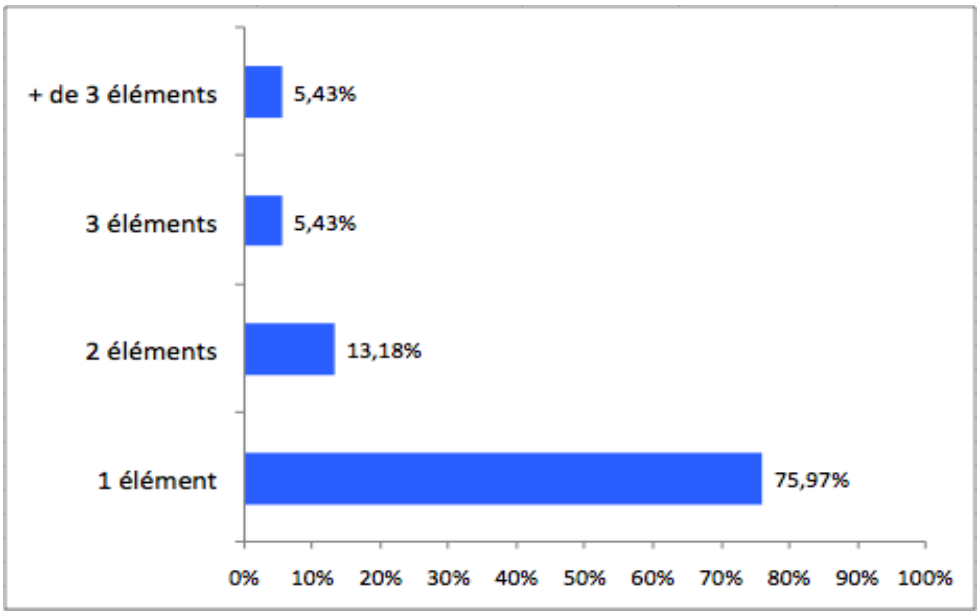


FIG. 7.16 – Nombre d'éléments contenus dans le répétable.

Sur la base des résultats obtenus, on peut dire que la répétition porte principale-

ment sur une seule unité lexicale, (dans 76% des cas) et on observe que ces unités sont souvent monosyllabiques (*la*, *un*, *de*, *et*, etc.) Dans 13% des cas, elle porte sur deux unités et dans seulement 5% des cas, le répétable contient trois éléments ou plus. Voici quelques exemples relevés dans le corpus :

Nel = 1 :

la **la** *qualification au niveau du pressing*

deux *euh* **deux** *unités importantes*

Nel = 2 :

ça peut **ça peut** *passer très bien*

il faut **il faut** *mettre un vin*

Nel = 3 :

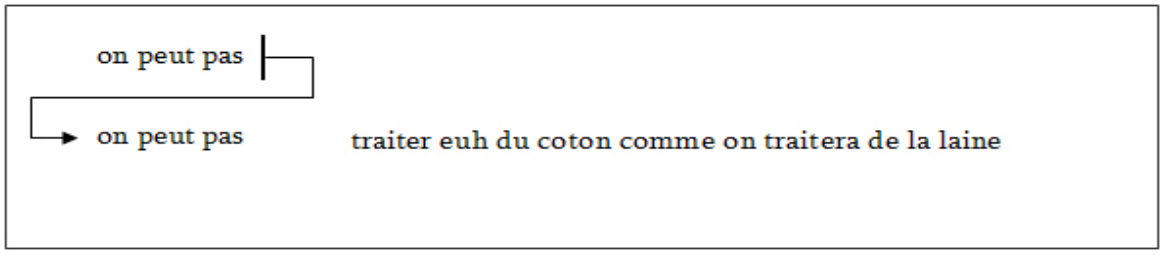
lorsque les circonstances sont plus favorables et que le **et que le** *marché le permet*

on parle de **on parle de** *l' ouvrée*

Nel = 4 :

le coq au vin euh est un de ces **est un de ces** *plats*

Il est à noter que pour les répétitions qui comportent plusieurs éléments dans le répétable (Nel > 1), la reprise (élément répété) a lieu toujours en revenant au début du syntagme (cf. 7.3.1). Exemple : [on peut pas] **on peut pas** *traiter euh du coton comme on traitera de la laine*



Catégories morpho-syntaxiques de la répétition (répétable unique)

Nous nous intéressons à présent à la nature du répétable en terme de catégorie morpho-syntaxique en nous basant sur l'étiquetage effectué par TreeTagger (voir 9.3.2). Le classement réalisé nous permet de faire une première remarque : les répétitions à répétable unique touchent plus fréquemment des mots « grammaticaux ». Les mots « grammaticaux » (ou mots « vides ») par opposition aux mots « lexicaux » (ou mots pleins) ont un contenu plus large et se définissent davantage par leurs fonctions syntaxiques que par leur contenu sémantique. Ils actualisent quand ils sont pronom, article, etc. Ils assemblent quand ils sont préposition ou conjonction par exemple. En général, la classe très large des mots « grammaticaux » englobe les déterminants, les pronoms, les coordonnants (par exemple les conjonctions de coordination ou les adverbes de liaison), ainsi que les subordonnants (par exemple les prépositions, les conjonctions de subordination etc.)

Ainsi, sur nos 98 cas de répétitions à répétable unique, on observe une grande proportion de mots grammaticaux qui sont touchés par la répétition. Le tableau qui suit décrit les différentes catégories morpho-syntaxiques du répétable et leurs proportions dans le corpus. Cette catégorisation précise des éléments concernés par la répétition nous fournit des informations sur le fonctionnement des répétitions : elles ne portent pas sur n'importe quelle catégorie grammaticale, ni sur n'importe quelle classe de mots.

Les catégories les plus concernées par la répétition sont les articles (définis [*le*, *la*, etc.] ou indéfinis [*un*, *une*, etc.]), les prépositions (*en*, *à*, *dans*, etc.), les « prépositions + articles » (*au*, *aux*, *du*, *des*).

Nombre d'éléments = 1			
Etiquette	Valeur	Effectif absolu	Effectif relatif
DET :ART	Article	29	29,59%
PRP	Préposition	21	21,53%
PRP :det	Préposition + article (du, des, aux)	14	14,59%
PRO :PER	Pronom personnel	9	9,18%
PRO :REL	Pronom relatif	5	5,10%
DET :POS	Proom possessif	4	4,08%
KON	Conjonction	4	4,08%
PRO :DEM	Pronom démonstratif	4	4,08%
ADV	Adverbe	3	3,06%
VER	Verbe	3	3,06%
NUM	Nombre	2	2,04%
Total		**98**	**100%**

TAB. 7.1 – Catégories touchées par la répétition (répétable unique)

Sur les 11 catégories représentées dans le tableau précédent, 9 sont des catégories appartenant à la classe des mots « grammaticaux » (pronom, préposition, conjonction etc.) et seulement 2 d'entre elles appartiennent à la classe des mots « lexicaux » (verbe, adverbe, nom etc.). Ce qui nous donne la proportion suivante :

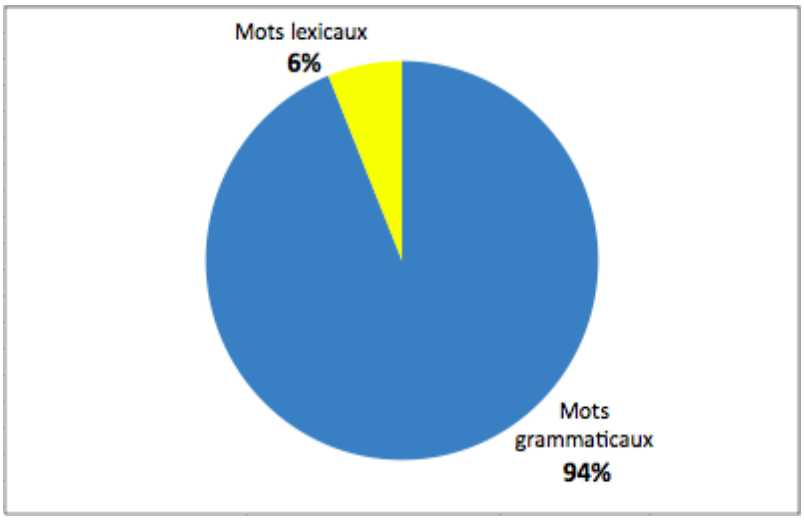


FIG. 7.17 – Répartition mots grammaticaux / mots lexicaux.

- Exemples de répétitions portant sur les mots grammaticaux :
  - Article (DET :ART)

    j'avais bien contourné **le le** *nord de l'Espagne*

  - Préposition (PRP)

    c'est toujours euh **en en en** *vigueur ça hein*

  - Préposition + article (PRP :det)

    en dehors évidemment **des des** *outils euh médiatiques locaux hein*

  - Pronom personnel (PRO :PER)

    on on *a trouvé que quand on leur donnait tout en polycops*

  - Pronom relatif (PRO :REL)

    tous les agents **qui qui** *font des contrôles*

  - Pronom possessif (DET :POS)

    de par **mon mon** *emploi je vais leur vendre un produit*

  - Conjonction (KON)

    en gros c'est ça **mais mais** *ça pose problème*

  - Pronom démonstratif (PRO :DEM)

    on a par exemple **ces ces** *fameux oeufs en meurette*

  - Nombre (NUM)

    donc **vingt-deux vingt-deux** *agents permanents*

- Exemples de répétitions portant sur des mots lexicaux
  - Verbes (VER)

    ces comités départementaux **sont** *un peu* **sont sont** *exactement le le reflet de ce qu'il y a à Paris*

  - Adverbe (ADV)

    je pense que c'est quand même **assez assez** *bien*

Catégories morpho-syntaxiques de la répétition (répétable à deux unités)

Sur les 17 occurrences de répétables formés à partir de deux unités, on relève que dans 53% des cas, il s'agit d'un pronom (personnel, démonstratif, etc.) suivi d'un verbe conjugué. Parmi ces formes, on distingue principalement les patrons morpho-syntaxiques suivants :

- Pronom personnel + verbe (PRO :PER + VER) [*30% des cas*]

  il faut il faut *faire tout ça*

  tu as tu as *juste à brancher ta machine*

- Pronom démonstratif + copule conjuguée (PRO :DEM + VER) [*23% des cas*]

  ça peut ça peut *passer très bien il y a des fiches pastoralisme*

  grâce à notre secrétaire comptable euh et **c'est c'est** *une très grosse partie et relativement ingrate de notre travail*

Au regard des cas relevés (seulement 17 occurrences de répétitions qui comportent deux éléments dans le répétable), il faudrait nuancer ces décomptes et effectuer ces

observations à partir de données plus importantes. Il en va de même des répétables constitués de trois unités ou plus, pour lesquels il est impossible de formuler des hypothèses de fonctionnement et des régularités puisqu'il s'agit de cas dont la fréquence d'apparition est très faible. Là encore, une étude sur un plus grand nombre de cas est nécessaire.

Combinatoire de la répétition

Après avoir classé nos répétitions en fonction de la longueur du répétable et de la catégorie morpho-syntaxique, nous proposons désormais d'observer la combinatoire du phénomène de répétition en reprenant la typologie établie au début de notre étude concernant les répétitions.

Cette typologie permet de classer la majeure partie de nos répétitions mais surtout elle présente l'avantage de combiner les critères d'observations : la succession des termes répétés (présence ou non d'autres éléments à l'intérieur de la répétition) ainsi que le nombre d'éléments répétés. Rappelons qu'une distinction est à opérer entre répétition « simple » (répétable suivi d'un seul élément répété) et répétition « multiple » (qui contient plusieurs éléments répétés).

Ces deux types de répétitions peuvent être « directes » c'est-à-dire produites en contiguïté, ou « associées », elles sont alors accompagnées d'autres marques de production : pause silencieuse, pause remplie, mots (parenthétiques), etc.

Le tableau suivant présente la répartition des différents types de répétitions rapportée à la totalité des occurrences du corpus :

	Directe	Associée	Total
Simple	98	15	113
Multiple	11	5	16
Total	109	20	129

TAB. 7.2 – Répartition des types de répétitions

On constate d'après ce tableau que les répétitions simples sont largement majoritaires puisqu'elles représentent, de par leur effectif, 88% du total. La répétition multiple (qui comporte plus d'un élément répété) ne représentant que 12% sur le total des occurrences. Ces chiffres nous éclairent sur le fonctionnement de la répétition : le répété unique va être plus souvent rencontré que plusieurs éléments répétés. Ainsi, dans la plupart des cas, lorsque le locuteur produit une répétition, il ne va répéter qu'une seule fois l'élément (répétable + répété).

– Répétitions simples (répété unique)

  il **il** *y aura pas d' authenticité derrière*

  (Nb de répétés = 1)

  c'est **c'est** *même pas la peine même de l' eau c' est pas bon*

  (Nb de répétés = 1)

– Répétitions multiples (plusieurs répétés)

  là où on entreposait avant dans les **dans les dans les** caves les huiles

  (Nb de répétés = 2)

  il y a euh des fiches euh euh sur euh sur la **la la la la** *botanique*

  (Nb de répétés = 4)

Enfin, un autre point relevé dans le tableau, la tendance des répétitions à apparaître

en association avec une autre marque de production, qu'il s'agisse de répétitions simples ou multiples. En effet, dans certains cas (16%) la répétition est associée à un autre phénomène de production : pauses remplies, parenthétiques, marqueurs discursifs, etc. (insérés entre le répétable et le répété).

- Répétitions associées avec pauses remplies

  c'est exactement **la même** euh euh **la même** *représentation en plus petit dans les départements*

- Répétitions associées avec parenthétique

  le théâtre lopesque **traduit un** grosso modo bien évidemment **traduit un** *préjugé*

- Répétitions associées avec marqueur discursif

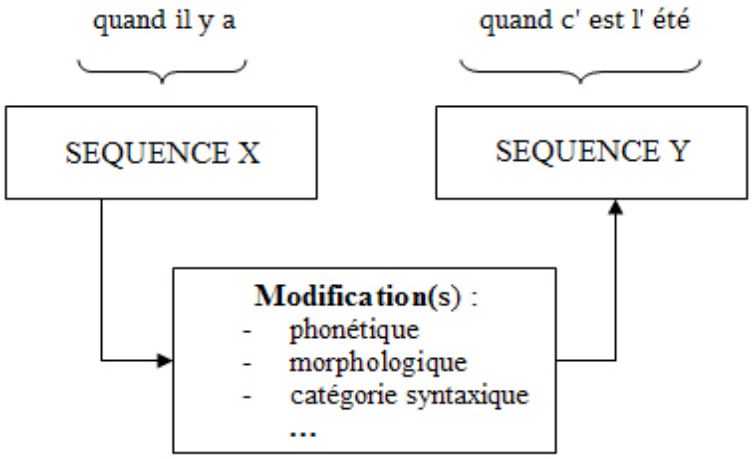
  des choses **qui** bon **qui qui** *sont jolies*

Autocorrections

Séquence d'origine et séquence corrigée

Nous relevons 37 occurrences d'autocorrections dans le corpus ce qui correspond à 13% sur le total des disfluences. Rappelons brièvement la terminologie et la formalisation employée par [Candéa, 2000b] (inspirée de [Levelt, 1983]) pour l'autocorrection, où X définit la séquence d'origine et Y la séquence autocorrigée.

Exemple :



La première remarque que l'on peut faire en observant les cas d'autocorrections est que la nature morpho-syntaxique de la séquence d'origine et de la séquence autocorrigée sont bien souvent similaires. Ainsi, dans 73% des cas, les catégories morpho-syntaxiques dégagées pour chaque séquence (X et Y) sont strictement identiques. Nous donnons ci-après quelques exemples de patrons identiques.

– PRO :CLI PRO :CLI

  on nous *avons une photocopieuse*

– PRO :DEM PRO :DEM

  ce cette *envie de voyager*

– ADV ADV

très *euh* **toujours** *prête à me défendre*

– PRO :REL VER :conj :s3 PRO :REL VER :conj :s3
un vin **qui soit qui ait** *assez de corps*

– PRO :CLI VER :conj :s3 PRO :CLI VER :conj :s3
on va on retourne *à l' hôpital*

Il est à noter également que certains des patrons que nous avons relevé diffèrent au niveau de l'étiquette morpho-syntaxique attribuée par TreeTagger, mais que la séquence d'origine et la séquence autocorrigée appartiennent néanmoins à la même catégorie de mot. Prenons un exemple simple pour illustrer ce cas de figure :

DET :ART :**ms3** DET :ART :**s3** NOM NOM NOM
le l' alternance stage école

Dans cet exemple, l'article *le* (DET :ART :ms3) est remplacé par l'article *l'* (DET :ART :s3). Si on adopte un niveau de détail moindre que celui réalisé dans l'étiquetage, on peut dire que ces deux formes appartiennent à la catégorie générique des articles et que la séquence d'origine et la réparation ont la même nature et assurent les mêmes fonctions dans l'énoncé.

L'autocorrection s'apparente à la répétition dans la mesure où les catégories morpho-syntaxiques de la séquence d'origine et de la séquence autocorrigée sont souvent semblables ou diffèrent au niveau de l'étiquetage mais appartiennent néanmoins à la même catégorie générique. On peut donc imaginer que, dans de nombreux cas, l'autocorrection peut être traitée de manière analogue à la répétition, et notamment pour l'adaptation des règles syntaxiques pour l'oral.

De la même façon, l'autocorrection peut rappeler certaines caractéristiques de l'amorce. En effet, les deux phénomènes se rejoignent en partie dans la mesure où il y a – dans les deux cas – une modification de la séquence d'origine, exprimée partiellement dans le cas de l'amorce, et complète dans le cas de l'autocor-

rection. Rappelons toutefois que la distinction que nous opérons entre ces deux types disfluences repose justement sur cette différence de mode production « complet/incomplet ».

Catégorie morpho-syntaxique de la séquence d'origine

Il s'agit d'observer à présent la catégorie morpho-syntaxique de la séquence d'origine (séquence qui fait l'objet de la correction) : quelles sont les unités qui sont le plus souvent touchées par l'autocorrection ?

Le tableau suivant présente l'effectif des différentes catégories morpho-syntaxiques de la séquence d'origine :

Séquences d'origine formées d'une seule unité	
Catégorie	Effectif absolu
Préposition	7
Article	4
Verbe	3
Pronom possessif	2
Adverbe	1
Pronom personnel clitique	1
Pronom démonstratif	1
Total	**19**

TAB. 7.3 – Catégories grammaticales de la séquence d'origine (une seule unité)

Le tableau ci-dessus montre que les catégories touchées par l'autocorrection (séquence d'origine formée à partir d'une seule unité) sont principalement des mots grammaticaux. En effet, dans 79% des cas, la séquence qui va faire l'objet d'une autocorrection porte sur un mot grammatical (pronoms, prépositions, etc.) Rappelons que nous avons eu les mêmes observations concernant la répétition décrite précédemment.

– Séquences d'origine formées d'**une seule unité** :

le l' *hiver* (Article)

le Romanée-Conti qui est issu **de du** *village de Vosne-Romanée* (Préposition)

mon un *vieux collègue de sciences naturelles* (Pronom possessif)

cette chienne qui était là euh **très** euh **toujours** *prête à me défendre* (Adverbe)

Lorsque la séquence d'origine est formée de plusieurs unités, il est moins évident d'établir une classification dans la mesure où les patrons relevés correspondent pour la grande majorité (83% des cas) à des hapax. Le tableau qui suit, présente les différents patrons relevés :

Séquences d'origine formées de plusieurs unités	
Patrons	Effectif absolu
Pro pers cli + verbe conj	3
Conj + pro pers cli + pro pers cli + verbe conj	1
Pro dém + verbe conj	1
Prép + art + prép	1
Pro pers cli + pro pers cli	1
Conj + pro pers cli + pro pers cli	1
Pro rel + verbe conj	1
Prép + verbe inf	1
Prép + art	1
Pro pers cli + pro pers cli + verbe conj + adv	1
Conj + pro pers cli + pro pers cli + verbe conj	1
Prép + nom propre	1
Art + nom	1
Nom + adj	1
Pro pers cli + verbe conj + adv + conj	1
Prép + art + nom + adj	1
Total	**18**

TAB. 7.4 – Patrons de la séquence d'origine (plusieurs unités)

– Séquences d'origine formées à partir de **plusieurs unités** :

j'avais *j'étais en maîtrise* (Pronom personnel clitique + verbe conjugué)

on sait pas quand *euh on a du mal à prévoir* (Pronom personnel clitique + verbe conjugué + adverbe + conjonction)

je peux parler **d'un de** *pas un de mes premiers voyages* (Préposition + article + préposition)

Autocorrections simples et complexes

Nous tentons à présent de classifier de manière plus fine nos cas d'autocorrections. Pour cela, nous proposons de distinguer deux grandes catégories d'autocorrections : les autocorrections simples et les autocorrections complexes. Cette distinction s'appuie essentiellement sur le nombre de traits linguistiques (morphologiques et phonétiques) qui sont modifiés entre la séquence d'origine et la séquence autocorrigée. Ce classement nous a semblé pertinent dans la mesure où les patrons morpho-syntaxiques de la séquence d'origine et celle qui est conservée au final sont souvent identiques et il n'est donc pas possible d'observer les traits modifiés en nous basant uniquement sur la composante morpho-syntaxique.

Dans les autocorrections simples, un seul trait est modifié. Il peut s'agir d'un trait grammatical (genre, nombre, personne, temps, etc.) ou phonétique, ou encore il peut s'agir d'un changement de catégorie morpho-syntaxique. Nous incluons aussi dans cette catégorie les cas d'autocorrections où aucun trait ne change, mais le locuteur opère un changement au niveau du lexique sans opérer de modification au niveau des traits linguistiques ou de la catégorie morpho-syntaxique.

Trait grammatical modifié (genre) : *toujours* **ce** *cette envie de voyager*

Trait phonétique modifié : *les* **comitons** *les comités départementaux*

Les autocorrections complexes englobent les cas d'autocorrections où plus d'un

trait est modifié. Comme cette catégorie se veut générique, nous incluons aussi les cas où le locuteur opère un changement de construction et dans lesquels l'analogie entre séquence d'origine et séquence autocorrigée est moins probante. Sont également considérées comme des cas complexes, les autocorrections qui s'étendent sur plusieurs mots et qui comptent plusieurs piétinements syntaxiques (accompagnés ou non de commentaires).

Plusieurs traits modifiés (ajout de lexique et forme élidé remplacée par sa forme neutre) :

d'être euh de pas pouvoir répondre euh aux questions

Changement de construction :

l'ONIC (...) *recrute le recrutement s'effectue euh* (...)

Plusieurs piétinements syntaxiques (retouches successives) :

à l'étage supérieur enfin au-dessus à un autre niveau

Il à noter que dans cette classification, nous ne proposons pas de distinguer les cas où l'autocorrection porte sur une seule unité ou un syntagme entier car ce phénomène comme celui de la répétition représente une reprise de tout un ensemble de traits morphologiques, syntaxiques et phonétiques à l'exception de ceux qui sont modifiés.

À partir de cette classification, nous donnons ci-après la répartition des types d'autocorrections. Les autocorrections simples sont majoritaires puisqu'elles représentent 70% du total des occurrences relevées.

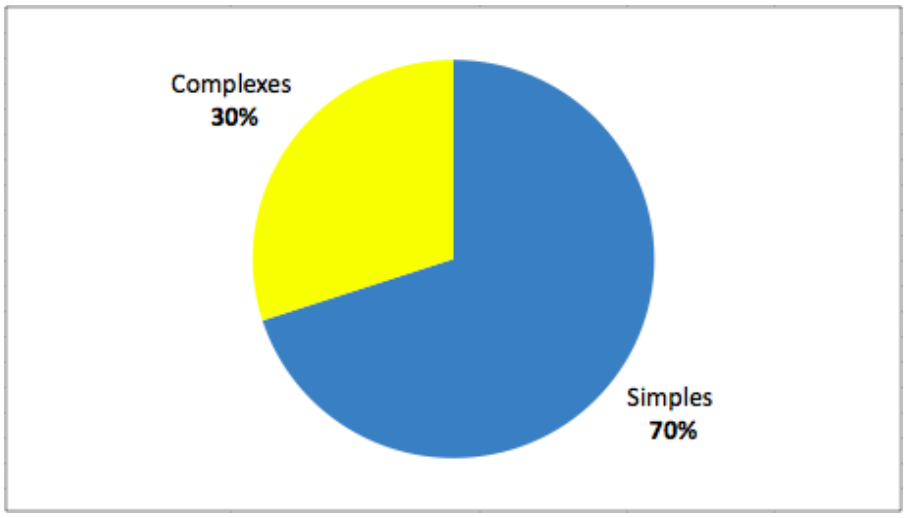


FIG. 7.18 – Répartition des autocorrections simples et complexes.

Autocorrections simples

Le diagramme qui suit présente les traits linguistiques modifiés dans l'autocorrection simple. A partir des résultats obtenus, on peut principalement noter que dans 38% des cas, la modification concerne un changement d'ordre lexical et dans 27% des cas, elle porte sur la catégorie morpho-syntaxique.

– Modification d'un trait portant sur le **choix du lexique** :

on **va** *on* **retourne** *à l' hôpital*

sont à **structures** *variables puisque* **dimensions** *variables*

– Modification d'un trait portant sur la **catégorie morpho-syntaxique** :

ça *devient euh* **tout** *devient difficile* (Pronom démonstratif ⟶ Pronom indéfini)

mon un *vieux collègue de sciences naturelles* (Pronom démonstratif ⟶ article)

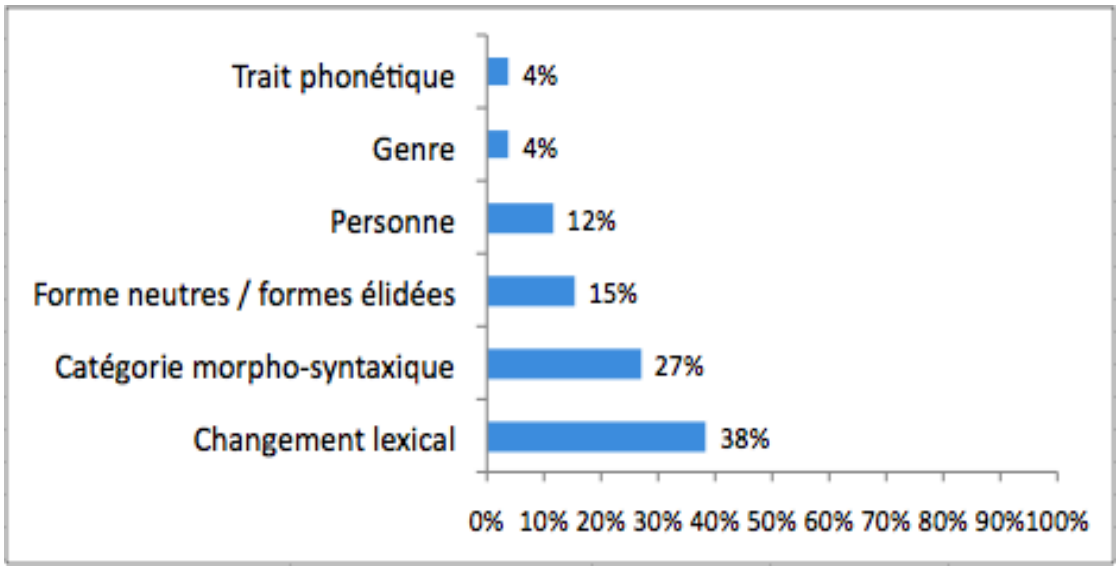


FIG. 7.19 – Traits linguistiques modifiés dans l'autocorrection simple.

– Modification d'un trait portant sur les **formes neutres et élidées** :

les habitants **de d'** *Aloxe-Corton*

de *mmh* **d'***assurer le revenu des producteurs*

– Modification d'un trait portant sur la **personne** :

que **je** *me penche que* **nous** *penchions ensemble sur le problème*

– Modification d'un trait portant sur le **genre** :

toujours **ce cette** *envie de voyager*

– Modification d'un trait **phonétique** :

les **comitons** *les* **comités** *départementaux*

Autocorrections complexes

– **Plusieurs traits modifiés** :

j'ai je voulais *passer à Santiago de Compostelle*

Dans cet exemple, trois traits sont modifiés : le temps (passé composé / imparfait), le lexique (verbe *avoir* → verbe *vouloir*) et la forme neutre qui est remplacée par sa forme élidée correspondante (*j'* → *je*).

d'être *euh* **de pas pouvoir** *répondre euh aux questions*

Ici, le locuteur opère un changement au niveau du lexique (verbes *être* et *pouvoir*) où est également inséré un mot (*pas*). De plus, un changement au niveau de la préposition est opéré : la forme élidée est remplacée par sa forme neutre correspondante (*d'* → *de*).

– **Changement de construction** :

l'ONIC (...) **recrute le recrutement** *s'effectue euh* (...)

L'exemple ci-dessus correspond à une reformulation où le locuteur modifie le verbe employé pour utiliser plutôt le substantif correspondant.

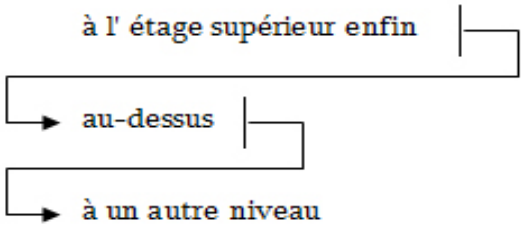
notre un de nos *principaux outils*

Dans cet exemple, le locuteur modifie une construction au singulier (*notre*) pour généraliser son propos à l'aide de la construction *un de nos* qui sous-entend qu'il n'y a pas un mais plusieurs *outils*.

– **Plusieurs piétinements syntaxiques** (retouches successives) :

à l' étage supérieur *enfin* **au-dessus** *à* **un autre niveau**

Dans cet énoncé, il faut au locuteur trois piétinements successifs (sémantiquement équivalent) pour exprimer le sens de son énoncé, et on assiste une nouvelle fois au phénomène de « rembobinage syntagmatique » portant ici sur un syntagme prépositionnel :


à l' étage supérieur enfin
au-dessus
à un autre niveau

Amorces

Typologie des amorces

Nous relevons 34 cas distincts d'amorces dans notre corpus d'étude. Rappelons que les amorces se traduisent par une interruption de morphèmes en cours d'énonciation et qu'elles sont codées dans le corpus au moyen d'un tiret accolé au mot. [Roubaud, 2004] considère que les amorces sont des unités de sens ainsi que des représentants de constituants syntaxiques. C'est en partie grâce au contexte que l'on peut identifier ces unités de sens et les syntagmes correspondants.

Nous avions vu dans la première partie de notre travail que l'on peut distinguer trois grandes catégories d'amorces qui peuvent nous permettre d'établir une première classification de nos cas.

– Amorces **complétées** : la complétion du mot amorcé se situe sur un même emplacement syntaxique.

  à **m-** *à mon anniversaire*

– Amorces **modifiées** : le mot amorcé n'est pas complété mais remplacé par un autre mot sur le même emplacement syntaxique.

  des **a-** *des sujets*

– Amorces **inachevées** : il n'y a pas de piétinement sur une même place syntaxique car ce qui suit le mot amorcé occupe une autre place syntaxique.

  ne sont pas si étendus que ça hein **pas-** *euh par exemple*

Les deux premières catégories (amorces complétées et modifiées) renvoient à une recherche lexicale s'effectuant sur une même place syntaxique tandis que la troisième catégorie (amorces inachevées) introduit une rupture syntaxique plus importante dans la mesure où le morphème amorcé n'est ni complété, ni modifié. En fonction

des catégories identifiées précédemment, nous obtenons la répartition suivante pour les cas relevés dans notre corpus :

Types	Effectif absolu	Effectif relatif
Amorces complétées	20	59%
Amorces modifiées	13	38%
Amorces inachevées	1	3%
Total	**34**	**100%**

TAB. 7.5 – Répartition des types d'amorces : effectif et pourcentage

Les amorces complétées sont majoritaires dans ce corpus puisqu'elles représentent 59% des cas ; viennent ensuite celles qui sont modifiées (38% des cas). Les amorces inachevées sont anecdotiques puisque nous ne relevons qu'une seule occurrence.

– Exemples d'amorces complétées :

*l'***en-** *l' envie de liberté*

*l'***ai-** *aigle royal*

c'est **p-** *c'est peut-être péjoratif*

– Exemples d'amorces modifiées :

avoir des petites **bou-** *fioles d'huile de partout*

un **com-** *un conseil central*

elles se sentent **remet-** *remises en question*

– Exemple d'amorce inachevée :

ne sont pas si étendus que ça hein **pas-** *euh par exemple*

Catégories morpho-syntaxiques de l'amorce

Nous nous intéressons ici à la catégorie morpho-syntaxique de l'amorce en nous aidant du contexte. En effet, l'étiquetage morpho-syntaxique bute sur ces éléments qui se voient le plus souvent attribuer une étiquette erronée. Il s'agit pour nous de reconstituer le morphème inachevé et de lui attribuer l'étiquette morpho-syntaxique correspondante. L'examen du contexte de l'amorce facilite l'interprétation et permet de déterminer à quelle catégorie morpho-syntaxique appartient le morphème interrompu. Dans le cas des amorces complétées, les plus nombreuses, il est plus facile d'identifier les unités de sens et les syntagmes qui correspondent aux morphèmes amorcés dans la mesure où ils sont complétés et qu'il est possible d'établir une corrélation entre l'élément amorcé et l'élément complété. En revanche, l'identification des amorces modifiées nécessite une part d'interprétation plus importante.

Sur les 34 cas d'amorces, un cas seulement n'est pas interprétable à l'aide du contexte et nous ne pouvons donc pas déterminer l'étiquette morpho-syntaxique de l'élément amorcé. Le cas en question appartient à la catégorie des amorces inachevées :

> *ne sont pas si étendus que ça hein* **pas-** *euh par exemple*

Les catégories morpho-syntaxiques les plus touchées (amorces complétées et modifiées confondues) concernent les mots lexicaux avec la répartition suivante :

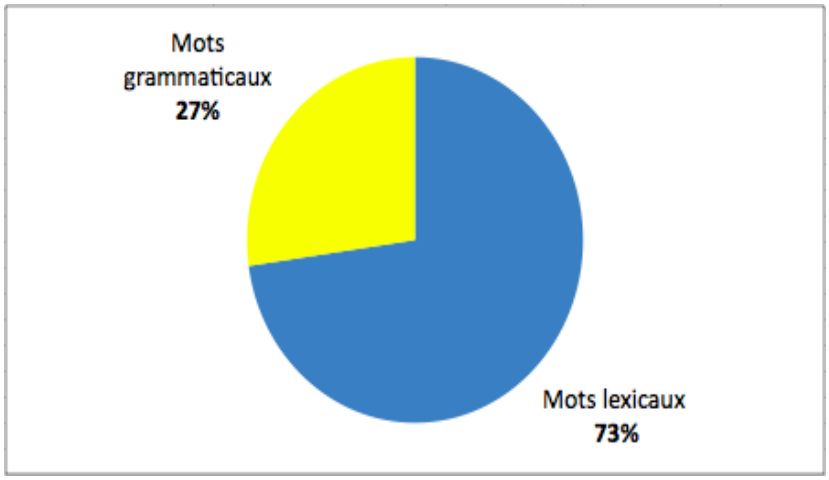


FIG. 7.20 – Classes de mots touchées par l'amorce.

Mots lexicaux	Effectif absolu	Effectif relatif
Nom	13	54%
Verbe conjugué	5	21%
Adverbe	3	13%
Adjectif	2	8%
Verbe participe passé	1	4%
Total	**24**	**100%**

TAB. 7.6 – Sous-catégorisation pour les mots lexicaux

Mots grammaticaux	Effectif absolu	Effectif relatif
Pronom démonstratif	4	44%
Pronom personnel clitique	2	22%
Pronom possessif	1	11%
Article	1	11%
Préposition	1	11%
Total	**9**	**100%**

TAB. 7.7 – Sous-catégorisation pour les mots grammaticaux

– Amorces portant sur les mots lexicaux :

nous **a-** *nous* **sommes** *vingt-deux permanents* (Verbe)

le **cli-** *la* **relation** *avec le client* (Nom)

*l'***au-** *l'***autre peut faire** (Adjectif)

et **p-** *et* **puis** *euh* (Adverbe)

– Amorces portant sur les mots grammaticaux :

c- cette *intervention* (Pronom démonstratif)

j- j'*ai eu des difficultés à répondre* (Pronom personnel)

d- *avec le client* **d'***avoir euh une relation intéressante* (Préposition)

Inachèvements

Catégories morpho-syntaxiques de l'inachèvement

L'inachèvement est un phénomène peu abordé dans les descriptions linguistiques et la définition qui en est faite reste vague. Il semble donc primordial de décrire ce phénomène à partir des exemples attestés dans notre corpus. Nous recensons 42 cas d'inachèvements, ce qui représente environ 14% des disfluences.

Rappelons qu'un énoncé inachevé est un énoncé auquel il manque un ou plusieurs éléments pour qu'il soit grammaticalement bien formé et interprétable sémantiquement. Nous avions exposé dans la partie précédente (cf. 2) les trois possibilités d'interprétations de l'inachèvement :

– le locuteur cherche ses mots
– changement de construction
– construction abandonnée mais reprise plus loin

Plus concrètement, on observe principalement dans le corpus un type bien précis d'inachèvements : les énoncés auxquels il manque un ou plusieurs constituants.

Par exemple, les énoncés qui suivent sont considérés comme inachevés car ils nécessitent une complétion après le pronom clitique qui annonce un syntagme verbal qui ne se produit pas :

> *mais moi* **je** *c'était juste l'année en plus où la majorité venait de passer à dix-huit ans*
>
> *euh dans lequel* **nous** *donc tous les agents qui qui font des contrôles sont sont de service investigation*

Il s'agit alors d'observer quelles catégories morpho-syntaxiques sont touchées par l'inachèvement. Le tableau suivant et le graphique associé présentent les différentes

catégories concernées par ce phénomène :

	Effectif absolu	Effectif relatif
Verbe conjugué	12	28,57%
Adverbe	6	14,29%
Article	6	14,29%
Pronom personnel clitique	6	14,29%
Préposition	3	7,14%
Nom	2	4,76%
Préposition + article	2	4,76%
Conjonction	1	2,38%
Pronom démonstratif	1	2,38%
Pronom indéfini	1	2,38%
Pronom relatif	1	2,38%
Verbe participe passé	1	2,38%
Total	**42**	**100%**

TAB. 7.8 – Catégories morpho-syntaxiques de l'inachèvement

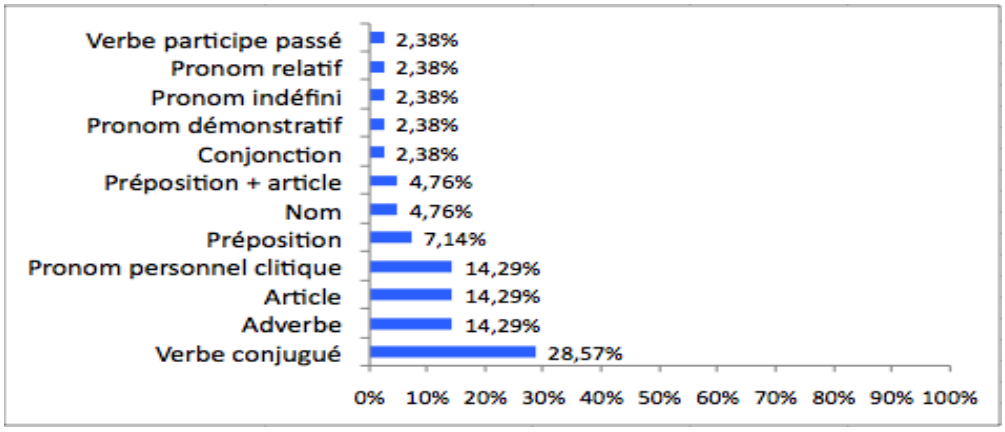


FIG. 7.21 – Répartition des catégories touchées par l'inachèvement

La première observation que nous pouvons faire, est que les constructions verbales sont dominantes dans l'inachèvement. En effet, dans plus de 28% des cas, l'inachèvement est marqué par une construction verbale incomplète comme dans les

exemples qui suivent :

vous savez que on **a** *c'est comme ça petit à petit en travaillant qu'on s'est aperçu*

*mais bon c'***est** *son CAP ne l'a p- elle ne l'a pas eu*

euh bon il y **a** *le meilleur moyen de communiquer*

On trouve un pourcentage identique (14,29%) d'inachèvements qui portent sur des adverbe, article ou pronoms clitiques :

c'est **quand-même** *euh on on a des des bons clients* (Adverbe)

je travaille dans **le** *enfin je suis née de parents de commerçants* (Article)

Après examen des précédents phénomènes de disfluences et au regard des résultats obtenus pour l'inachèvement, nous pourrions émettre l'hypothèse suivante : lorsqu'une disfluence est produite, soit elle touche un mot grammatical et on observe fréquemment une répétition du mot à l'identique (répétition) ou une correction (autocorrection), soit elle porte sur un mot lexical et il s'agit alors le plus souvent d'un mot amorcé ou d'un syntagme inachevé. Il reste bien évidemment à vérifier cette hypothèse à partir d'un plus grand volume de données afin d'infirmer ou de confirmer celle-ci.

7.4 Conclusion

Notre étude théorique et expérimentale des disfluences a montré que celles-ci présentent des régularités sur lesquelles une analyse automatique peut utilement se reposer. En effet, si les disfluences ne sont pas des phénomènes irréguliers tel qu'on pourrait le penser au premier abord, le degré de régularité varie cependant d'un phénomène à l'autre (les inachèvements par exemple sont moins réguliers que les répétitions). Ce type de remarque nous permettra ainsi d'orienter nos techniques de traitement automatique mises en IJuvre par la suite.

Cette partie a également été l'occasion de présenter notre vision théorique de la structure des disfluences. La constitution de la banque de données d'exemples dans cette représentation a permis de mettre en évidence une complexité structurelle et une certaine logique de production des énoncés oraux. Elle permet de plus de disposer d'une liste non exhaustive mais rigoureuse des phénomènes étudiés, guide précieux et nécessaire pour la conception d'un module robuste de traitement automatique de la langue parlée. Elle pourra également fournir une base de comparaison avec d'autres corpus oraux français ou étrangers.

Enfin, nous avons largement mis en avant la notion de rembobinage syntagmatique, présente sur plusieurs exemples. Cette caractéristique récurrente des disfluences, conjuguée aux remarques de l'analyse linguistique menée dans cette partie, offre une piste de travail pour l'implémentation, par l'analyse des productions orales à l'aide de la notion de « chunks ».

Troisième partie

Automatisation

Chapitre 8

Rappels de travaux antérieurs : premières expérimentations

8.1 Introduction

Après avoir dressé un large panorama des propriétés du langage oral, des différents formalismes pouvant être utilisés pour sa représentation, et des différentes approches dans le domaine de l'analyse linguistique du langage parlé, nous proposons dans cette partie de rendre compte des traitements effectués pour mener une tâche d'analyse syntaxique de surface sur notre corpus de travail. Pour ce faire nous revenons brièvement sur les résultats obtenus lors d'une première étude que nous avions réalisée durant nos travaux de master.

La motivation première du début de nos travaux est justifiée par l'idée selon laquelle les disfluences perturbent l'analyse syntaxique des productions orales (idée également défendue dans les travaux de [Ferreira *et al.*, 2004] notamment), ralentissant ainsi les avancées en terme de développement logiciel dans le domaine des technologies vocales (reconnaissance automatique de la parole pour l'aide aux personnes handicapées, pour l'apprentissage des langues, etc.).

En effet, la plupart de systèmes précédemment décrits amènent à une opération de « filtrage » des disfluences consistant à mettre en mémoire le corpus d'entrée,

supprimer les disfluences, puis construire les structures syntaxiques des énoncés où les disfluences ont été effacées. De fait, la question posant le problème de la prise en compte des disfluences sans les supprimer n'a reçu que peu d'attention de la part des chercheurs.

Nous commençons donc par rappeler les étapes réalisées dans le cadre d'une étude précédant ce travail de thèse, avant de présenter en détail les modifications et évolutions apportées depuis.

8.2 Rappel de travaux antérieurs

A l'occasion de travaux antérieurs à ceux menés dans cette étude [Bove, 2005], nous avons expérimenté le traitement automatique des disfluences au sein d'un système existant initialement prévu pour l'écrit (TiLT[1]), à partir de corpus de parole simulé (corpus *Vocalia Bourse* et *Pharmacie*[2]). Ce travail a essentiellement consisté à mettre au point des règles syntaxiques adaptées, en vue de la réalisation d'un prototype de traduction parole-parole ([Bove *et al.*, 2006]). Nous rappelons ci-après les résultats de cette expérimentation.

8.2.1 Implémentation expérimentale de règles grammaticales en dépendance

L'application TiLT est exécutée en appelant un profil (*i.e* un fichier regroupant l'ensemble des fichiers spécifiques à utiliser) adapté au traitement souhaité. Par exemple, si l'on souhaite traiter un texte anglais, le profil utilisé appellera des données de segmentation, des lexiques et des grammaires appropriées, aux données anglaises. De la même manière pour l'oral, il a fallu définir un profil adapté pour la prise en compte les phénomènes de disfluences par l'analyseur. Ce profil a supposé la création de fichiers adaptés au traitement de l'oral et nous a donc amené à

[1]Voir [Bove, 2005] pour les descriptions détaillées
[2]Voir [Bove, 2005] pour les descriptions détaillées

nettoyer le corpus (pour l'exploiter plus facilement), manipuler des lexiques, mettre en place des règles de grammaires, etc.

Aussi, à partir d'une étude sur corpus, nous avons montré comment des modifications du lexique et de la grammaire ont permis de traiter les cas les plus simples (pauses remplies, répétitions de mots isolés, etc.). D'autres cas plus complexes comme répétitions et autocorrections de syntagmes ont nécessité la mise au point d'un mécanisme de contrôle sémantique permettant de limiter la combinatoire mais sans solutions entièrement satisfaisantes. Cette étude a mis également en évidence la difficulté de traitement de certaines unités plus complexes : amorces, inachèvements, autocorrections complexes et la nécessité d'observer plus finement ces phénomènes afin de dégager des régularités qui faciliteraient leur traitement.

Notons que le corpus sur lequel nous avons réalisé cette étude est constitué de deux corpus issus de projets de France Télécom Division R&D : *Vocalia Dialogue Bourse* et *Pharmacie*. Le premier projet concerne le développement d'un système de dialogue homme-machine destiné à simuler des transactions boursières ; le second vise la mise au point d'un prototype de traduction parole-parole pour des interactions client/vendeur dans le domaine pharmaceutique. Le corpus Vocalia comprend environ 20 000 mots ; le corpus Pharmacie, toujours en cours d'élaboration, comprend à l'heure actuelle plus de 4 500 mots.

Voici deux exemples simples de règles syntaxiques mises au point pour traiter certains cas de disfluences :

– Répétition de déterminants (articles définis) :

> RègleAttachement DETDEF_ORAL REPETITION_ORAL
>
> Schéma GN-D >/> GN-D
>
> Exemple : *en descendant de l'autobus euh ma femme a mis* **le : le** *pied par terre elle a fait un faux mouvement +*

– Répétition de prépositions :

> RègleAttachement PREP_ORAL REPETITION_ORAL
>
> Schéma GP-S >/> GP-S
>
> Exemple : *je vais vous l'indiquer c'est le Sillon +* **sur** *euh* **sur** *le Sillon à Saint-Malo*

Par exemple, la première règle concerne le phénomène de répétition (*REPETITION_ ORAL*), et porte sur la répétition de déterminants définis (*DETDEF_ ORAL*). Ainsi lorsque deux déterminants définis (*GN-D*) se succèdent, le second va « consommer » (>/>) le premier pour ne former qu'un seul arbre.

Les figures 8.1 et 8.2 donnent respectivement un exemple de représentation syntaxique arborescente avant et après mise en place des règles pour la phrase *ben je je viens vous voir* (mêlant marqueurs et répétition « simple »). Notons que la figure 8.1 comporte trois arbres partiels ; ceux-ci se retrouvent correctement regroupés une fois les règles mises en place (8.2).

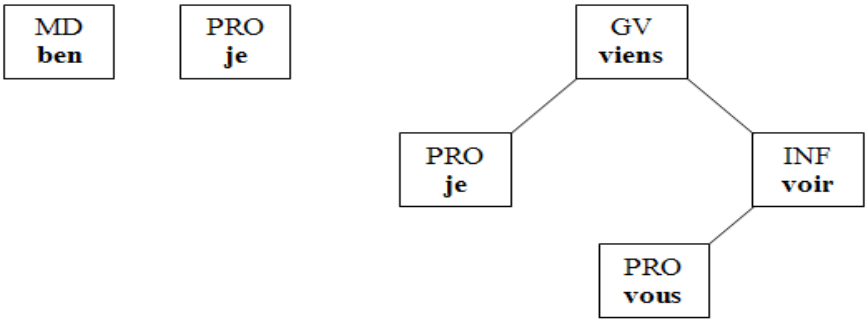


FIG. 8.1 – Représentation syntaxique **avant** mise en place des règles

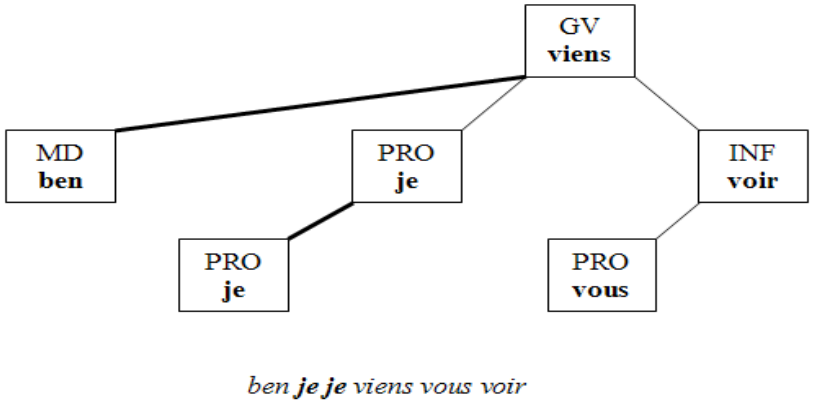


FIG. 8.2 – Représentation syntaxique **après** mise en place des règles

L'exemple ci-dessus est un cas de répétition **simple** : l'empan de la répétition est un mot unique, presque exclusivement un mot-outil (article, préposition, etc.) monosyllabique introducteur de *chunk* ([Henry *et al.*, 2004]). Ce type de répétition produit un *chunk* inachevé (dans l'exemple ci-dessus, le premier *je*). Comme nous l'avions expliqué précédemment (cf. 2.3.3), il témoigne généralement de la difficulté de mise en place du lexique (ici le choix du verbe).

Notons que nous avons considéré les répétitions simples et les autocorrections immédiates selon le même type de traitement étant donnée la similarité fonctionnelle assez forte des deux phénomènes.

Un autre type de répétition concerne des unités **complexes**, concernant des syntagmes entiers :

a) *euh je voudrais connaître* **l'analyse technique l'analyse technique** *sur le CAC-40*

De tels cas sont plus complexes à traiter, particulièrement lorsqu'ils font intervenir une autocorrection :

b) *je voudrais savoir la* **la valeur la cotation** *de l'action euh Aventis*

Il faut être capable de repérer que certains syntagmes doivent être regroupés et que la recherche des fonctions syntaxiques qu'ils remplissent doit tenir compte de ce regroupement. Dans l'exemple ci-dessus, l'arbre représentant le syntagme *la valeur* est attaché à l'arbre représentant *la cotation*, qui remplit la fonction d'objet direct du verbe *savoir*.

On peut ajouter dans la grammaire des règles permettant d'attacher un groupe (nominal, verbal, ...) à un autre groupe du même type, comme il a été fait pour les introducteurs de *chunks*. Toutefois, des suites consécutives de déterminants ou prépositions ne permettent guère d'autres interprétations que celle d'une répétition, alors que pour des syntagmes complexes, ce type de règle est beaucoup trop tolérant et serait susceptible de générer une explosion combinatoire. On ne peut en effet pas imposer comme contrainte que la source et la cible de la répétition soient de structure absolument identique. L'exemple ci-dessous (variante du b) est tout à fait légitime :

c) * *je voudrais savoir la* **la valeur le volume d'échanges** *de l'action euh Aventis*

Nous avons commencé à intégrer des critères de taille et de structure destinés à limiter la combinatoire. En effet, il semble, en première approximation, que des arbres très profonds ou très larges aient peu de chances d'être la source d'auto-corrections. Cette hypothèse est cohérente avec un modèle psychologique qui attribue à un mécanisme de self-monitoring la production des auto-corrections ([Levelt, 1989]) : interruption et mise en place d'une auto-correction demandent un délai relativement bref au locuteur, incompatible avec la durée de structures complexes. Grâce à un outil de gestion des grammaires développé dans l'équipe Langues Naturelles, nous avons par exemple pu modifier automatiquement les règles de grammaire exploitées par l'analyseur de manière à ce que les arbres représentant des groupes verbaux avec compléments ne puissent plus être utilisés par des règles de reprise

de groupes verbaux.

8.2.2 Ajout du contrôle sémantique

L'analyseur TiLT est basé sur le formalisme des grammaires de dépendance, tout nIJud d'un arbre syntaxique correspond donc à un mot (simple ou composé) de l'énoncé initial. L'analyse syntaxique s'effectue après une phase d'analyse lexicale, qui attribue à tout mot de l'énoncé un ensemble d'interprétations lexicales, en fonction du lexique utilisé. La notion d'interprétation lexicale correspond à un ensemble d'informations telles que lemme, description morpho-syntaxique, mais aussi « unité sémantique ».

Pour limiter l'attachement d'arbres correspondant à des syntagmes nominaux ou verbaux, nous avons introduit une phase de contrôle basé sur la sémantique. Par exemple, dans l'énoncé (b), c'est la proximité des sens de « valeur » et de « cotation » (dans un contexte applicatif donné) qui va permettre de valider l'hypothèse que *la cotation* est une reprise de *la valeur* :

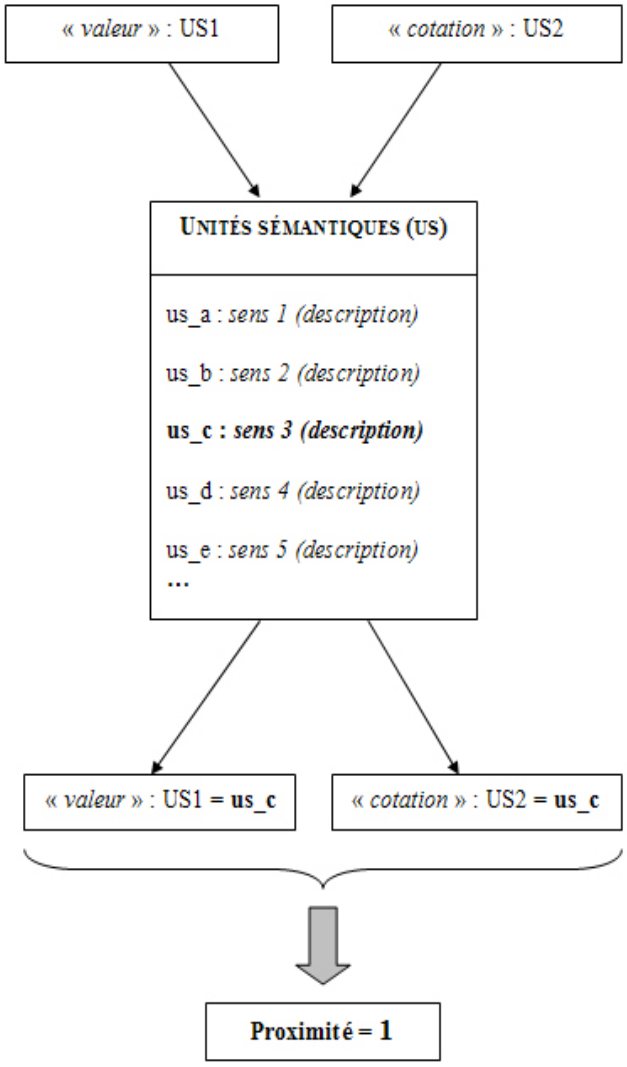


FIG. 8.3 – Mécanisme (simplifié) de contrôle sémantique

L'unité sémantique (US) permet de représenter le sens de l'interprétation. Les unités sémantiques sont organisées dans un réseau sémantique dans lequel elles sont reliées par des relations d'hyperonymie, d'appartenance, etc. Il est alors possible de définir la notion de proximité sémantique entre deux unités sémantiques US1 et US2. Cette proximité vaut 1 si les unités sémantiques sont identiques, et elle est calculée en fonction de paramètres tels que nombre, sens et type des relations à parcourir pour passer de l'une à l'autre dans le réseau.

Le contrôle sémantique n'est cependant pas suffisant pour analyser d'autres phénomènes tels que les amorces ou encore les inachèvements. :

a) *oui le* **d-** *je veux le détail*

Le traitement des amorces implique des modifications au niveau du système de reconnaissance vocale. En effet, pour l'instant, celui-ci fournit le mot (évidemment erroné) le plus proche dans son lexique en fonction du modèle de langage sous-jacent, et le système TiLT n'a aucun moyen de savoir qu'il s'agissait d'une amorce. Ce problème constitue un défi pour tous les systèmes de reconnaissance à l'heure actuelle, car le nombre d'amorces potentielles est très grand (pratiquement chaque suite de phonèmes commençant chaque mot plein), et la fréquence de chaque amorce trop faible pour pouvoir entraîner efficacement un modèle de langage statistique. Il est probable que d'autres critères (prosodiques notamment) devraient être pris en compte.

Le système n'est également pas en mesure pour l'instant de traiter les inachèvements, car une modification des règles qui permettrait à la plupart des syntagmes de rester incomplets aboutit à une explosion combinatoire extrêmement rapide.

b) **qu'est-ce que vous avez d'efficace contre** + *bon je ne sais pas si ce que j'ai c'est vraiment la grippe mais* (...)

Face à ces phénomènes, l'analyseur ne peut que proposer des arbres syntaxiques

partiels recouvrant l'énoncé, mais ne permettant pas d'aboutir à une solution globale pour celui-ci.

Aussi, malgré cette première expérience de traitement automatique des disfluences, il reste encore beaucoup de travail à faire en ce qui concerne celles dont le mode de fonctionnement syntaxique est éminemment plus complexe. Ceci dénote clairement l'un des problèmes majeur qui est que les phénomènes de productions constituent une classe très hétérogène. Les pauses remplies, par exemple, correspondent à un ensemble relativement clos ; tandis que les autres phénomènes tels que répétitions ou autocorrections de syntagmes, inachèvements, etc. sont beaucoup plus délicats à appréhender ne serait-ce qu'en terme statistiques. Ils n'ont en effet aucune forme lexicale fiable, et ils varient beaucoup en taille : de quelques mots à des énoncés plus longs.

8.3 Conclusion

Bien que fournissant de nombreuses pistes d'analyses, notre première approche du problème de prise en compte des disfluences présentait un certain nombre de limites et ce pour plusieurs raisons.

D'une part, les corpus utilisés n'étaient pas parfaitement spontanés dans la mesure où ceux-ci ont été constitués dans le cadre d'une simulation d'interaction dans le cadre d'une expérience de faux magicien d'Oz[3] (homme-machine pour *Vocalia Bourse*, et homme-homme pour le corpus *Pharmacie*). Ce type de données n'a donc pas le même degré de « spontanéité » que le CRFP que nous utilisons dans le cadre de notre étude. Les contenus diffèrent ainsi en terme de complexité des disfluences (dans le CRFP les répétitions et autocorrections de syntagmes sont

[3]Le terme magicien d'Oz désigne une expérience de simulation de dialogue homme-machine au cours de laquelle un individu (le compère) joue le rôle du système informatique. Les expériences de vrai magicien d'Oz sont celles où le locuteur n'est pas informé de l'existence du compère, et celles de faux magicien celles où il es précisé au locuteur q'un être humain remplace l'ordinateur.

plus nombreuses, les imbrications de disfluences plus longues, etc.)

D'autre part, ce travail nécessite une étude préalable suffisamment approfondie des patrons syntaxiques des chunks disfluents qui n'a été faite qu'ultérieurement ([Piu, 2006]). L'analyse et l'implémentation menées s'avéraient en ce sens quelque peu prématurées. De plus, cette étude antérieure a mis en avant une autre piste de traitement que nous n'avions pas exploité : travailler sur les segments disfluents au niveau des chunks avant de passer aux relations en dépendances des énoncés disfluents. L'expérience d'implémentation de règles syntaxiques en dépendance s'avère effectivement trop avancée dès lors que les étapes précédentes, d'étiquetage et d'analyse syntaxique partielle ne sont pas suffisamment maîtrisées. Dans cette perspective, nous pensons donc qu'il est préférable – dans l'état actuel des techniques de traitement de l'oral– de se consacrer plus spécifiquement à une analyse syntaxique de surface.

Chapitre 9

Analyse syntaxique partielle pour l'oral

9.1 Introduction

Le traitement automatique des disfluences doit être réalisé à différents niveaux en prenant garde de ne pas précipiter chacune des phases nécessaires à celui-ci. Dans le cadre de notre travail, nous ne pouvons légitimement pas prétendre réaliser l'intégralité d'une chaîne de traitement automatique (prosodique, syntaxique, sémantique, etc.) sans perdre inévitablement en qualité d'analyse.

Nous choisissons ici de nous consacrer à l'analyse syntaxique de surface (ou *shallow parsing*) via les étapes d'étiquetage puis de regroupement en chunks de notre corpus (y compris les phénomènes de disfluences) tels que nous le décrivons plus loin. Dans la littérature, plusieurs études menées sur le français (notamment par [Goulian, 2000] ; [Antoine *et al.*, 2003] ; [Bourigault, 2007], etc.) ont clairement montré à quel point la réalisation des ces étapes est incontournable. En effet, elles permettent de faciliter l'extraction de relations de dépendance, et autres traitements ultérieurs (analyse sémantique, pragmatique, etc.) qui nécessitent des résultats préliminaires robustes.

L'analyse que nous proposons se veut résolument applicative. Il s'agit de confronter

ici nos hypothèses théoriques avec des données non artificielles (dans notre cas, des séquences extraites du CRFP), malgré la rareté de ces ressources dans le domaine de l'oral.

Dans cette optique, l'analyse syntaxique consiste à associer automatiquement à la chaîne découpée en unités, une représentation des groupements structurels et des relations fonctionnelles existants entre des unités.

9.2 Chaîne de traitement axée sur les disfluences

Les traitements réalisés dans cette étude visent à obtenir une analyse syntaxique dite « robuste » des séquences disfluentes, de façon à pouvoir, à terme, s'intégrer dans une chaîne de traitement plus large adaptée aux corpus oraux.

Dans une telle approche, on ne cherche plus à analyser entièrement un énoncé, mais à en faire une analyse partielle, laissant de côté les phénomènes les plus ambigus, tels que les rattachements prépositionnels. L'idée qui sous-tend ce type d'approche est que l'analyse complète n'est pas toujours nécessaire dans certaines applications du traitement automatique de la langue, telle que l'extraction d'information, par exemple. De tels systèmes sont généralement organisés en « chaînes de traitement », dans lesquels plusieurs modules (module de découpage en mots, module d'étiquetage morpho-syntaxique, module d'analyse syntaxique partielle) sont organisés de manière séquentielle.

Il s'agit précisément du schéma retenu ici. En effet, ce type d'organisation permet de décomposer le problème global en une série de problèmes plus simples.

Néanmoins, elle possède un inconvénient qui provient du fait que chaque module n'a qu'une vision locale et commet des erreurs qui auraient pu être évitées en prenant en compte les modules suivants. Un exemple typique est celui de l'étiquetage morpho-syntaxique qui effectue des erreurs d'étiquetage propagées en cascade qui

auraient pu être évitées au niveau de l'analyse syntaxique partielle. Cette contrainte nécessite de prendre en compte les éventuelles erreurs par le biais de systèmes de désambiguïsation par exemple.

A l'instar de [Boufaden *et al.*, 1998], ou plus récemment [Bourigault, 2007], nous optons ainsi pour une approche modulaire, de façon à pouvoir intervenir sur chaque étape de façon indépendante. En effet, le traitement que nous proposons ici correspond à une analyse procédurale « en cascade » (cf. [Bourigault, 2007]). De cette façon les séquences disfluentes du corpus sont traitées via plusieurs modules successifs. L'entrée d'un module étant la sortie du module précédent. Les diverses étapes nécessaires à la chaîne complète (classiques en TAL) sont schématisées sur la figure suivante :

FIG. 9.1 – Étapes de traitements classiques en TAL

L'idée de base consiste donc à élaborer un système constitué de plusieurs composants autonomes (en terme d'implémentation) qui interagissent ensuite au niveau de l'analyse du corpus.

Concernant la première étape, nous utilisons comme point de départ la sous-partie du CRFP (plus de 8500 mots) annotée prosodiquement de façon semi-automatique ([Campione, 2001]). Le postulat de départ est que ce type de corpus constituerait la sortie idéale de l'annotation prosodique. Partant de ce point, nous nous concentrons essentiellement sur les phases d'étiquetage et d'analyse syntaxique partielle que nous développons dans les points suivants. Les étapes d'extraction de dépendance et d'analyse sémantique constituent les étapes ultérieures idéales pour terminer

l'ensemble de la chaîne ; celles-ci représentent un travail à part entière, qui dépasse le champ d'investigation mené dans notre étude.

L'analyse se déroule en trois passes principales, elles même divisées en sous-tâches :

- Étiquetage morpho-syntaxique
  - Étiquetage
  - Apprentissage
  - Post-étiquetage
- Détection préliminaire « brute » des phénomènes de disfluences
  - Détection des pauses remplies
  - Détection des répétitions et autocorrections
- Regroupement en chunks
  - Segmentation des séquences non-disfluentes
  - Segmentation des séquences disfluentes

Les modules sont développés dans les langages Perl et Bash, selon une méthode empirique faisant appel aux connaissances grammaticales tout en réalisant de nombreux tests sur le corpus de travail et ce, sans faire référence à une théorie linguistique particulière. Pour mettre au point la stratégie d'analyse, nous avons opté pour la confrontation systématique au corpus par essai/erreur et l'alternance entre observation et implémentation. Les tests menés sur le corpus de développement permettent d'améliorer la couverture et la précision des algorithmes de reconnaissance. Cette approche se justifie par le fait que les énoncés traités présentent des configurations syntaxico-discursives assez peu décrites dans les théories linguistiques (cf. 5). Ces énoncés requièrent donc une analyse en corpus détaillée ainsi qu'un développement de procédures de traitement automatique empruntant peu aux descriptions linguistiques classiques. D'ailleurs, [Bourigault, 2007] rappelle à juste titre que " *les problèmes pratiques liés à la reconnaissance automatique de la*

structure syntaxique d'une phrase et les problèmes théoriques liés à la description syntaxique sont de deux ordres différents ". De telles descriptions syntaxiques ne fournissent en effet que peu d'indices pour la mise en place d'un système automatique.

Même si le développement de modules ne recours pas directement aux travaux des descriptions syntaxiques, cela implique cependant de connaître ces travaux. Il est en effet indispensable de pouvoir reconnaître des phénomènes syntaxiques identiques face à la multitude des configurations syntaxiques possibles. La modélisation et l'élaboration des traitements sur corpus doivent passer par l'identification correcte de ces configurations.

La démarche de développement n'est pas exclusivement guidée par l'observation du corpus. En effet, l'utilisation du corpus de développement est essentielle pour « diagnostiquer » au mieux l'inventaire des multiples configurations de surface. Mais dans le même temps, il faut être en mesure de prévoir des règles pour les configurations non attestées dans le corpus de développement mais qui ne signifie pas pour autant qu'elles n'existent pas.

Le développement des traitements ne peut donc se faire sans avoir également recours à la connaissance de la grammaire de la langue. C'est elle qui permet, après observation du corpus, de dépasser les configurations relevées, pour anticiper des règles de reconnaissance dont la couverture dépassera à son tour les cas observés. La figure suivante donne l'architecture globale simplifiée du prototype d'analyseur, dont chacune des « briques » de traitement sera ensuite détaillée.

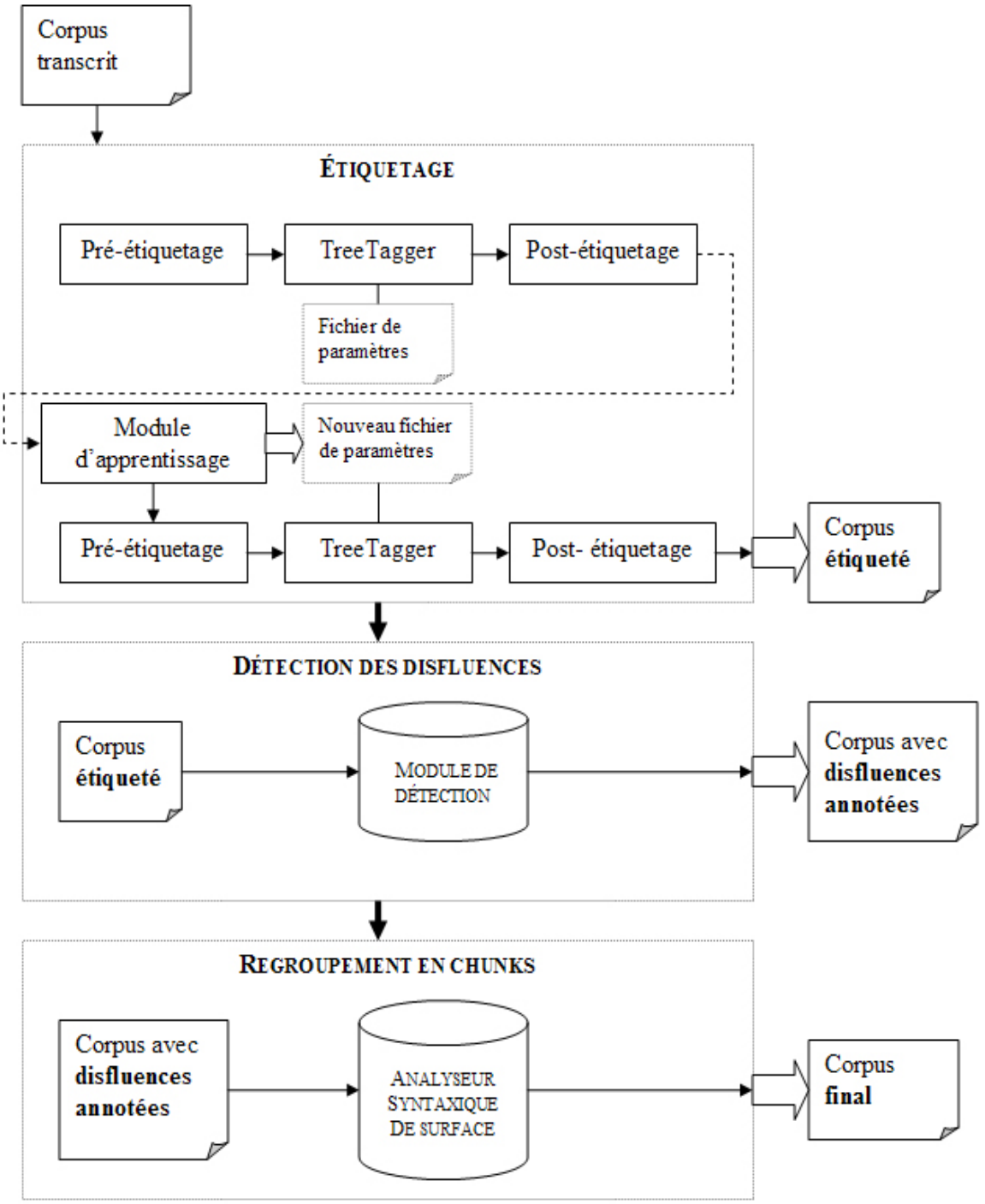


FIG. 9.2 – Architecture générale de l'analyseur

9.3 Mise en place de la phase d'étiquetage

9.3.1 Choix d'un étiqueteur

L'annotation morpho-syntaxique automatique (ou *tagging*) est une phase indispensable pour de nombreuses applications en TAL. Il s'agit d'automatiser autant que faire se peut l'ajout d'informations linguistiques au corpus brut.

[Véronis, 1998] rappelle à cet égard que :

> " *Quelle que soit la technologie sur laquelle ils sont basés, les étiqueteurs atteignent à l'heure actuelle des performances très satisfaisantes. Les résultats publiés sont couramment de l'ordre de 95% d'étiquettes correctes, et l'on trouve même des chiffres supérieurs* ".

Les performances restent toutefois difficiles à quantifier. D'une part la finesse du jeu d'étiquette varie d'un système à l'autre. On retrouve néanmoins une régularité dans la forme des étiquettes qui contiennent la partie du discours du mot concerné, accompagnée d'un certain nombre d'informations morphosyntaxiques (genre, nombre, temps, personne, etc.). D'autre part, même lorsque des étiquettes sont identiques, Véronis (1998) explique que l'extension de ces étiquettes peut être très différente d'un système à l'autre. C'est particulièrement le cas pour les catégories difficiles (adverbes, adjectifs « indéfinis », etc.).

De plus, il convient de noter qu'une grande partie des mots (environ 60%) ne sont pas ambigus et un simple accès à un lexique produit donc un étiquetage en grande partie correct ([Véronis, 1998]).

Les méthodes, pourtant sophistiquées, utilisées dans les étiqueteurs n'apportent une amélioration que de 50% environ par rapport à cet étiquetage trivial. [Pelurson, 2006] précise que plusieurs stratégies sont proposées pour annoter automatiquement les mots par des étiquettes morphosyntaxiques. De nombreux outils sont fondés sur

des systèmes à base de règles ([Brill, 1992]). D'autres implémentent des méthodes probabilistes [Schmid, 1994], [Church, 1988], [Cutting *et al.*, 1992], [Kempe, 1993] ou encore inspirées des réseaux de neurones [Federici et Pirrelli, 1994].

Leur technologie est le plus souvent basée sur l'examen de contextes locaux (tels que les trigrammes de mots), qui résolvent mal les cas qui demanderaient une analyse syntaxique plus globale, et en particulier la reconnaissance des dépendances à distance dans la phrase. [Véronis, 1998] rajoute que dans de nombreux étiqueteurs la théorie linguistique sous-jacente est très rustique, car fortement inspirée de la grammaire scolaire, et de nombreux cas d'étiquetage considérés comme « corrects » semblent assez critiquables sur le plan purement linguistique.

On peut toutefois considérer, malgré ces réserves, que les systèmes d'étiquetage en parties du discours sont opérationnels, notamment dans plusieurs applications liées au langage comme la recherche d'informations. Ils permettent d'avoir une information syntaxique suffisante pouvant être utile dans le traitement d'un énoncé. Partant de ces considérations, notre choix s'est naturellement porté sur l'étiqueteur TreeTagger qui nous a semblé, après expérimentation, particulièrement robuste, rapide et disponible facilement pour la mise en IJuvre de notre prototype.

9.3.2 TreeTagger

Développé au sein du projet TC (*Textcorpora and Erschliessungswerkzeuge / textual corpora and tools for their exploration*) à l'institut de linguistique computationnelle de l'université de Stuttgart ([Schmid, 1994]), TreeTagger[1] est un système d'annotation de catégories morphosyntaxiques permettant d'étiqueter des textes dans différentes langues (anglais, français, allemand, italien, grec, et ancien français). Il est possible d'adapter l'étiqueteur à d'autres langues, à condition de disposer d'un lexique et d'un corpus manuellement annoté. Le programme TreeTagger

[1] http ://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

entraîné pour le français est en charge de la segmentation et de l'étiquetage grammatical des mots dans les corpus.

Cet étiqueteur stochastique, publiquement disponible à des fins de recherche, est indépendant de la langue. Il présente l'intérêt d'être ouvert, en ce sens qu'il est possible d'effectuer en amont, à sa place, une partie du travail de segmentation en mots et d'étiquetage. C'est ce que nous avons fait pour l'adapter à notre corpus comme nous l'expliquerons plus loin.

Malgré ce que peut laisser penser son nom, TreeTagger ne fait pas référence à une éventuelle représentation arborée de la syntaxe. Ceci tient uniquement au fait que cet étiqueteur choisit les étiquettes qu'il applique aux mots à l'aide un arbre de décision. En effet, à la différence d'autres méthodes d'étiquetage probabiliste, qui ont des difficultés à estimer exactement de petites probabilités à partir d'une quantité limitée de données pour l'apprentissage, TreeTagger évite le problème des données rares en utilisant un arbre à décisions binaires, qui détermine la taille appropriée du contexte nécessaire pour estimer les probabilités de transition. Les contextes possibles sont non seulement des trigrammes, bigrammes et unigrammes, mais aussi d'autres types de contextes.

Le système comprend également un module de segmentation ; lors de l'étiquetage, le lemme et certaines informations morpho-flexionnelles (temps pour les verbes, type du déterminant, etc.) sont calculées. Les résultats de l'étiquetage sont affichés sous forme de triplets correspondant respectivement à la forme fléchie, la catégorie morpho-syntaxique et le lemme.

Il est à noter que lorsque l'étiqueteur arrive sur des unités inconnues du catégoriseur, il ne se risque pas à reconstruire le lemme mais attribue alors une valeur arbitraire `<unknown>` en lieu et place de celui-ci.

Il s'agit donc d'un outil qui calcule les catégories grammaticales, les informations

ce	PRO:DEM	ce
que	KON	que
nous	PRO:PER	nous
disions	VER:subp	dire
la	DET:ART	le
semaine	NOM	semaine
dernière	ADJ	dernier
(...)	(...)	(...)
dans	PRP	dans
la	DET:ART	le
comedia	NOM	<unknown>
lopesque	ADJ	<unknown>
en	PRP	en
général	ADJ	général

FIG. 9.3 – Exemple de corpus étiqueté par TreeTagger

morphosyntaxiques (selon les langues), mais aucune « structure syntaxique », ni « complète » ni « partielle », que ce soit en terme de chunking ou de relations de dépendance ([Aubry *et al.*, 2007]).

TreeTagger se rapproche des taggers n-grammes traditionnels ([Church, 1988], [Kempe, 1993]) mais utilise un arbre de décision binaire (évoqué ci-dessus) obtenu par entraînement, illustré par cet exemple donné par [Schmid, 1994] dans sa présentation de TreeTagger :

Le choix se fait à partir des probabilités, en maximisant la probabilité que l'étiquette « tag » apparaisse à la position i sachant un historique.

Par exemple la probabilité pour l'exemple de la figure 9.4 peut se formuler de la façon suivante :

$$P(tag_m = Nom | tag(-2) = dterminant, tag(-1) = adjectif) = 0,7$$

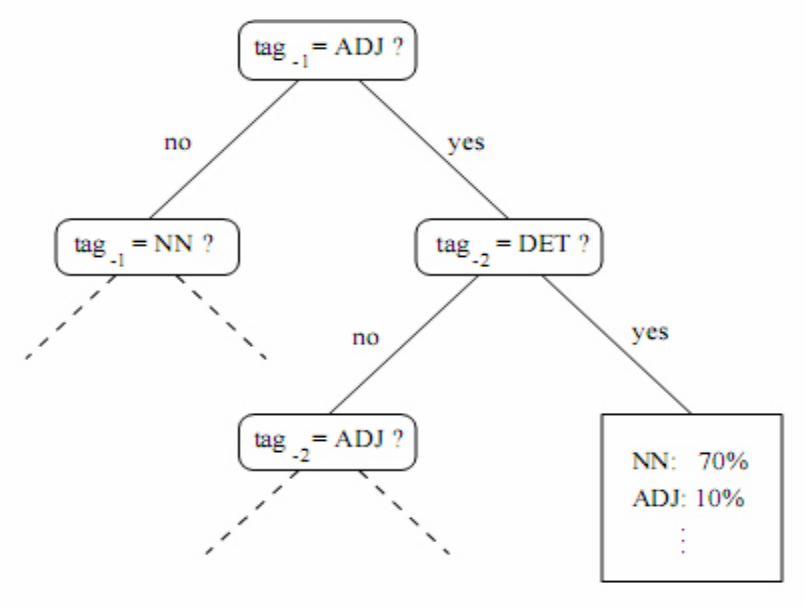


FIG. 9.4 – Exemple d'arbre de décision utilisé par TreeTagger

En d'autres termes, étant donné un mot m, dont les deux mots précédents sont respectivement un déterminant suivi d'un adjectif, m a une probabilité de 70% d'être un Nom et d'être ainsi étiqueté comme tel. Exemple :

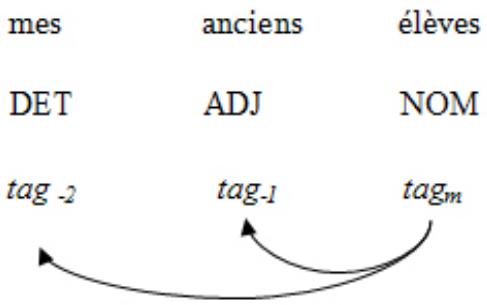


FIG. 9.5 – Exemple d'examen de contexte par TreeTagger

L'arbre de décision a pour feuilles des listes de probabilités. L'entraînement comme l'analyse sont réalisés le plus souvent à partir de trigrammes de mots. Durant la

recherche d'un mot, c'est en premier lieu le lexique qui est examiné. Si le mot s'y trouve, le vecteur des probabilités d'étiquettes est renvoyé. Sinon, le TreeTagger essaie de prédire l'étiquette correcte d'après les dernières lettres du mot (probabilités de suffixe).

Le lexique implémenté dans TreeTagger contient la liste des possibilités d'étiquetage pour chaque mot, l'étiqueteur en utilise principalement deux par défaut :

- un lexique de formes pleines associées à des lemmes, contenant les deux millions d'entrées du corpus *Penn Treebank*[2]. (Les mots dont la fréquence relative était inférieure à 1% ont été supprimés : ils étaient le plus souvent dus à des erreurs d'étiquetage).

- Un lexique de suffixes organisé dans une structure arborescente. Il permet de gérer efficacement les « non mots » (*i.e* les cas `<unknown>`), et ajoute une certaine fiabilité à la robustesse de Treetagger par l'usage de la méthode probabiliste.

La recherche d'un mot dans le lexique démarre par une recherche dans le premier fichier (avec changement de la casse du mot si la recherche s'avère infructueuse avec la casse originelle) ; puis dans le second si le mot n'a pas été trouvé dans le premier.

La particularité de cette méthode réside dans le calcul de la probabilité de transition qui n'est autre que la probabilité d'une étiquette par rapport aux étiquettes précédentes.

L'étiqueteur fonctionne à partir de deux programmes : train-tree-tagger, qui génère un fichier paramètre à partir d'un lexique et d'un corpus manuellement balisé, et tree-tagger, qui prend un fichier paramètre, un fichier texte en arguments ainsi que

[2]http ://www.cis.upenn.edu/ treebank/

certaines options facultatives, et qui permet d'étiqueter les textes.

En terme de performance, TreeTagger annonce pour l'anglais écrit un taux d'étiquetage situé dans la moyenne haute des étiqueteurs : 96,34% sur le *Penn Treebank* (http ://www.cis.upenn.edu/~treebank). Ce score peut être imputable à cette utilisation à la fois d'un arbre de décision (qui rappelle les étiqueteurs « par règles ») et de méthodes statistiques ([Aubry *et al.*, 2007]).

Par ailleurs, l'étude de [Pelurson, 2006] dans le domaine de la traduction automatique a permis d'évaluer des étiqueteurs sur un corpus annoté issu du Hansard (débats parlementaires canadiens). Ce corpus est composé de 4 528 phrases, de 107 165 mots. Ces mots ont été étiquetés manuellement par le RALI[3] à l'aide d'un jeu d'étiquettes riche et détaillé (183 étiquettes en tout telles que `AdjQ-Compar` pour adjectif qualificatif de comparaison, `Verb-aux-PAST-plur-p1` pour verbe auxiliaire à la première personne du passé).

Les étiqueteurs donnent une étiquette à tous les mots du texte ; l'auteur a donc pris comme critère d'évaluation le pourcentage de mots bien étiquetés (*i.e* avec la même étiquette que celle du fichier original pour le mot). L'auteur annonce un taux d'étiquetage correct pour TreeTagger de 86.70%.

9.3.3 Modification et adaptation de TreeTagger

La chaîne de traitement de l'analyse syntaxique traditionnelle (voir figure 9.1) comprend au minimum deux modules : l'analyseur morpho-lexical, qui, partant de ressources lexicales considérées comme exhaustives, produit une liste de *tokens* (« mots » arbitraires) munis chacun d'une ou plusieurs étiquettes, suivi de l'analyseur syntaxique proprement dit, qui, pour chaque énoncé, produit zéro, une ou

[3] http ://rali.iro.umontreal.ca/

plusieurs analyses ([Vergne et Guiguet, 1998]).

Notre travail s'est justement focalisé sur ces deux étapes principales. Nous commençons par expliquer de quelle manière nous avons adapté un outil existant (TreeTagger) pour ensuite présenter la grammaire utilisée pour mener l'analyse en chunks des énoncés de notre corpus. L'idée sous-jacente étant de voir dans quelle mesure TreeTagger est capable d' « apprendre » les disfluences et d'examiner la pertinence d'une analyse en chunks des énoncés disfluents.

Adaptation de TreeTagger

Pour mener à bien le traitement des disfluences à l'aide de TreeTagger, nous avons modifié la structure initiale des fichiers sources fournis.

En effet, dans sa version standard, TreeTagger propose un script générique réalisant les étapes de segmentation, pré-étiquetage, étiquetage et post-étiquetage. Or, nous souhaitions utiliser uniquement le module d'étiquetage (dont les fichiers sources binaires ne sont pas modifiables). Aussi, nous avons divisé le programme en plusieurs sous-programmes pour ne garder que la phase d'étiquetage et personnaliser ensuite les étapes de segmentation, pré-étiquetage et post-étiquetage.

Étiquetage du corpus (avec pré- et post- étiquetage personnalisés)

Segmentation et pré-étiquetage

Dans un premier temps, le corpus donné en entrée à l'analyseur est pré-étiqueté : il est découpé en mots (*token*), et à chaque mot est associée une catégorie grammaticale (nom, verbe, adjectif, ...).

Grâce à la dissociation des différentes étapes de traitement effectuée sur TreeTagger, nous avons également ajouté un certain nombre de traitements additionnels

pour le pré-étiquetage :

- Formatage des balises d'annotation prosodiques (pauses, allongements syllabiques, intonations montantes, descendantes, etc.) et d'annotations spécifiques à la transcription (segments inaudibles, incompréhensibles, etc.) : cette considération est destinée à faire en sorte qu'elle ne soit pas prise en compte lors de l'étiquetage (l'étiquetage de celles-ci seraient forcément erroné, l'analyseur cherchant à attribuer un étiquette à toute forme rencontrée) et qu'elle ne perturbe pas ainsi l'étiquetage des tokens environnants.

- Formatage de certaines locutions de façon à ce qu'elles soient considérées comme un token à part entière et non divisées en autant de mots qui la composent (*de temps en temps*, *au jour le jour*, etc.)

- Séparation – à l'aide de règles génériques – des constructions formées de substantifs suivis de l'adverbe *là* (*moment-là*, *époque-là*, etc.), ainsi que de certaines spécificités de constructions verbales (*vas-y*, *paraît-il*, etc.) qui n'étaient pas reconnues par le tagger.

Exemple :

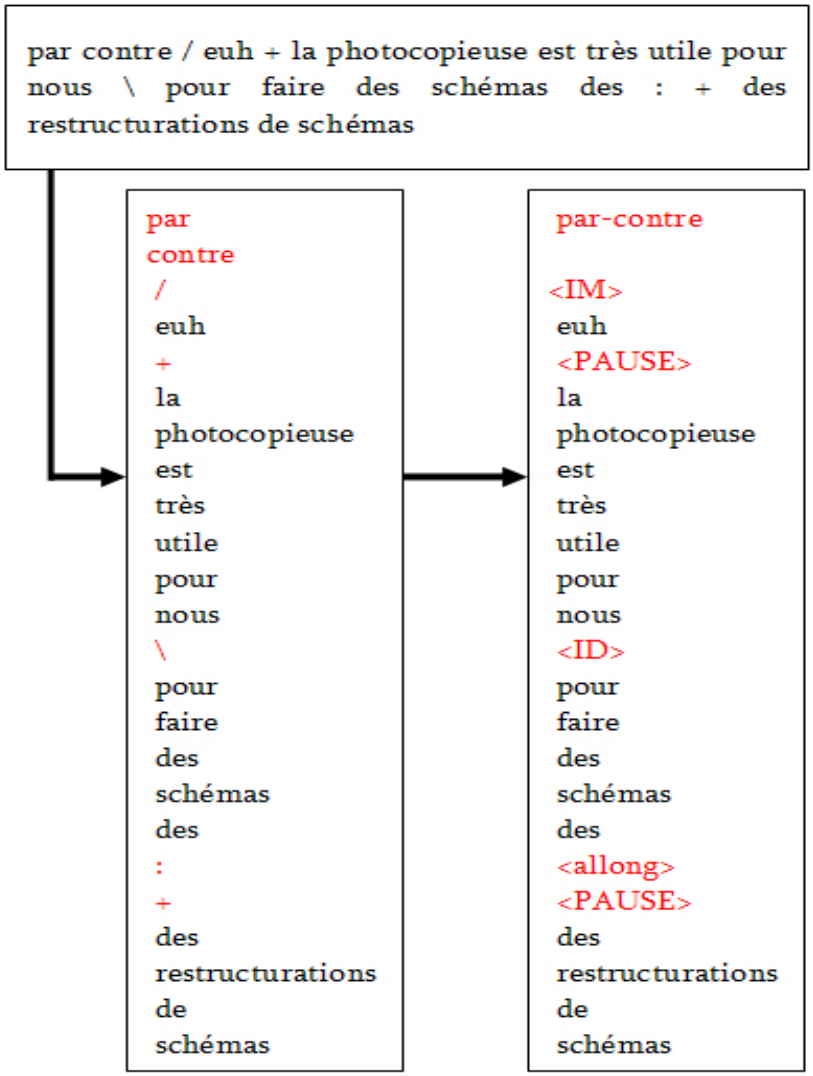


FIG. 9.6 – Extrait de corpus avant et après pré-étiquetage

Étiquetage

La phase suivante consiste à exécuter le module d'étiquetage de TreeTagger (fichier binaire non modifiable) sur la sortie de la passe précédente (pré-étiquetage). Précisons que nous appelons Treetagger avec trois options : `-token`, `-lemma` et `-sgml`, ainsi qu'un fichier de paramètres (dans notre cas celui pour le français proposé par défaut lors de l'installation de l'étiqueteur).

Nous appelons donc Treetagger à l'aide de la commande suivante :

```
tree-tagger french.par -token -lemma -sgml <corpus_préétiqueté>
```

Le fichier de paramètres pour le français a été développé par Achim Stein[4], qui ne donne pas d'évaluation quant au taux d'étiquetages corrects (mais il donne néanmoins un taux d'environ 92% pour le fichier qu'il a développé pour l'ancien français, soit 4% de moins que ce qui est annoncé pour l'anglais).

Si aucune option n'est précisée, TreeTagger imprime uniquement les catégories. Avec l'option -token, il imprime au début de chaque ligne le mot étiqueté, séparé de son étiquette par une tabulation. Avec l'option `-lemma`, il imprime le lemme si celui-ci lui est connu, et `<unknown>` dans le cas contraire (et dans les deux cas de figure à droite de la catégorie, et séparée de celle-ci par une tabulation). Rappelons que TreeTagger, du fait de sa nature probabiliste et de l'utilisation du lexique des suffixes, donne systématiquement une catégorie même si le mot ne fait pas partie de son lexique et il est donc incapable d'associer le lemme correspondant.

Enfin, l'utilisation de l'option `-sgml` permet de passer des balises au tagger, et de les récupérer intactes dans le fichier de sortie (*i.e* non suivies d'une catégorie et d'un lemme). Cela nous a permis de ne pas étiqueter les marques de prosodie formatées lors du pré- étiquetage (ex : `<PAUSE>`, `<allong>`, etc.).

[4]http ://www.uni-stuttgart.de/lingrom/stein/forschung/resource.html

par-contre	PRP	<unknown>
<IM>		
euh	INT	euh
<PAUSE>		
la	DET:ART	le
photocopieuse	NOM	photocopieuse
est	VER:pres	être
très	ADV	très
utile	ADJ	utile
pour	PRP	pour
nous	PRO:PER	nous
<ID>		
pour	PRP	pour
faire	VER:infi	faire
des	PRP:det	du
schémas	NOM	schéma
des	PRP:det	du
<allong>		
<PAUSE>		
des	PRP:det	du
restructurations	NOM	restructuration
de	PRP	de
schémas	NOM	schéma

FIG. 9.7 – Exemple de sortie étiquetée

Post-étiquetage

A partir du corpus étiqueté, nous effectuons un post-traitement qui permet d'ajouter un certain nombre d'informations supplémentaires. Il s'agit principalement ici de corriger les sorties de TreeTagger et d'adapter les étiquettes récupérables. Les étiquettes qui n'étaient pas comprises dans le système d'annotation adopté par TreeTagger ont été rajoutées automatiquement à l'aide d'expressions régulières :

– Formatage des amorces (ajout du tag `<amorce>` en lieu et place du lemme) pour faciliter leur détection lors des traitements ultérieurs.

les	DET :ART	les
médias	NOM	médias
euh	INT	euh
natio-	VER :pper	<unknown> → natio- VER :pper **<amorce>**
nationaux	ADJ	national

– Suppression de la distinction de temps des verbes pour ne garder que la distinction de mode.

VER :futu
VER :impf
VER :pres
VER :simp
⟶ **VER :conj**

– Distinction des adverbes de négation

ne	ADV	ne → ne	**ADV :ne**	ne
pas	ADV	pas →	pas **ADV :pas**	pas

– Distinction pronoms clitiques / pronoms disjoints (par examen du contexte d'apparition)

je PRO :PER je → je **PRO :CLI** je

tuPRO :PER tu →tu**PRO :CLI** tu

etc.

Contexte 1

nous PRO :PER nous → nous **PRO :CLI** nous

Contexte 2

nous PRO :PER nous → nous **PRO :DSJ** nous

– Attribution d'une étiquette spécifique aux pauses remplies

euh INT euh → euh **EUH** euh

– Distinction genre (f : féminin ; m : masculin), nombre (s : singulier ; p : pluriel) et personne (1 : 1ère pers ; 2 : 2ème personne ; etc.) pour les pronoms et verbes

je	PRO :CLI	je	→	je	**PRO :CLI :s1**	je
ils	PRO :CLI	ils	→	ils	**PRO :CLI :p3**	ils
moi	PRO :DSJ	moi	→	moi	**PRO :DSJ :s1**	moi
la	DET :ART	le	→	la	**DET :ART :fs3**	le
est	VER :conj être		→	est	**VER :conj :s3** être	
arrivaient	VER :conj arriver		→	arrivaient	**VER :conj :p3** arriver	

– etc.

Ces éléments ont été ajoutés dans la mesure où la sortie de la passe précédente (avec les étiquettes originales de l'étiqueteur) est sous spécifiée concernant ces informations. Cette dernière passe est destinée à faciliter l'exploitation ultérieure de la sortie du module précédent pour la phase de regroupement en chunks.

par-contre	KON	par-contre
<IM>		
euh	EUH	euh
<PAUSE>		
la	DET:ART:fs3	le
photocopieuse	NOM	photocopieuse
est	VER:conj:s3	être
très	ADV	très
utile	ADJ	utile
pour	PRP	pour
nous	PRO:DSJ:p1	nous
<ID>		
pour	PRP	pour
faire	VER:infi	faire
des	PRP:det:p3	du
schémas	NOM	schéma
des	PRP:det:p3	du
<allong>		
<PAUSE>		
des	PRP:det:p3	du
restructurations	NOM	restructuration
de	PRP:de	de
schémas	NOM	schéma

FIG. 9.8 – Exemple de sortie suite au post-étiquetage

Ainsi à chaque passe, l'analyseur ajoute/modifie les informations morpho-syntaxiques, en s'appuyant sur les informations placées lors des passes antérieures. Notons que pour les cas d'ambiguïté restants, la séquence est conservée.

Entraînement sur le corpus de développement

TreeTagger propose également dans sa distribution standard un module d'apprentissage destiné à faire « apprendre » au système l'étiquetage de formes inconnues par rapport à celles des corpus classiques.

L'entraînement s'effectue avec la commande suivante :

```
train-tree-tagger <lexique> <fichier de classes ouvert> <fichier
d'entrée> <fichier de sortie>
```

Quatre fichiers sont donc nécessaires, et nous avons ainsi mis en place un script permettant de les générer automatiquement :

– **<lexique>**

Il s'agit d'un fichier contenant un lexique pleine forme. On retrouve un format similaire à celui requis par les autres étiqueteurs : une forme par ligne, chaque occurrence d'un mot étant suivie par une tabulation et un ensemble de paires tag-lemme, elles-mêmes séparées par des espaces.

Exemple :

conseil	NOM	conseil		
conseilla	VER :conj :s3	conseiller		
conseillai	VER :conj :s1	conseiller		
conseillaient	VER :conj :p3	conseiller		
conseillais	VER :conj :s1	conseiller	VER :conj :s2	conseiller
conseillait	VER :conj :s3	conseiller		
conseillant	VER :ppre	conseiller		
...	...	...		

Pour notre analyse, nous avons couplé le lexique généré à partir du corpus avec

le lexique issu du projet MULTEX de façon à enrichir la phase d'entraînement sur un plus grand nombre d'entrées.

– **<fichier de classes ouvert>**

Ce fichier contient une liste d'étiquettes susceptibles d'être affectées aux mots inconnus (mots non inclus dans le lexique).

Exemple (pour le système d'annotation du projet Penn Treebank) :

```
FW JJ JJR JJS NN NNS NP ...
```

Dans notre cas nous avons volontairement laissé ce fichier vide dans la mesure où nous contrôlons le traitement des mots inconnus lors de la phase de post-étiquetage.

– **<fichier d'entrée>**

Il correspond au fichier contenant des données balisées manuellement (un mot par ligne). Chaque ligne contient un token et un tag, séparés par une tabulation. Les marques de ponctuation sont considérées comme des tokens et doivent être étiquetées comme tels.

Nous avons mis en place un script supplémentaire pour générer ce fichier à partir du corpus initialement étiqueté (avant entraînement) auquel nous appliquons automatiquement un certain nombre de modifications après examen des erreurs d'étiquetage dans le fichier initial.

– **<fichier de sortie>**

Il s'agit du fichier dans lequel les paramètres résultants de l'apprentissage tagger sont enregistrés. Celui-ci sera alors utilisé ensuite comme nouveau fichier de paramètres pour l'étiquetage « post-apprentissage » afin d'étudier les améliorations éventuelles apportées par l'entraînement.

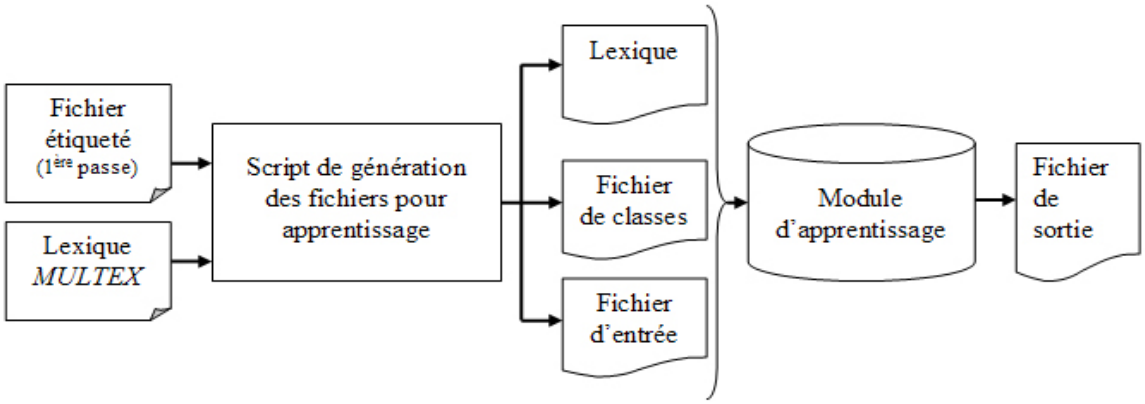


FIG. 9.9 – Organisation de la phase d'apprentissage (génération des fichiers et sortie du module)

D'autres paramètres – optionnels – sont également proposés, que nous n'avons utilisé pour nos traitements.

L'intérêt majeur de la personnalisation d'un certain nombre d'étapes dans la chaîne de traitement pour l'étiquetage, réside dans le fait de pouvoir examiner le « comportement » de TreeTagger face aux disfluences afin d'étudier la capacité potentielle de l'algorithme sous-jacent à « apprendre » celles-ci.

Étiquetage post-apprentissage

Dès la phase d'entraînement terminée, nous procédons ensuite à un nouveau passage de passes successives « pré-étiquetage / étiquetage / post-étiquetage » en utilisant le nouveau fichier de paramètre créé à l'occasion de l'apprentissage afin de générer un nouveau corpus étiqueté.

Comme en témoigne le tableau suivant (évaluation de l'étiquetage des amorces), l'une des difficultés résiduelle principale après apprentissage, concerne les amorces dont l'étiquette associée est souvent erronée, car leur identité est généralement

difficile à déterminer (les locuteurs prononcent les premières lettres d'un mot, ce qui ne suffit pas toujours à prédire le mot « prévu »).

N°	Correcte (√) / Erronée (×)	Catégorie attribuée	Catégorie attendue	N°	Correcte (√) / Erronée (×)	Catégorie attribuée	Catégorie attendu
1	×	VER:pper	PRO:DEM	39	√	NOM	
2	√	VER:pper		40	×	NOM	VER:pper
3	√	VER:pper		41	×	NOM	ADV
4	×	VER:pper	PRO:IND	42	×	NOM	VER:pres
5	×	NOM	PRO:IND	43	√	NOM	
6	×	NOM	PRP	44	×	VER:pper	ADJ
7	√	NOM		45	×	NOM	VER:pper
8	×	NOM	*ambigu*	46	×	VER:pper	DET:ART
9	×	NOM	*ambigu*	47	×	VER:pper	VER:infi
10	√	NOM		48	√	NOM	
11	×	NOM	DET:POS	49	×	NOM	VER:pres
12	×	VER:pper	ADV	50	×	VER:pper	ADV
13	√	VER:pper		51	×	VER:pper	PRO:PER
14	√	NOM		52	×	NOM	ADJ
15	×	VER:pper	DET:ART	53	√	NOM	
16	√	NOM		54	×	VER:pper	ADV
17	×	NOM	PRO:PER	55	×	NOM	ADV
18	×	VER:pper	PRO:DEM	56	×	VER:pper	PRP
19	×	VER:pper	PRO:DEM	57	×	NOM	VER:infi
20	×	VER:pper	ADV	58	×	NOM	*ambigu*
21	√	NOM		59	√	NOM	
22	×	VER:pper	ADV	60	×	NOM	*ambigu*
23	√	NOM		61	√	NOM	
24	×	NOM	PRP	62	×	VER:pper	VER:pres
25	×	VER:pper	PRP	63	√	ADJ	
26	×	VER:pper	NOM	64	×	NOM	*ambigu*
27	√	NOM		65	√	NOM	
28	×	NOM	*ambigu*	66	√	NOM	
29	×	VER:pper	ADJ	67	×	NOM	*ambigu*
30	√	NOM		68	×	NOM	DET:POS
31	√	NOM		69	×	NOM	VER:pres
32	√	NOM		70	×	NOM	VER:pres
33	√	NOM		71	×	NOM	VER:pres
34	√	NOM		72	√	NOM	
35	×	VER:pper	VER:pres	73	√	NOM	
36	×	NOM	ADJ	74	×	VER:pper	PRO:DEM
37	√	NOM		75	×	VER:pper	PRO:DEM
38	√	NOM		76	√	NOM	

Les cases où nous avons noté une catégorie attendue « ambigu » correspond justement aux cas où l'amorce est difficilement identifiable car composée d'une seule lettre et apparaissant dans un contexte où il n'est pas possible de déterminer précisément la catégorie touchée.

La figure suivante montre la proportion de cas d'amorces correctement et incorrectement étiquetés.

Au final, le corpus étiqueté servira de base pour l'étape suivante d'identification « brute » des disfluences avant de procéder au regroupement en chunks de celles-ci.

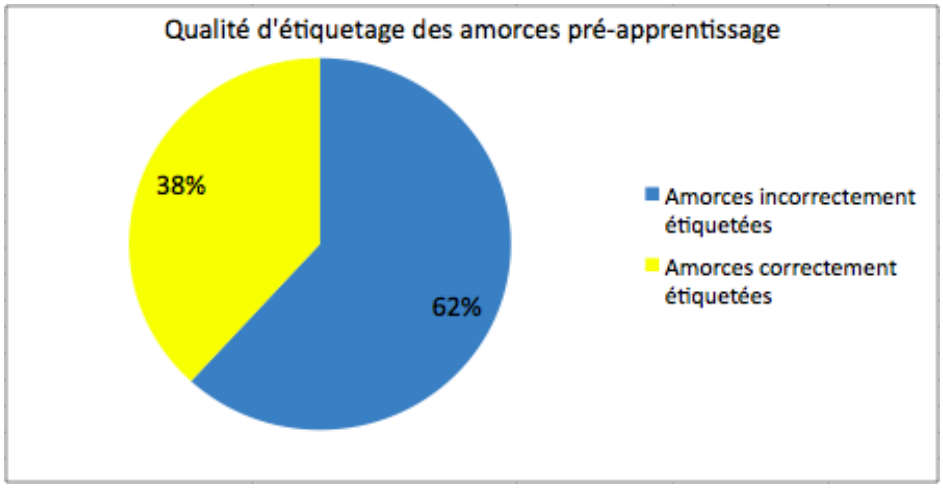


FIG. 9.10 – Proportion d'amorces correctement et incorrectement étiquetées

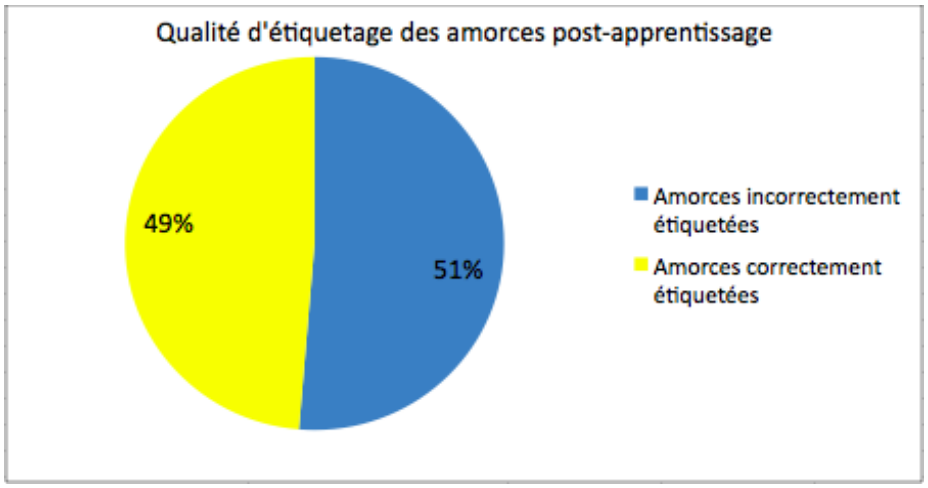


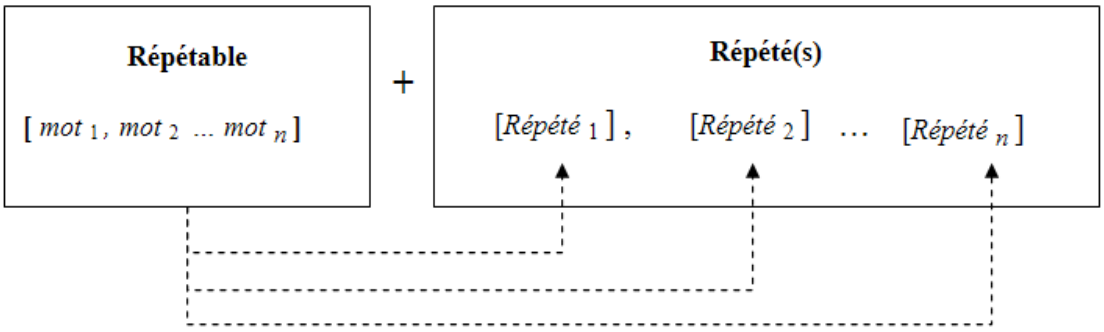
FIG. 9.11 – Proportion d'amorces correctement et incorrectement étiquetées après le module d'apprentissage

9.4 Détection « brute » des disfluences

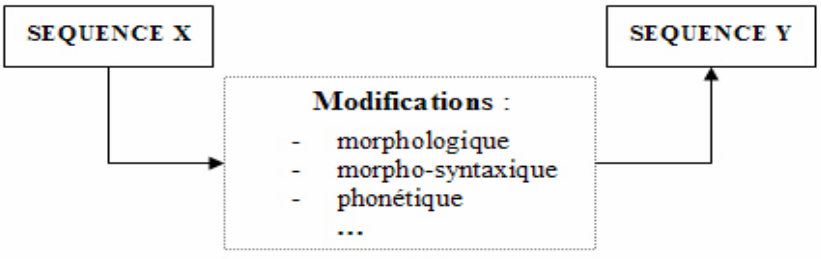
Cette étape intermédiaire entre étiquetage et regroupement en chunks, est destinée à signaler des zones potentiellement disfluentes à partir du résultat de l'étiquetage, en préalable de la phase de chunking. De même que la phase de chunking permet – notamment – de dégager des « îlots de confiance » (voir 9.5), ce module doit permettre de déterminer des points de « doute » au sein des séquences traitées. Il s'agit ensuite de s'appuyer sur ces zones pour améliorer la robustesse de l'analyseur au moment du regroupement en chunks.

Il convient de préciser que la détection est principalement focalisée sur les répétitions et autocorrections (du fait de la proximité structurelle des deux phénomènes). Rappelons brièvement la structure générale de ces deux catégories de disfluences :

Répétition :



Autocorrection :



Par ailleurs, nous avons déjà évoqué le statut des pauses silencieuses qui ne constituent pas des disfluences à proprement parler. Le traitement des amorces est trivial dans la mesure où le phénomène est déjà marqué par le processus de transcription qui attribue un tiret aux unités touchées. Il suffit donc de filtrer ces mots par rapport aux autres mots non amorcés pour les détecter ; cette étape est réalisée au moment de l'étiquetage. De même pour les pauses remplies, il s'agit simplement de lister toutes les formes possibles observées dans le corpus (*euh*, *hum*, etc.).

Sans revenir en détail sur les approches en TAL pour le traitement des disfluences, rappelons qu'il existe diverses tendances qui utilisent soit des techniques superficielles à base de N-grammes et de patrons, soit la syntaxe pour ce même traitement et généralisent par conséquent son utilisation à tous les phénomènes de production. Or, les expérimentations que nous avons menées sur le corpus ([Bove, 2008]) nous laissent à penser que nous pouvons tout à fait combiner ces techniques pour traiter les répétitions et autocorrections puisque ces phénomènes, par leur nature intrinsèque, ne nécessitent pas d'informations syntaxiques approfondies. La combinaison des techniques vise ainsi à améliorer le rapport coût de traitement / efficacité dans le traitement.

Nous nous appuyons donc à la fois sur :

- Des informations de **structure** : *i.e* l'identité de chaque mot et des mots qui le succèdent et le suivent (contexte immédiat). Des études comme celles de [McKelvie, 1998] par exemple, ont curieusement négligé cette source d'informations pourtant essentielle d'un point de vue pratique.
- Des informations **morpho-syntaxiques** : *i.e* les catégories morpho-syntaxiques des mots et leurs successions possibles.

En plus de ces techniques, nos choix de recourir à la technique de N-grammes de mots et d'identification de patrons syntaxiques ont fortement été guidés par deux

remarques :

- D'une part la tendance générale dégagée dans les travaux de ([Kurdi, 2003]) selon laquelle la fréquence d'un patron est inversement proportionnelle à sa taille. En d'autres termes, les patrons les plus petits sont les plus fréquents. L'analyse linguistique réalisée dans la partie précédente corrobore cette tendance.
- D'autre part, et dans le même ordre d'idée, nous avons noté lors de l'étude linguistique menée dans le chapitre précédent, que le nombre de répétés dans une répétition était le plus souvent unique (75,97% des cas) et plus rarement supérieur à 3 éléments (5,43% des cas).

  À ce propos, la littérature dans le domaine des pathologies du langage nous a apporté une information précieuse pour le développement de notre prototype. Concernant le bégaiement, il existe différents critères permettant de différencier un sujet non-bègue d'un sujet bègue ; parmi ces critères il apparaît que la différence entre les disfluences des sujets normaux et les disfluences pathologiques (ou bégaiements) des sujets atteints de troubles du langage porte sur la fréquence des répétitions de différentes unités (phonèmes, lexèmes, etc.). Plusieurs auteurs ([Conture, 1991], [Faure, 1980]) ont ainsi introduit un « seuil d'innacceptabilité », au delà duquel les répétitions portant sur des éléments larges doivent être considérés comme « anormales ». Plus précisément, ([Pfauwadel, 2000] [notamment]) considère par exemple que la répétition de mots ne sera considérée comme de nature bègue que si le nombre d'instances est supérieure à cinq unités ou plus. Ce chiffre corrobore notamment les observations relevées dans notre corpus pour les répétitions puisque le nombre maximum de répétés observé est de quatre éléments.

La figure suivante présente le module de détection mis en place :

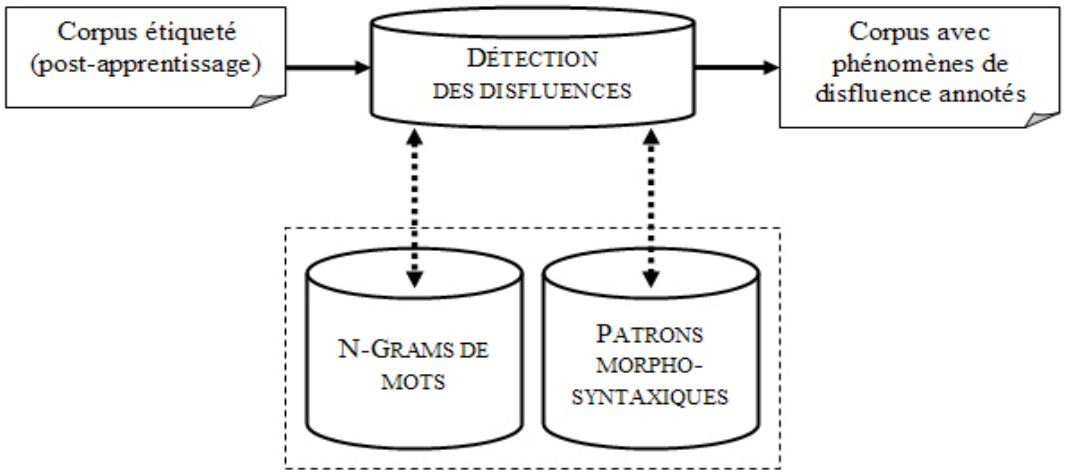


FIG. 9.12 – Module de détection « brute » des disfluences

Les différents phénomènes de production ont fait l'objet d'application des diverses techniques évoquées précédemment. De plus, une précaution doit être prise en amont de ces traitements. En effet, en examinant les différents cas de disfluences dans le corpus de développement, nous avons relevé (de même que dans d'autres études du domaine) plusieurs cas d'énoncés où des éléments sont insérés entre les différentes parties de la disfluence (qu'il s'agisse des répétitions ou des autocorrections). Par exemple, il peut s'agir de pauses remplies (PR) (*euh*, *hum*, etc.) ou encore de marqueurs discursifs (MD) (*ben*, *mais*, *donc*, *enfin*, etc.).

Ces éléments éloignent ainsi les couples « répétés - répétables » ou « séquence d'origine - séquence corrigée ». Bien que les pauses remplies et les marqueurs discursifs ne constituent pas simplement du « bruit » sans intérêt ([Engel *et al.*, 2002] par exemple, ont montré qu'ils pouvaient servir d'indices dans la détection des frontières d'énoncés), ils empêchent ici le module de détection de repérer plusieurs cas pourtant pertinents, détériorant ainsi la performance du système.

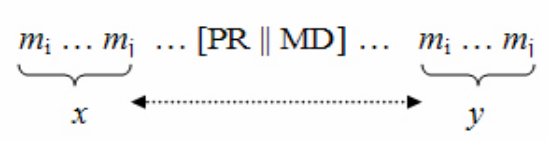


FIG. 9.13 – Distance entre les correspondances des parties de la disfluence

Pour pallier à cela, nous localisons ces éléments dans un premier temps, nous les mémorisons et enfin nous les restituons dans le corpus une fois l'exécution du module achevée (de façon à préserver la structure initiale du corpus). Ceci nous permet de réduire facilement la distance entre deux correspondances des parties d'une disfluence.

Exemple :

on se voit **euh** *on se voit*

$\longrightarrow$ <**Dis**> *on se voit on se voit* </**Dis**>

$\longrightarrow$ <**Dis**> *on se voit* **euh** *on se voit* </**Dis**>

Les intervalles x et y de la figure représentent des séquences de mots qui constituent respectivement la première et la seconde partie de la disfluence. La détection des répétitions consiste ainsi à chercher des N-grammes de mots identiques. Cependant, lorsqu'une correspondance est trouvée, le nombre de mots contenus dans le répété doit être limité. Eu égard aux remarques précédentes, nous avons trouvé qu'un ensemble de six mots (pauses remplies et marqueurs discursifs exclus) sont suffisants. Ainsi dans une répétition, le répétable repéré (et de fait l'ensemble du ou des répétés) doit contenir au plus six mots (*i.e* $x \leq 6$ et $y \leq 6$).

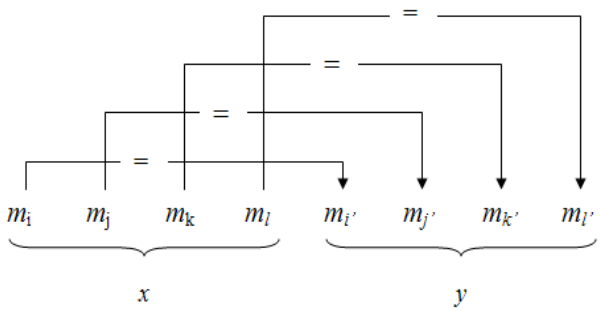


FIG. 9.14 – Exemple de N-grammes de mots pour les répétitions

Nous avons également intégré la technique de reconnaissance de patrons, dans la mesure où le corpus présente des informations structurales basées sur l'identité des mots, à côté des informations morpho-syntaxiques présentes sur les catégories. Deux types de patrons ont ainsi été utilisés :

- Des patrons **simples** : patrons uniquement basés sur les informations structurales où l'on vérifie uniquement l'identité du mot et son emplacement dans la chaîne traitée.
- Des patrons **mixtes** : patrons qui combinent l'information de structure à l'information morpho-syntaxique. Le traitement se fait non seulement en considérant l'identité des éléments mais aussi avec leurs catégories morpho-syntaxiques et leurs lemmes.

L'exemple suivant illustre un patron de détection simple, avec deux mots w_i et w_{i+1} (mot suivant immédiatement i) où l'étiquette et le lemme de w_i sont identiques à ceux de w_{i+1}, mais où les mots sont différents.

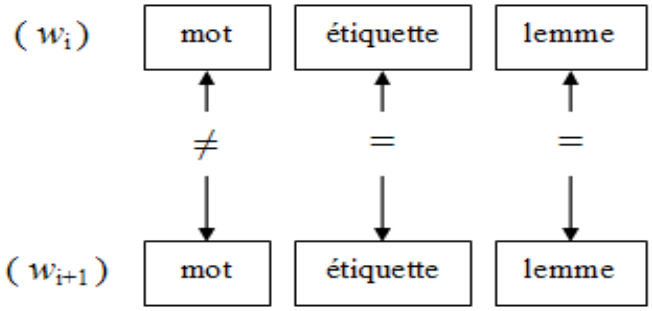


FIG. 9.15 – Exemple de patron pour les autocorrections

Exemple de segment reconnu à partir de ce patron :

⟶ `ce PRO :DEM ce  cette PRO :DEM ce  envie NOM envie`

9.5 Mise en place de la phase de regroupement en chunks

L'analyse menée précédemment (cf. II) sur les segments disfluents du CRFP a mis en avant le phénomène de « rembobinage syntagmatique » synonyme de retour au début d'un chunk avant de modifier ou corriger la production du locuteur.

Il s'agit ici d'exploiter une piste de l'analyse en chunks de séquences disfluentes, à l'instar des études réalisées pour l'écrit ([Aït-Mokhtar et Chanod, 1997]) ou dans des situations de parole conversationnelle en dialogue ([Goulian, 2002]) par exemple.

9.5.1 L'analyse en chunks

Les travaux de Steven Abney sont parmi les plus connus parmi ceux publiés ces quinze dernières années dans le domaine de l'analyse syntaxique robuste.
[Bourigault, 2007] rappelle à ce propos que la renommée de ces travaux "*tient autant à leur intérêt propre qu'au fait qu'ils ont été publiés à un moment où, face aux difficultés rencontrées par les approches classiques pour réaliser des analyseurs*

syntaxiques utiles, on s'intéresse plus ouvertement à l'analyse syntaxique robuste".

Le concept de chunk est donc issu des travaux de thèse d'Abney sur la structure du groupe nominal anglais ([Abney, 1991]) qui ont débouché entre autre sur le développement d'un analyseur syntaxique : CASS[5] (*Cascaded Analysis of Syntactic Structure*). Il vise à développer une analyse syntaxique qui soit à la fois plus rapide et plus fiable que ce que réalisent les analyseurs développés dans un cadre classique, sans abaisser le niveau d'exigence en terme de qualité et de profondeur des analyses.

Il s'agit d'une analyse syntaxique qui se base sur les différentes parties du discours, et où les chunks sont définis à partir des têtes sémantiques (*major heads*) des principaux types de groupes syntaxiques (SN, SV, SP, SAdj, SAdv). Tout mot plein est une tête sauf s'il est situé entre un mot fonctionnel (déterminant, préposition, ...) et le mot plein que ce mot fonctionnel sélectionne.

Jacques Vergne, opposé à l'approche classique en analyse syntaxique automatique, donne également une définition (non exhaustive) pour le chunk en français écrit (qu'il appelle « syntagme non récursif » ou *SNR* ([Vergne, 1999])) :

> "[un chunk] *est constitué d'un élément central, le plus souvent un nom (ou un pronom tonique) ou un verbe (conjugué, infinitif, participe présent ou passé), entouré éventuellement de ses éléments périphériques*".

Par exemple, un syntagme nominal contient obligatoirement un nom (tête du chunk) et commence généralement par un déterminant. Il est également marqué par un genre et un nombre homogènes sur tous ses composants variables en flexion. Un syntagme verbal commence par un verbe, parfois un pronom personnel ou un adverbe de négation, et contient nécessairement un verbe. C'est précisément cette

[5] http ://www.vinartus.net/spa/

définition que nous garderons à l'esprit pour mener notre phase d'implémentation.

Eléments périphériques possibles :

- Chunk nominal : conjonction de coordination et/ou de subordination, préposition, déterminant, adjectif épithète antéposé ou postposé, adverbe antéposé à l'adjectif épithète ;
- Chunk verbal : conjonction de coordination et/ou de subordination, préposition, tous les pronoms atones (sujet, objet et autres) antéposés ou postposés, négations, auxiliaire, adverbe le plus souvent postposé, adjectif attribut avec la copule être, adverbe antéposé à l'adjectif attribut.

La structure interne d'un chunk est relativement figée. Les mots fonctionnels qu'il contient entretiennent des relations de dépendance avec la tête lexicale et les contraintes d'accords sont généralement assez fortes. L'exemple suivant – extrait du *Bourgeois Gentilhomme* de Molière (repris de [Vergne, 1999]) – est constitué de cinq chunks que l'on peut permuter, mais au sein desquels les éléments ont une place fixe :

[*Belle Marquise*], [*vos beaux yeux*] [*me font*] [*mourir*] [*d'amour*]

Celui-ci est composé d'un premier chunk nominal (*belle marquise*), d'un second (*vos beaux yeux*), de deux chunks verbaux (*me font* et *mourir* que l'on peut trouver regroupé sous la forme d'un chunk unique) et enfin d'un chunk prépositionnel (*d'amour*).

Le segment choisi pour être le segment relié par les relations de dépendance est non pas le mot, mais bien ce « groupe de mots ». Un chunk est donc constitué par la séquence des mots entre le mot fonctionnel et le mot tête sélectionné. Les chunks sont non récursifs et ont une structure syntaxique qui est un sous-graphe connecté de l'arbre syntaxique de l'énoncé, mais ce n'est pas nécessairement un constituant

syntaxique intégral. Didier Bourigault souligne par ailleurs que ce concept joue un rôle clé pour la conception de l'analyseur syntaxique, et il s'est en partie appuyé sur celui-ci pour développer l'analyseur Syntex ([Bourigault, 2007]).

9.5.2 Grammaire de chunking dans le cas de productions orales

La structure générale d'un énoncé correspond généralement aux liens entre ses chunks. Nous utiliserons donc naturellement cette notion de chunks non récursifs pour l'implémentation de notre prototype. Toutefois, si en français écrit l'ordre des mots est relativement respecté dans un chunk, il n'en est pas de même de l'ordre des chunks à l'oral ([Antoine et Goulian, 2001]).

Les méthodes d'analyse classiquement proposées (qui suppriment le plus souvent les disfluences des énoncés à traiter) sont basées sur l'idée d'une grammaticalité illusoire des productions orales. Or, les arguments linguistiques de l'analyse menée sur notre corpus de développement (notamment celui rejoignant la notion d'« entassement paradigmatique » de Claire Blanche-Benveniste) attestent au contraire que les disfluences présentent des régularités sur lesquelles peut être menée une analyse automatique.

La correspondance entre les syntagmes de Blanche-Benveniste et les chunks d'Abney est frappante. A tel point qu'une grammaire de chunking apparaît tout à fait envisageable pour les énoncés oraux, avec l'avantage essentiel de réduire le nombre d'éléments de l'énoncé.

Comme le rappelle ([Goulian, 2002]), de nombreuses méthodes d'analyse structurelle de surface utilisant la notion de chunks ont démontré leur capacité à analyser de façon robuste des textes portant sur des domaines relativement larges ([Abney, 1996] ; [Aït-Mokhtar et Chanod, 1997] ; [Guiguet et Vergne, 1997]).

Nous adaptons ainsi la définition des chunks d'Abney (Abney, 1991) à notre problé-

matique pour développer un analyseur minimal, adapté à nos besoins. Les groupes grammaticaux considérés sont minimaux et non récursifs, bien qu'il n'y ait pas de définition unifiée de ce type de structure (car comme le note [Kahane, 2002] les constituants non récursifs sont absents des travaux de linguistique théorique). Il s'agit dans un premier temps de faire une passe de regroupement en chunks sous-spécifiée, ne cherchant à caractériser que les chunks non-disfluents comme le présentait (Goulian, 2002 ou Antoine et al., 2003) de façon à obtenir des îlots de confiance (cf. les " *islands of certainty* " d'Abney).

Nous proposons six groupes de chunks :

– SN (Syntagmes nominaux) :

```
<SN> j'_PRO :CLI :s1_je </FSN>
<SN> trois_NUM_trois mois_NOM_mois </FSN>
<SN> ma_DET :POS :fs3_mon 4L_NOM_4L </FSN>
```

– SV (Syntagmes verbaux) :

```
<SV> suis_VER :conj :s1_suivre|être restée_VER :pper_rester</FSV>
<SV> ai_VER :conj :s1_avoir dû_VER :pper_devoir changer_VER :infi_changer
</FSV>
```

– SAdj (Syntagmes adjectivaux) :

```
<SAdj> capable_ADJ_capable </FSAdj>
<SAdj> toute_PRO :IND_tout seule_ADJ_seul </FSAdj>
```

– SAdv (Syntagmes adverbiaux) :

```
<SAdv> pratiquement_ADV_pratiquement </FSAdv>
<SAdv> plus_ADV_plus longtemps_ADV_longtemps </FSAdv>
```

– SConj (Conjonctions [coordination, subordination], pronoms relatifs) :

```
<SConj> mais_KON_mais </FSConj>
<SConj> qui_PRO :REL_qui </FSConj>
```

– SPrep (Syntagmes prépositionnels) :

```
<SPrep> au_PRP :det :ms3_au <SN> Portugal_NAM_Portugal </FSN> </FSPrep>
<SPrep> sans_PRP_sans <SV> revenir_VER :infi_revenir <SAdv> d'ailleurs_ADV_d'ailleurs
</FSAdv> </FSV> </SPrep>
```

Notre analyse résulte de l'application d'une séquence finie et ordonnée de transducteurs. Chaque transducteur est utilisé pour introduire (dans la chaîne d'entrée étiquetée lors de la passe précédente d'étiquetage morpho-syntaxique) des marqueurs de délimitation autour des instances d'un chunk particulier. Les chunks sont décrits par un ensemble de règles exprimées sous forme d'expressions régulières. Ces règles reposent essentiellement sur les catégories grammaticales et les chunks déjà délimités.

Exemples :

– Règle pour les SPrep :

```
(PRP[^]* )+(PRO[^&]* )*SN\d+& (.*?)&FSN\d+
```

– Règle pour les SN :

```
(PRO :IND )*(PRO :DEM )*(DET[^&]* )0,1(NUM )*(ADV[^&]* )*(ADJ )*(KON
(ADV[^&]* )*ADJ )*(NAM|NOM)( NAM| NOM)*(ADV[^&]*)*( ADJ)*( KON( ADV[^&]*)*
ADJ)*
```

– Règle pour les SV :

```
(ADV :ne )0,1(DET[^&]* )0,1(PRO :DEM )0,1(PRO :CLIREF[^&]+ )0,1(PRO :REL
)0,1(PRO :DSJ[^&]* )0,1VER[^&]+( ADV[^&]*)*( ADJ)*( KON( ADV[^&]*( ADV[^&]*)*))*(
ADJ)*( VER :pper)0,1( ADJ)*( ADV[^&]*)*( KON ADV[^&]*)*( VER :infi)*
```

– Règle pour les SAdv :

```
(ADV[^&]+ )+ADJ( KON ADJ)*
```

– etc.

Le schéma suivant détaille les étapes d'analyse du module d'analyse syntaxique.

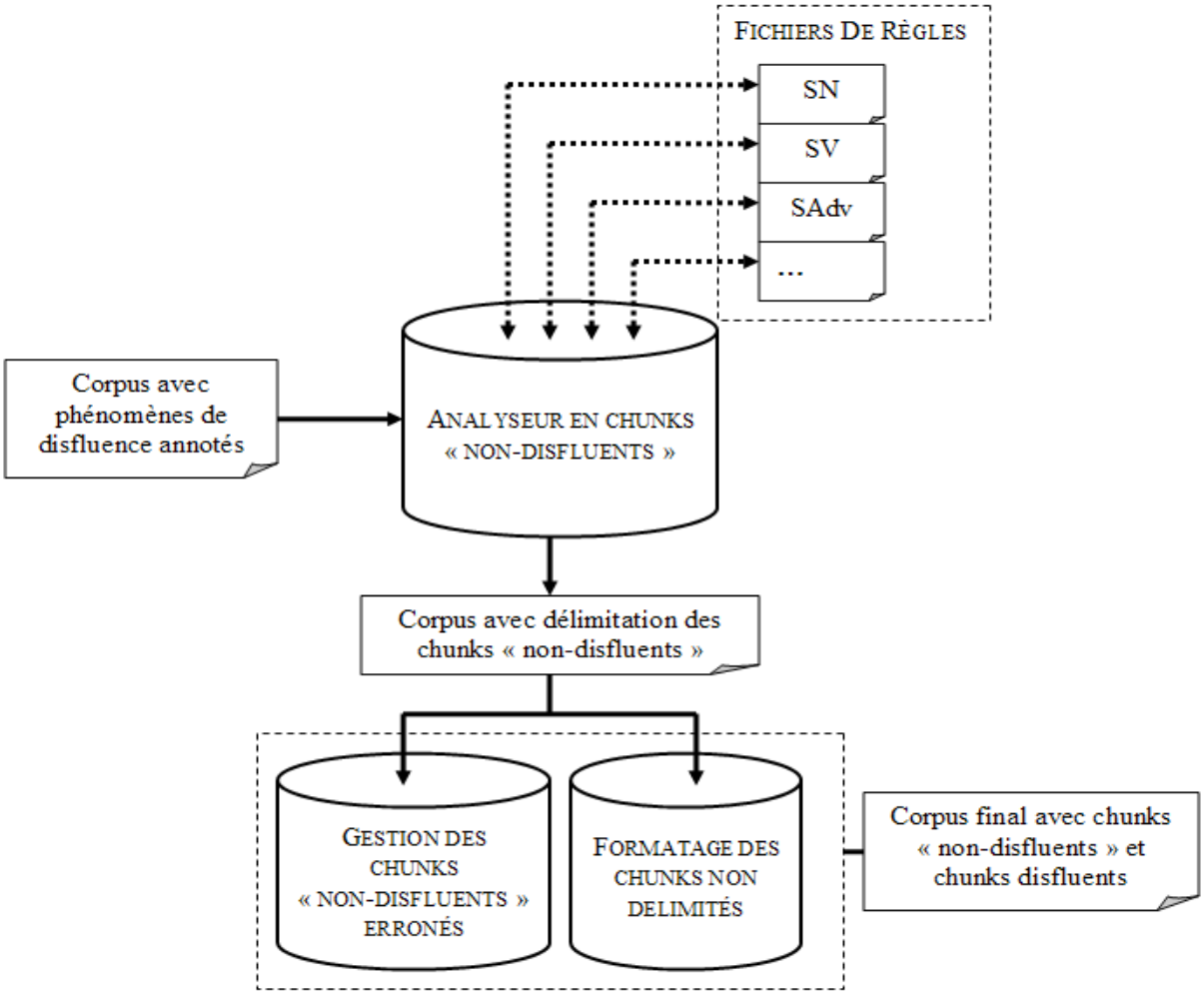


FIG. 9.16 – Architecture du module de segmentation

Il convient de remarquer que la segmentation n'est jamais destructrice à la différence des approches sélectives (par exemple [Bear *et al.*, 1992]) qui proposent d'effacer la partie de l'énoncé altérée par la disfluence. L'idée consiste à conserver l'ensemble de l'information véhiculée par le message oral, permettant ainsi une identification plus fine des intentions du locuteur.

Exemple de séquence non-disfluente analysée :

j'étais au bord de la mer

→ <SN> j'_PRO :CLI :s1_je </FSN> <SV> étais_VER :conj :s1_être </FSV>
<SPrep> au_PRP :det :s3_au <SN> bord_NOM_bord </FSN> </FSPrep> <SPrep>
de_PRP :de_de <SN> la_DET :ART :fs3_le mer_NOM_mer </FSN> </FSPrep>

La portée volontairement limitée du regroupement en chunks garantit une robustesse correcte tout en autorisant une analyse détaillée des énoncés oraux. Aussi, à l'inverse des approches à base de pré-traitements correctifs (par exemple Heeman, 1997), aucun élément n'est ignoré. De plus, la segmentation repose sur une connaissance syntaxique entièrement indépendante de la tâche.

Les zones non segmentées correspondent ensuite à des segments comportant une disfluence.

(...) *car les la la scolarité* (...)

→ <SConj> car_KON_car </FSConj> les_DET :ART :p3_le la_DET :ART :fs3_le*REP*
<SN> la_DET :ART :fs3_le scolarité_NOM_scolarité </FSN>

Néanmoins, il peut y avoir des cas où l'analyseur regroupe les unités en îlots de confiance de façon erronée car présentant une disfluence ambiguë avec un cas d'énoncé non-disfluent, où la séquence de catégories morpho-syntaxiques est similaire :

on on rend (...)

→ <SN> on_PRO :CLI :s3_on*REP* on_PRO :CLI :s3_on rend_VER :conj :s3_rendre
</FSN>

qui pourrait être confondu avec une séquence du type :

* *on le rend* (...)

*<SN> on_PRO :CLI :s3_on le_PRO :CLI :s3_le rend_VER :conj :s3_rendre
</FSN>

A ce stade du traitement, il reste alors trois cas de figures possibles pour les segments disfluents :

1. Les séquences n'entrant dans aucune configuration de chunk définie.
2. Les séquences contenant une amorce.
3. Les séquences regroupées en chunks malgré la présence d'une disfluence.

C'est précisément dans ces types de cas que la phase de détection préliminaire des disfluences intervient. Si le chunk comporte une ou plusieurs unité(s) signalée(s) comme disfluente(s) (indépendamment du phénomène concerné), le chunk est alors corrigé et délimité comme étant disfluent.

Ainsi, le croisement entre les séquences non segmentées (synonyme de disfluence), les chunks comportant une amorce, et les séquences segmentées à tort (car disfluentes) nous permettent d'obtenir un corpus où sont distingués les chunks « non-disfluents » des chunks disfluents.

Dans le cas de l'exemple précédent, nous obtenons un chunk catégorisé comme disfluent :

<SN> on_PRO :CLI :s3_on*REP* on_PRO :CLI :s3_on rend_VER :conj :s3_rendre
</FSN>

→ <Dis> on_PRO :CLI :s3_on*REP* on_PRO :CLI :s3_on rend_VER :conj :s3_rendre
</FDis>

De cette façon, il s'agit d'utiliser la structure syntaxique au niveau local minimum des syntagmes, sans chercher à effectuer de rattachements syntaxiques entre eux. De cette façon, il est ensuite possible de réaliser une analyse linguistique riche de l'énoncé tout en conservant une robustesse appréciable de celui-ci. L'avantage d'une analyse syntaxique partielle est d'être suffisamment générale pour ne pas être dépendante d'une application spécifique.

9.5.3 Évaluation

D'un point de vue méthodologique, l'évaluation a été réalisée sur un corpus de test extrait également du CRFP, et qui représente plus de 25% du corpus de développement. Précisons que nous n'avons pas utilisé le corpus constitué pour la création de la base de données d'arbres marcottés ; en effet, bien qu'une partie de ce corpus soit étrangère au corpus de développement, une autre partie était déjà extraite de ce même corpus de développement. Or, nous souhaitions avoir recours à un corpus n'ayant fait l'objet d'aucune manipulation préalable, de façon à garantir la fiabilité des résultats de l'évaluation.

Les tests ont porté dans un premier temps sur la détection des disfluences et plus particulièrement des répétitions et des autocorrections (les pauses remplies et les amorces présentant une détection triviale, tandis que les inachèvements sont plus délicats à identifier).

Puis, nous avons procédé à l'évaluation de la grammaire, d'une part pour la segmentation des chunks non-disfluents, et d'autre part pour les chunks disfluents. Nous utilisons classiquement trois critères d'évaluation : le rappel (R), la précision (P) et la F-mesure (Fm). Le but étant bien sûr de maximiser les trois scores.

La précision représente une mesure de l'efficacité du système par rapport au nombre de cas traités. Pourtant, elle n'est pas suffisante pour caractériser le comportement

global du système[6]. En revanche, le rappel tient compte de cet aspect, en indiquant la performance en terme de traitements corrects par rapport au nombre total des cas à traiter.

Les formules utilisées pour le calcul de ces métriques sont donc :

$$\mathbf{P} = \frac{\textit{Nombre d'éléments trouvés corrects}}{\textit{Nombre d'éléments trouvés}}$$

$$\mathbf{R} = \frac{\textit{Nombre d'éléments trouvés corrects}}{\textit{Nombre total d'éléments}}$$

Les deux métriques ne sont pas indépendantes. Il y a une forte relation entre elles, et il est délicat d'apprécier la qualité d'un algorithme qui fournirait une bonne précision et un mauvais rappel ou inversement. De plus, comment choisir le meilleur compromis lorsque la précision et le rappel sont pratiquement d'égale importance ?

Une des méthodes utilisées dans plusieurs domaines (dont le TAL) est de maximiser la moyenne harmonique de la précision et du rappel. On appelle cette moyenne la F-Mesure et elle se calcule ainsi :

$$\mathbf{F1} = \frac{2(\mathbf{P}\times\mathbf{R})}{(\mathbf{P}+\mathbf{R})}$$

La F-Mesure permet donc de faciliter la lecture du résultat de l'évaluation en calculant la moyenne des deux premières mesures. Nous présentons ci-après les

[6]Rappelons qu'une précision de 100% n'indique pas toujours un fonctionnement parfait. Par exemple, un système qui ne traite que 2 cas sur un total de 10, même pour une précision de 100% (2 réponses correctes pour 2 cas traités), ne représente pas un système satisfaisant.

résultats obtenus grâce à cette méthode.

	Précision	Rappel	F-mesure
Répétitions	91%	97%	94%
Autocorrections	83%	57%	67%

TAB. 9.1 – Évaluation de la détection des répétitions et des autocorrections

	Précision	Rappel	F-mesure
Chunks non-disfluents	96,12%	96,22%	96,16%
Chunks disfluents	90,28%	82,59%	86,27%

TAB. 9.2 – Évaluation de la grammaire de chunking

Ces résultats appellent quelques remarques. Tout d'abord, concernant l'étape de détection, plusieurs raisons expliquent les scores obtenus pour les répétitions, et surtout la forte baisse du rappel pour les auto-corrections.

Les erreurs restantes pour les répétitions proviennent principalement des cas ambigus entre les répétitions disfluentes et les répétitions d'emphase (*j'ai été en contact* **très très** *jeune* (...)). Pour les auto-corrections en revanche, la difficulté majeure concerne leur aspect structurel, où le schéma de construction des auto-corrections est souvent proche de ceux des énumérations (par exemple : **des plantes des fleurs** *euh extrêmement euh euh rares*). Dans ce deuxième cas en effet, le second élément énuméré (*des fleurs*) n'est pas une correction du premier (*des plantes*).

Ce type d'erreur de détection est ainsi reporté lors du regroupement en chunks, qui se voit, de plus, accru par des erreurs d'étiquetage en amont de l'analyse (malgré la phase d'apprentissage de TreeTagger). Les cas d'étiquetages erronés constituent à ce propos l'une des causes principales d'erreur pour la segmentation

des chunks non-disfluents. Voici quelques exemples d'erreurs recontrées, par type de disfluences :

- Répétitions

Exemple de répétition détectée disfluente dûe à la simple présence de l'emphase :

```
<Dis> ça_PRO :DEM :ça_cela s'_PRO :CLIREF :3_se est_VER :conj :s3_être
<SAdv> très_ADV_très très_ADV_très bien_ADV_bien </FSAdv> passé_VER :pper_passer
</FDis>
```

- Autocorrections

Exemple non détecté comme autocorrection dûe à la confusion possible avec certains cas d'énumérations à la structure semblable :

```
<SN> nous_PRO :CLI :p1_nous </FSN> <SV> avions_VER :conj :p1_avoir </FSV>
<SN> nous_PRO :CLI :p1_nous </FSN> <SV> <SAdv> n'_ADV :ne_ne </FSAdv> étions_VER :conj
<SAdv> pas_ADV :pas_pas très_ADV_très </FSAdv> <SAdj> tranquilles_ADJ_tranquille
</FSAdj> </FSV>
```

Toutefois, l'analyseur fournit une segmentation aux performances correctes compte tenu de la nature des données traitées. Malgré tout, des améliorations en différents points de l'analyse restent encore nécessaires, notamment dans une perspective de traitements ultérieurs prenant en entrée le corpus généré à l'issue de nos traitements. Par exemple, le recouvrement des disfluences à l'aide d'un contrôle sémantique pourrait limiter encore les erreurs issues des passes précédentes.

9.6 Conclusion

Ce chapitre a été consacré à la présentation du module d'analyse syntaxique de surface des disfluences. Le premier chapitre de cette partie était destiné à rappeler les prémices des traitements réalisés dans le second chapitre. Nous y avons

brièvement présenté les résultats de travaux ultérieurs menés dans un cadre précis d'expérimentation de traduction parole/parole. Les travaux d'analyse en dépendance d'un corpus oral conversationnel simulé, suggère une approche intéressante d'un point de vue pratique, mais quelque peu prématurée pour prétendre à une réelle robustesse au moment où celle-ci fut menée.

Aussi, les modules développés dans ce mémoire constituent les bases manquantes d'une telle analyse, en prévision d'un analyseur qui dépasse l'état de prototype. Nous avons ainsi proposé différents modules destinés à l'analyse syntaxique de surface basée sur corpus, reposant sur trois étapes principales :

- Étiquetage morphosyntaxique du corpus adapté aux données à traiter.
- Détection « brute » de disfluences pour dégager des « îlots de doute »
- Segmentation en chunks (non-disfluents et disfluents) du corpus.

En effet, dans le cadre qui nous intéresse, l'analyse syntaxique de surface revêt une importance particulière. Elle assure en particulier la généricité de l'analyse qui ne repose pour cette étape que sur un minimum de contraintes ne faisant pas intervenir de connaissances liées au domaine d'application.

Une étude ultérieure intéressante consisterait à examiner si les chunks possèdent une identité sémantique suffisante, ce qui permettrait de poser que la détermination de ces liens entre chunks peut être étayée par des connaissances sémantiques supplémentaires. Il est évident que celles-ci seront plus faciles à mettre en œuvre à partir du moment où le domaine de l'application permettra de réduire l'ambiguïté.

Conclusion

Cette thèse a été consacrée à l'étude et au traitement automatique des disfluences du français parlé spontané. De prime abord, de telles productions orales apparaissent d'une complexité décourageante pour l'analyse syntaxique automatique. L'étude que nous avons proposée essaie toutefois de montrer qu'en procédant avec méthode, cette complexité s'estompe.

L'hypothèse de base sur laquelle nous avons bâti notre démarche porte sur le fait que ces phénomènes, qui ne constituent pas un « bruit » de nature aléatoire, obéissent manifestement à une organisation syntaxique précise dont nous avons tenté de rendre compte.

Nous avions pour cela un double objectif :

- D'une part, donner une vision théorique d'ensemble du français parlé, focalisé sur les disfluences et leur analyse linguistique détaillée.
- D'autre part, identifier une démarche nous permettant l'élaboration d'un module d'analyse syntaxique de surface pour l'oral spontané.

Afin de répondre à cet objectif nous avons commencé par exposer les particularités du français parlé, avant de fournir une typologie détaillée des phénomènes de disfluence, illustrée de nombreux exemples extraits du *Corpus de Référence du Français Parlé* (CRFP). Nous avons poursuivi en situant notre travail par rapport aux différentes alternatives existantes dans le domaine du traitement des corpus oraux. Il s'agissait de mettre en parallèle les études théoriques et pratiques menées

ces dernières années, en présentant leurs avantages et leurs limites.

Après avoir positionné notre travail, nous nous sommes définis quatre tâches principales qui ont guidé notre travail :

- Identifier un cadre formel pour la représentation structurelle des disfluences et proposer l'analyse linguistique détaillée de celles-ci ;
- Adapter un analyseur morpho-syntaxique existant (TreeTagger) pour l'analyse syntaxique subséquente ;
- Concevoir et réaliser un module d'analyse syntaxique de surface pour les transcriptions d'oral spontané ;
- Évaluer les performances de notre analyseur.

Ces objectifs ont constitué l'objet de la démarche générale définie dans ce mémoire.

En consacrant une large part de cette thèse à la partie linguistique, nous avons dégagé précisément les caractéristiques de la langue parlée et la richesse structurelle des procédés que nous employons, consciemment ou inconsciemment lorsque nous nous exprimons à l'oral. C'est sur la base de l'identification précise de ces structures que doit en effet reposer la conception d'un module d'analyse syntaxique automatique de l'oral. Nous avons essayé dans notre étude linguistique et expérimentale de tenir compte le plus possible des aspects dus à la nature de la tâche (spontanéité des productions, parole monologique, etc.). Nous avons choisi de faire une étude cas par cas des disfluences afin de mieux comprendre la structure de tels phénomènes.

Cette partie a notamment abouti à la mise en place d'une base de données d' « arbres marcottés » visant à décrire les données linguistiques et les processus de production des disfluences mis en IJuvre dans leurs usages sur un autre support que

les enregistrements audio ou les transcriptions papier. La représentation adoptée a permis de mettre en évidence un phénomène de rembobinage syntagmatique, nous indiquant une piste de travail selon laquelle le *chunk* est le lieu privilégié des disfluences à l'oral.

C'est précisément cette piste que nous avons suivie pour développer un module fournissant l'analyse syntaxique de surface de notre corpus de travail. Nous avons présenté l'étiqueteur morphosyntaxique choisi, et les justifications de ce choix. Cette partie constituait également l'occasion d'expliquer les stratégies d'adaptation de l'outil aux données utilisées. Les traitements proposés dans ce travail ont été facilités par un contexte où le corpus utilisé correspond à ce que l'on pourrait attendre d'une sortie idéale en reconnaissance vocale (transcription fidèle de la parole du locuteur). Nous avons ainsi eu recours à un corpus transcrit comportant des informations pour certains phénomènes (cf. amorces de mots par exemple) qu'un système automatique n'est pas en mesure de fournir dans l'état actuel des connaissances.

Nous avons ensuite décrit les différents mécanismes mis en oeuvre pour exploiter de façon optimale la sortie de l'étiqueteur et procéder à l'analyse syntaxique. Ceux-ci se basent sur une détection préliminaire « brute » des disfluences donnant ainsi des indices sur les zones potentiellement problématiques pour la suite du traitement. Ces indices ont constitué une des informations utilisées pour procéder ensuite au regroupement en chunks du corpus, en distinguant d'une part les constructions syntaxiques correctes des chunks comportant une disfluence.

Les procédures développées étant – en l'état – indépendantes du domaine d'application, celles-ci peuvent être appliquées de manière générale dans d'autres types de corpora.

Nous avons obtenu des résultats préliminaires corrects qui témoignent d'un poten-

tiel considérable en terme de robustesse pour des traitements de niveau supérieur. Néanmoins, nous avons remarqué certaines faiblesses dans la chaîne d'analyse.

Tout d'abord, en ce qui concerne l'analyse morpho-syntaxique, le processus d'apprentissage réalisé sur TreeTagger ne suffit pas à lever tous les cas d'ambiguïtés, propageant ainsi des erreurs en cascade au moment du regroupement en chunks. De même, au niveau des constructions syntaxiques correctes, il reste certains oublis de constructions, telles que les locutions adverbiales par exemple.

Une autre faiblesse de l'analyse syntaxique, concerne le recouvrement des disfluences. Il reste encore des cas de disfluence qui ne sont pas correctement recouverts, particulièrement en ce qui concerne les autocorrections et les inachèvements. Pour le premier phénomène une difficulté majeure concerne sa nécessaire dissociation avec les cas d'énumérations ou de juxtaposition. Pour le second, c'est sa nature imprédictible qui en fait le phénomène le plus délicat à appréhender, et pour le quel nous n'avons pas de solution satisfaisante.

Ainsi, il est nécessaire d'effectuer d'autres évaluations de l'analyseur sur d'autres corpora afin de compléter les procédures de recouvrement. Plusieurs améliorations peuvent donc être ajoutées. Dans le point suivant présentant des projets de travaux futurs, nous présentons un plan de ces perspectives d'améliorations.

Si l'analyse automatique du français parlé spontané ne saurait se passer du pouvoir structurant de la syntaxe, elle ne pourra cependant pas, pour traiter ces constructions, reposer uniquement sur des considérations grammaticales. Le rôle des marqueurs discursifs par exemple, ne semble intéressant que si on les envisage à un niveau supérieur à la syntaxe. Par ailleurs, l'étude de ces constructions encourage la mise en place d'un module d'analyse sémantique, pour résoudre notamment des cas de disfluences plus complexes comme l'inachèvement.

L'étude réalisée a de plus permis de mettre en évidence les indices linguistiques

dont la prise en compte peut s'avérer pertinente dans la conception d'un système informatique de traitement de la langue parlé. En effet, les chunks correspondent souvent à des unités de sens représentant les objets de l'univers. La segmentation facilite donc la transition vers des traitements sémantiques (expérimenté dans [Bove, 2005]), ou sémantico-pragmatiques ([Antoine et Goulian, 2001]) ultérieurs.

A ce titre, les grammaires de dépendances et les formalismes basés sur cette notion ([Kahane, 2002]) semblent présenter un réel avantage. En envisageant la structure de l'énoncé selon les relations diverses que ses éléments entretiennent entre eux, cette approche permet une analyse suffisamment souple pour le traitement de l'oral. En effet, un des atouts majeurs des grammaires de dépendance est de pouvoir construire des structures incomplètes et donc d'intégrer directement le caractère hétérogène de la parole spontanée. Si la pertinence de la notion de dépendance semble intéressante, ce travail demande à être poursuivi et approfondi afin d'utiliser au mieux le principe de dépendance par rapport à un système d'analyse automatique finalisé de l'oral (par exemple au niveau de l'interaction Homme-Machine). La spécification de ce formalisme devra être mené conjointement à une réflexion sur le choix d'une architecture qui tienne compte de sa possible intégration dans un système de traitement automatique de la parole (reconnaissance vocale par exemple) pour aboutir in fine à un système d'analyse hybride utilisant une segmentation en chunks suivie d'une analyse des dépendances entre les segments identifiés, à l'instar de travaux tels que ([Aït-Mokhtar et Chanod, 1997], [Guiguet et Vergne, 1997]) pour l'écrit.

Cette étude a donc montré la faisabilité de l'intégration des phénomènes de l'oral dans l'analyse syntaxique, mais de nombreux problèmes restent évidemment à optimiser et à résoudre (pour les autocorrections et inachèvements, par exemple). La prise en compte plus large du contexte afin de favoriser la recherche des relations

de dépendances entre les têtes lexicales associées à ces unités peut être ensuite envisagée à un niveau sémantique. Par ailleurs, la gestion d'un nombre d'énoncés plus important pourra faire également l'objet de travaux futurs.

Bibliographie

[Abney, 1991] Abney, S. (1991). Parsing by chunks. *In* Berwick, R., Abney, S. et Tenny, C., éditeurs : *Principle-Based Parsing*. Kluwer Academic Publishers.

[Abney, 1996] Abney, S. (1996). Partial parsing via finite state cascades. *In Proceedings of the ESSLLI 96 Robust Parsing Workshop*, pages 337–344, Prague (République Tchèque).

[Adda-Decker *et al.*, 2004] Adda-Decker, M., Habert, B., Barras, C., Adda, G., Boula de Mareuil, P. et Paroubek, P. (2004). Une étude des disfluences pour la transcription automatique de la parole spontanée et l'amélioration des modèles de langage. *In Actes des XXVèmes Journées d'Étude de la Parole*, Fes (Maroc).

[Aït-Mokhtar et Chanod, 1997] Aït-Mokhtar, S. et Chanod, J. P. (1997). Incremental finite-state parsing. *In Proceedings of the 8th Conference on Applied Natural Language Processing, ANLP-97*, pages 72–79, Washington (USA).

[Allen et Guy, 1974] Allen, D. E. et Guy, R. F. (1974). *Conversation analysis : the sociology of talk*. The hague : mouton.

[Allen et Schubert, 1991] Allen, J. F. et Schubert, L. K. (1991). The trains project. Rapport technique, Computer science department., University of Rochester, Rochester.

[Allwood *et al.*, 1989] Allwood, J., Nivre, J. et Ahlsen, E. (1989). Speech management : on the non-written life of speech. *Gothenburg papers in theoretical linguistics*, 58.

[Antoine et Goulian, 2001] Antoine, J. Y. et Goulian, J. (2001). Étude des phénomènes d'extraction en français parlé sur deux corpus de dialogue oral finalisé. *Traitement automatiques des langues*, 42(2):413–440.

[Antoine *et al.*, 2003] Antoine, J. Y., Goulian, J. et Villaneau, J. (2003). Quand le TAL robuste s'attaque au langage parlé : analyse incrémentale pour la compréhension de la parole spontanée. *In Actes de la IXème conférence Traitement Automatique des Langues Naturelles*, Batz-sur-mer (France).

[Apotheloz et Zay, 1999] Apotheloz, D. et Zay, F. (1999). Incidents de la programmation syntagmatique : reformulations micro et macro-syntaxiques. *Cahiers de linguistique française*, 21:10–34.

[Assie, 2005] Assie, D. (2005). Analyse syntaxique automatique de corpus oraux retranscrits.

[Aubry *et al.*, 2007] Aubry, N., Bisazza, A., Cozler, C., Ezzat, M., Jean, C. et Tharrault, S. (2007). Alignalco, outil d'alignement de terminologie multilingue. Projet de master professionnel, http ://www.crim.fr/travaux_etudiants/2006-2007/alignalco/.

[Barberis et Maurer, 1998] Barberis, J. M. et Maurer, B. (1998). Sur le « ratage » en discours oral. *L'information grammaticale*, 77:43–47.

[Baude *et al.*, 2005] Baude, O., Blanche-benveniste, C., Calas, M. F., Cordereix, P., De Lamberterie, I., Goury, L., Jacobson, M., Marchello-Nizia, C. et Mondada, L. (2005). Guide des bonnes pratiques pour la constitution, l'exploitation, la conservation et la diffusion des corpus oraux. Délégation générale de la langue française et aux langues de france, http ://www.culture.gouv.fr/culture/dglf/guide_corpus_oraux_2005.pdf.

[Bear *et al.*, 1992] Bear, J., Dowding, J. et Shriberg, E. (1992). Integrating multiple knowledge sources for detection and correction of repairs in human-

computer dialog. *In Proceedings of the 30th annual meeting of the Association for Computational Linguistics*, Newark, Deleware (USA).

[Becker *et al.*, 1999] Becker, R., Boye, J., Carter, D., Lewin, I., Rayner, M. et Wiren, M. (1999). Language processing for spoken dialogue systems : is shallow parsing enough ? *In Accessing information in spoken audio : proceedings of ESCA ETRW workshop*, pages 37–42, Cambridge (UK).

[Beguelin, 2000] Beguelin, M. J. d. (2000). *De la phrase aux énoncés : grammaire scolaire et descriptions linguistiques*. Deboeck Duculot.

[Benzitoun, 2004] Benzitoun, C. (2004). L'annotation syntaxique de corpus oraux constitue-t-elle un problème spécifique ? *In Actes de la VIIIème conférence Rencontre des Etudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, pages 13–22, Fes (Maroc).

[Benzitoun *et al.*, 2004] Benzitoun, C., Campione, E., Deulofeu, H.-J., Henry, S., Teston, S., Valli, A. et Véronis, J. (2004). L'analyse syntaxique de l'oral : problèmes et méthode.

[Blanche-Benveniste, 1990] Blanche-Benveniste, C. (1990). *Le français parlé. Études grammaticales*. Paris : Collection Sciences du Langage, CNRS éditions.

[Blanche-Benveniste, 2000] Blanche-Benveniste, C. (2000). *Approches de la langue parlée en français*. Collection LŠessentiel Français. Éditions Ophrys.

[Blanche-Benveniste, 2003] Blanche-Benveniste, C. (2003). La naissance du syntagme dans les hésitations et répétitions du parler. *In* Champion, H., éditeur : *Le sens et la mesure. Hommages à Benoît de Cornulier*, pages 40–55.

[Blanche-Benveniste *et al.*, 1979] Blanche-Benveniste, C., Borel, B., Deulofeu, J., Durand, J., Giacomi, A., Loufrani, C., Meziane, B. et Pazeri, N. (1979). Des grilles pour le français parlé. *Recherches sur le français parlé*, 2:163–204.

[Blanche-Benveniste *et al.*, 1984] Blanche-Benveniste, C., Deulofeu, J., Stefanini, J. et Eynde, K. v. d. (1984). *Pronom et syntaxe. L'approche pronominale et son application au français.* Paris : Selaf.

[Blanche-Benveniste et Jeanjean, 1987] Blanche-Benveniste, C. et Jeanjean, C. (1987). *Le français parlé. Transcription et édition.* Paris : Didier érudition.

[Blasco-Dulbecco, 2004] Blasco-Dulbecco, M. (2004). Quelques éclairages sur le sujet de type "moi je" à l'oral. *Recherches sur le français parlé*, 18:127–144.

[Bock, 1982] Bock, J. K. (1982). Towards a cognitive psychology of syntax : information processing contributions to sentence formulation. *Psychological Review*, 89(1):1–48.

[Boomer, 1965] Boomer, D. S. (1965). Hesitation and grammatical encoding. *Language and Speech*, 8:148–158.

[Boufaden *et al.*, 1998] Boufaden, N., Delisle, S. et Moulin, B. (1998). Analyse syntaxique robuste de dialogue retranscrits : peut-on vraiment traiter l'oral à partir de l'écrit ? *In Actes de la Vème conférence Traitement Automatique des Langues Naturelles*, Paris (France).

[Bourigault, 2007] Bourigault, D. (2007). Un analyseur syntaxique opérationnel : Syntex. Mémoire d'Habilitation à Diriger les Recherches, Université Toulouse le Mirail.

[Bove, 2005] Bove, R. (2005). Impact des disfluences pour l'analyse syntaxique automatique de l'oral.

[Bove, 2008] Bove, R. (2008). A tagged corpus-based study for repeats and self-repairs detection in french transcribed speech. *In* Sojka, P., Horak, A., Kopecek, I. et Pala, K., éditeurs : *Actes de la XIème conférence internationale Text, Speech and Dialogue (TSD 2008)*, pages 269–276. Springer-Verlag, Brno (République Tchèque).

[Bove *et al.*, 2006] Bove, R., Chardenon, C. et Véronis, J. (2006). Prise en compte des disfluences dans un système d'analyse syntaxique automatique de l'oral. *In Actes de la XVème conférence Traitement Automatique des Langues Naturelles*, pages 103–111, Louvain (Belgique).

[Bove et Piu, 2007] Bove, R. et Piu, M. (2007). Annotation des disfluences dans les corpus oraux. *In Actes de la XIème conférence Rencontre des Etudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, pages 397–406, Toulouse (France).

[Brill, 1992] Brill, E. (1992). A simple rule-based part of speech tagger. *In Proceedings of the third conference on Applied Natural Language Processing*, Trento (Italie).

[Buoe et Waibel, 1996] Buoe, F. D. et Waibel, A. (1996). Learning to parse spontaneous speech. *In Proceedings of the International Conference on Spoken Language Processing*, volume 2, pages 1153–1156, Philadelphie (USA).

[Campione, 2001] Campione, E. (2001). *Étiquetage prosodique semi-automatique de corpus oraux : algorithmes et méthodologie*. Thèse de doctorat, Université de Provence.

[Campione et Véronis, 2004] Campione, E. et Véronis, J. (2004). Pauses et hésitations en français spontané. *In Actes des XXVèmes Journées d'Étude de la Parole*, pages 109–112, Fes (Maroc).

[Candéa, 2000a] Candéa, M. (2000a). Les « euh » et les allongements dits « d'hésitation » : deux phénomènes soumis à certaines contraintes en français oral non lu. *In Actes des XXIIèmes Journées d'Étude de la Parole*, Aussois (France).

[Candéa, 2000b] Candéa, M. (2000b). *Contribution à l'étude des pauses silencieuses et des phénomènes dits « d'hésitation » en français oral spontané*. Thèse de doctorat, Université de Paris III.

[Cappeau et Seijido, 2005] Cappeau, P. et Seijido, M. (2005). Les corpus oraux en français. Inventaire 2005 v.1.0, http ://www.culture.gouv.fr/culture/dglf/recherche/corpus_parole/presentation_inventaire.pdf.

[Carbonell et Hayes, 1983] Carbonell, J. G. et Hayes, P. J. (1983). Recovery strategies for parsing extragrammatical language. *American journal of computational linguistics*, 9:123–146.

[Carre *et al.*, 1991] Carre, R., Dégremont, J. F., Gross, M., Pierrel, J. M. et Sabah, G. (1991). *Langage humain et machine*. Paris : presses du CNRS.

[Chanet, 2003] Chanet, C. (2003). Fréquence des marqueurs discursifs en français parlé : quelques problèmes de méthodologie. *Recherches sur le français parlé*, 18:83–104.

[Church, 1988] Church, K. (1988). A stochastic parts program and noun phrase parser for unrestricted text. *In Proceedings of the second Conference on Applied Natural Language Processing*, Austin (USA).

[Clerc-Renaud *et al.*, 2004] Clerc-Renaud, J., Vasilescu, I., Candéa, M. et Adda-Decker, M. (2004). Étude acoustique et perceptive des hésitations autonomes multilingues. *In Actes des XXVèmes Journées d'Étude de la Parole*, Fes (Maroc).

[Conture, 1991] Conture, E. G. (1991). Young stutterers' speech production : A critical review. *In* Peters, H. F. M., W., H. et Starkweather, C. W., éditeurs : *Speech motor control and stuttering.*, page 365Ű384. Elsevier/Excerpta Medica : Amsterdam.

[Core et Schubert, 1998] Core, M. et Schubert, L. K. (1998). Implementing parser metarules that handle speech repairs and other disruptions. *In* Cook, D., éditeur : *Proceedings of North American Chapter of the Association for Computational Linguistics annual meeting*, Sanibel Island, Florida (USA).

[Core et Schubert, 1999] CORE, M. et SCHUBERT, L. K. (1999). A model of speech repairs and other disruptions. Working notes of the AAAI fall symposium on psychological models of communication in collaborative systems.

[Cutting *et al.*, 1992] CUTTING, D., KUPIEC, J., PEDERSEN, J. et SIBUN, P. (1992). A pratical part-of-speech tagger. *In Proceedings of the third Conference on Applied Natural Language Processing*, newark, Delaware (USA).

[De Smedt et Kempen, 1987] DE SMEDT, K. et KEMPEN, G. (1987). Incremental sentence production, self-correction and coordination. *In* KEMPEN, G., éditeur : *Natural Language Generation*. Kluwer.

[Delais-Roussarie et Choi-Jonin, 2004] DELAIS-ROUSSARIE, E. et CHOI-JONIN, I. (2004). Existent-ils des indices intonatifs de segmentation en unités macro-syntaxiques ? *In Actes des XXVèmes Journées d'Étude de la Parole*, Fes (Maroc).

[DELIC, 2004] DELIC, E. (2004). Présentation du corpus de référence du français parlé. *Recherches sur le français parlé*, 18:11–43.

[Deulofeu, 2004] DEULOFEU, J. (2004). Traitement automatique de l'oral. Cours de Master « Technologies du langage » (2e année). Université de Provence.

[Dubois *et al.*, 1994] DUBOIS, J., GUESPIN, L., GIACOMO, M., MARCELLESI, C., MARCELLESI, J. B. et MÉVEL, J. P. (1994). *Dictionnaire de linguistique et des sciences du langage*. Paris : Larousse.

[Eklund, 2004] EKLUND, R. (2004). *Disfluency in swedish humanŰhuman and humanŰmachine travel booking dialogues*. Thèse de doctorat, Linköping University (Suède).

[Engel *et al.*, 2002] ENGEL, D., CHARNIAK, E. et JOHNSON, M. (2002). Parsing and disfluency placement. *In Proceedings of the ACL-02 conference on Empirical Methods in Language Processing*, volume 10, pages 49–54, Morristown (USA).

[Faure, 1980] FAURE, M. (1980). Results of a contrastive study of hesitation phenomena in french and german. *In* DECHERT, H. W.and Raupach, M., éditeur : *Temporal Variables in Speech*, pages 287–290. The Hague : Mouton.

[Federici et Pirrelli, 1994] FEDERICI, S. et PIRRELLI, V. (1994). ontext-sensitivity and linguistic structure in analogy-based parallel networks. *Current issues in mathematical linguistics*, pages 353–362.

[Ferreira *et al.*, 2004] FERREIRA, F., LAU, E. F. et BAILEY, K. G. D. (2004). Disfluencies, language comprehension and tree adjoining grammars. *Cognitive Sciences*, 28:721–741.

[Finkler, 1997] FINKLER, W. (1997). A descriptive view of human self-corrections as the basis of a constructive approach to automatic self-corrections during incremental generation. *In Computational Psycholinguistics'97*, San Francisco (USA).

[Forget, 2000] FORGET, D. (2000). Les insertions parenthétiques. *Revue québécoise de linguistique*, 28(2):15–28.

[Fornel et Marandin, 1996] FORNEL, M. et MARANDIN, J. M. (1996). L'analyse grammaticale de l'auto-réparation. *Le gré des langues*, 10:6–68.

[Fox Tree, 1995] FOX TREE, J. E. (1995). The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of memory and language*, 34:709–738.

[Gala Pavia, 2003] GALA PAVIA, N. (2003). *Un modèle d'analyseur syntaxique robuste fondé sur la modularité et la lexicalisation de ses grammaires.* Thèse de doctorat, Université de Paris XI Orsay.

[Garrett, 1980] GARRETT, M. F. (1980). Levels of processing in sentence production. *In* PRODUCTION, B. B., éditeur : *Speech and talk*, volume 1, pages 177–220. Academic Press.

[Godfrey *et al.*, 1992] Godfrey, J., Holliman, E. et McDaniel, J. (1992). Switchboard : Telephone speech corpus for research and development. *In Proceedings of the International Conference on Spoken Language Processing*, page 517Ű520.

[Goto *et al.*, 1999] Goto, M., Itou, K. et Hayamizu, S. A. (1999). A real-time filled pause detection system for spontaneous speech recognition. *In Proceedings of the european conference on speech communication and technology (Eurospeech'99)*, pages 227–230, Budapest (Hungary).

[Goulian, 2000] Goulian, J. (2000). Analyse linguistique détaillée pour la compréhension automatique de la parole spontanée. *In Actes de la VIIème conférence Traitement Automatique des Langues Naturelles*, Lausanne (Suisse).

[Goulian, 2002] Goulian, J. (2002). *Stratégie d'analyse détaillée pour la compréhension automatique robuste de la parole.* Thèse de doctorat, Université de Bretagne Sud.

[Goulian *et al.*, 2002] Goulian, J., Antoine, J. Y. et Poirier, F. (2002). Compréhension automatique de la parole et tal : une approche syntaxico-sémantique pour le traitement des inattendus structuraux du français parlé. *In Actes de la Vème conférence Rencontre des Etudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, Nancy (France).

[Gravier *et al.*, 2006] Gravier, G., Huet, S. et Sebillot, P. (2006). Utilisation de la linguistique en reconnaissance de la parole : un état de l'art. Rapport technique, INRIA, Rennes.

[Guiguet et Vergne, 1997] Guiguet, E. et Vergne, J. (1997). Syntactic analysis of unrestricted french. *In Proceedings of the international conference on Recent Advances in Natural Languages Processing*, pages 276–281, Tzigov Chark (Bulgaria).

[Halliday, 1985] HALLIDAY, M. A. K. (1985). *Spoken and written language.* Oxford : Oxford University Press, Oxford (UK).

[Hübener *et al.*, 1996] HÜBENER, K., JOST, U. et HEINE, H. (1996). Speech recognition for spontaneously spoken german dialogues. *In Proceedings of the International Conference on Spoken Language Processing*, volume 1, pages 212–215, Philadelphie (USA).

[Heeman, 1997] HEEMAN, P. A. (1997). *Speech repairs, intonational boundaries and discourse markers : modeling speaker's utterances in spoken dialog.* Thèse de doctorat, University of Rochester (USA).

[Heeman et Allen, 1994] HEEMAN, P. A. et ALLEN, J. F. (1994). Detecting and correcting speech repairs. *In Proceedings of the 32nd annual meeting of the Association for Computational Linguistics*, pages 295–302, Las Cruces (New Mexico).

[Heeman *et al.*, 1996] HEEMAN, P. A., LOKEN-KIM, K. et ALLEN, J. F. (1996). Combining the detection and correction of speech repairs. *In Proceedings of the 4rd International Conference on Spoken Language Processing*, pages 358–361, Philadelphie (USA).

[Henry, 2002a] HENRY, S. (2002a). Quelles répétitions à l'oral ? esquisse d'une typologie. *In Actes des IIèmes journées de Linguistique de Corpus*, Lorient (France).

[Henry, 2002b] HENRY, S. (2002b). Étude des répétitions en français parlé spontané pour les technologies de la parole. *In Actes de la Vème conférence Rencontre des Etudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, pages 467–476, Nancy (France).

[Henry *et al.*, 2004] HENRY, S., CAMPIONE, E. et VÉRONIS, J. (2004). Répétitions et pauses (silencieuses et remplies) en français spontané. *In Actes des XXVèmes Journées d'Étude de la Parole*, Fes (Maroc).

[Henry et Pallaud, 2003] HENRY, S. et PALLAUD, B. (2003). Word fragments and repeats in spontaneous spoken french. *In Proceedings of Ddisfluency In Spontaneous Speech workshop*, pages 77–80, Gøteborg (Sweden).

[Henry et Pallaud, 2004] HENRY, S. et PALLAUD, B. (2004). Amorces de mots et répétitions : des hésitations plus que des erreurs en français parlé. *In Actes des VIIèmes Journées internationales d'Analyse de Données Textuelles*, Louvain-la-neuve (Belgique).

[Hindle, 1983] HINDLE, D. (1983). Determinic parsing of syntactic non-fluencies. *In Proceedings of the XXIst annual meeting of the Association for Computational Linguistics*, pages 123–128, Cambridge, Massachusetts (USA).

[Hirschberg et Nakatani, 1994] HIRSCHBERG, J. et NAKATANI, C. (1994). A corpus-based study of repair cues in spontaneous speech. *Jasa*, pages 1603–1616.

[Hockett, 1973] HOCKETT, C. F. (1973). Where the tongue slips, there slip i. *In* FROMKIN, V. A., éditeur : *Speech errors as linguistic evidence*, pages 93–119. The Hague : Mouton.

[Jorgensen, 2007] JORGENSEN, F. (2007). The effects of disfluency detection in parsing spoken language. *In* NIVRE, J., KAALEP, H. J., MUISCHNEK, K. et KOIT, M., éditeurs : *Proceedings of the 16th nordic conference of computational linguistics*, pages 240–244, Tartu (Estonie).

[Kahane, 2002] KAHANE, S. (2002). Grammaire d'unification sens-texte. vers un modèle mathématique articulé de la langue. Document de synthèse d'habilitation à diriger des recherches.

[Kasl et Mahl, 1965] KASL, S. V. et MAHL, G. F. (1965). The relationship of disturbances and hesitations in spontaneous speech to anxiety. *Journal of personality and social psychology*, 1:425–433.

[Kempe, 1993] KEMPE, A. (1993). A probabilistic tagger and an analysis of tagging errors. Rapport technique, Institut für maschinelle sprachverarbeitung, Universität stuttgart.

[Kleiber, 2003] KLEIBER, G. (2003). Faut-il dire adieu à la phrase ? *L'information grammaticale*, 98:17–22.

[Kurdi, 2003] KURDI, M. Z. (2003). *Contribution à l'analyse du langage oral spontané*. Thèse de doctorat, Université de Grenoble I.

[Lavie, 1995] LAVIE, A. (1995). *GLR* : a robust grammar focused parser for spontaneously spoken language*. Thèse de doctorat, School of computer science, Carnegie Mellon University, Pittsburgh.

[Levelt, 1983] LEVELT, W. J. M. (1983). Monitoring and self-repair in speech. *Cognition*, 14:41–104.

[Levelt, 1989] LEVELT, W. J. M. (1989). *Speaking : from intention to articulation*. Cambridge massachusetts : the MIT press, Cambridge (USA).

[Lickley, 1994] LICKLEY, R. J. (1994). *Detecting disfluency in spontaneous speech*. Thèse de doctorat, University of Edinburgh. (Scotland).

[Liu, 2003] LIU, Y. (2003). Word fragment identification using acoustic-prosodic features in conversational speech. *In HLT-NAACL student research workshop*, pages 37–42, Edmonton (Canada).

[Liu *et al.*, 2003] LIU, Y., SHRIBERG, E. et STOLCKE, A. (2003). Automatic disfluency identification in conversational speech using multiple knowledge sources. *In Proceedings of the european conference on speech communication and technology*, Geneva (Suisse).

[Luzzati, 2004] LUZZATI, D. (2004). Le fenêtrage syntaxique : une méthode d'analyse et d'évaluation de l'oral spontané. MIDL workshop.

[Maclay et Osgood, 1959] MACLAY, H. et OSGOOD, C. E. (1959). Hesitation phenomena in spontaneous english speech. *Word*, 15:19–44.

[Martinie, 1999] MARTINIE, B. (1999). *Étude syntaxique des énoncés réparés en français parlé*. Thèse de doctorat, Université Paris X.

[Martinie, 2000] MARTINIE, B. (2000). Remarques sur la syntaxe des énoncés réparés en français parlé. *Recherches sur le français parlé*, 16:189–206.

[McKelvie, 1998] MCKELVIE, D. (1998). The syntax of disfluency in spontaneous spoken language. *In* SAMPSON, G. et MCCARTHY, D., éditeurs : *Corpus linguistics : reading in a widening discipline*. London (England).

[Morel et Danon-Boileau, 1998] MOREL, M. A. et DANON-BOILEAU, L. (1998). *Grammaire de l'intonation. L'exemple du français*. Collection bibliothèque de faits de langues. Éditions Ophrys, Paris-Gap.

[Pallaud, 1999] PALLAUD, B. (1999). Lapsus et phénomènes voisins dans la langue parlée. problèmes d'identification. *Recherches sur le français parlé*, 15:9–40.

[Pallaud, 2002] PALLAUD, B. (2002). Les amorces de mots comme faits autonymiques en langage oral. *Recherches sur le français parlé*, 17:79–102.

[Pelurson, 2006] PELURSON, A. L. (2006). Prise en compte d'informations sémantiques dans un système de traduction phrase-based.

[Pfauwadel, 2000] PFAUWADEL, M. C. (2000). *Un manuel du bégaiement*. Solal, Marseille (France).

[Piu, 2006] PIU, M. (2006). Annotation des disfluences dans les corpus oraux.

[Quimbo *et al.*, 1998] QUIMBO, F. C., KAWAHARA, T. et DOSHITA, S. (1998). Prosodic analysis of fillers and self-repair in japanese speech. *In Proceedings of the International Conference on Spoken Language Processing*, Sydney (Australia).

[Rose et Riccardi, 1999] ROSE, R. C. et RICCARDI, G. (1999). Modeling disfluency and background events in asr for a natural language understanding task. *In Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, Phoenix, Arizona (USA).

[Roubaud, 2004] ROUBAUD, M. N. (2004). Du bon usage des amorces dans la transcription des corpus. *Recherches sur le français parlé*, 18:163–184.

[Roulet, 2003] ROULET, E. (2003). Les relations de discours rhétoriques et praxéologiques dans la description des propriétés des constituants parenthétiques. *In Table ronde : nouveaux développements dans les recherches sur les relations de discours et leurs marqueurs, 8th International Pragmatics Conference*, Toronto (Canada).

[Sabio, 1996] SABIO, F. (1996). *Description prosodique et syntaxique du discours en français : données et hypothèses.* Thèse de doctorat, Université de Provence.

[Schegloff *et al.*, 1977] SCHEGLOFF, E. A., JEFFERSON, G. et SACKS, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language*, 53:361–382.

[Schmid, 1994] SCHMID, H. (1994). Probabilistic part-of-speech tagging using decision trees. *In Proceedings of International Conference on New Methods in Language Processing*, Manchester (UK).

[Shriberg *et al.*, 1997] SHRIBERG, E., BATES, R. et STOLCKE, A. (1997). A prosody-only decision-tree model for disfluency detection. *In* KOKKINAKIS, G., FAKOTAKIS, N. et DERMATAS, E., éditeurs : *Proceedings of the european conference on speech communication and technology (Eurospeech'97)*, volume 5, pages 2383–2386, Rhodes (Grèce).

[Shriberg, 1994] SHRIBERG, E. E. (1994). *Preliminaries to a theory of speech disfluencies.* Thèse de doctorat, University of California, Berkeley (USA).

[Siu et Ostendorf, 1996] SIU, M. H. et OSTENDORF, M. (1996). Modeling disfluencies in conversational speech. *In Proceedings of the International Conference on Spoken Language Processing*, volume 1, pages 386–389, Philadelphie (USA).

[Spilker *et al.*, 2000] SPILKER, J., KLARNER, M. et GÖRZ, G. (2000). Processing self-corrections in a speech-to-speech system. *In* WAHLSTER, W., éditeur :

Verbmobil : foundations of speech-to-speech translation, pages 131–140. Springer-Verlag, Berlin (Germany).

[Stolcke et Shriberg, 1996] STOLCKE, A. et SHRIBERG, E. (1996). Statistical language modelling for speech disfluencies. *In Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 405–408, Atlanta (USA).

[Stouten et Martens, 2004] STOUTEN, F. et MARTENS, J. P. (2004). Benefits of disfluency detection in spontaneous speech recognition. Cost 278 and ISCA tutorial and research workshop (ITRW) on robustness issues in conversational interaction.

[Strassel, 2003] STRASSEL, S. (2003). Simple metadata annotation specification linguistic data consortium. Annotation guide, version 5.0., http ://www.ldc.upenn.edu/projects/mde/.

[Taylor et Cameron, 1987] TAYLOR, T. J. et CAMERON, D. (1987). *Analysing conversation : rules and units in the structure of talk.* Oxford : Pergamon, Oxford.

[Tesniere, 1959] TESNIERE, L. (1959). *Éléments de syntaxe structurale.* Éditions Klincksieck.

[Teston et Véronis, 2004] TESTON, S. et VÉRONIS, J. (2004). Recherche de critères formels pour l'identification automatique des particules discursives.

[Valli et Véronis, 2000] VALLI, A. et VÉRONIS, J. (2000). Étiquetage grammatical de corpus oraux : problèmes et perspectives. *Revue française de linguistique appliquée*, 4(2):113–133.

[Vergne, 1999] VERGNE, J. (1999). Étude et modélisation de la syntaxe des langues à l'aide de l'ordinateur, analyse syntaxique automatique non combinatoire : synthèse et résultats. Mémoire d'Habilitation à Diriger les Recherches, Université de Caen Basse-Normandie.

[Vergne et Guiguet, 1998] VERGNE, J. et GUIGUET, E. (1998). Regards théoriques sur le « tagging ». *In Actes de la Vème conférence Traitement Automatique des Langues Naturelles*, Paris (France).

[Villemonte de la Clergerie et Rajman, 2003] Villemonte de la CLERGERIE, E. et RAJMAN, M. (2003). Évolutions en analyse syntaxique. *Traitement Automatique des Langues*, 44(3):7–14.

[Véronis, 1998] VÉRONIS, J. (1998). Annotation automatique de corpus : état de l'art. *In Colloque international "questions de méthode dans la linguistique de corpus"*, pages 7–9, Perpignan (France).

[Véronis, 2004] VÉRONIS, J. (2004). Le traitement automatique des corpus oraux. *Traitement Automatique des Langues*, 45(2):7–14.

[Willems, 1981] WILLEMS, D. (1981). *Syntaxe, grammaire et sémantique. Les constructions verbales.* Publications de la Faculté des Lettres de Gand, Gand.

Résumé

Le but de cette thèse est d'étudier de façon détaillée l'impact des disfluences en français parlé (répétitions, auto-corrections, amorces, etc.) sur l'analyse syntaxique automatique de l'oral, et de proposer un modèle théorique permettant de les intégrer dans cette analyse. Notre axe de recherche se fonde sur l'hypothèse selon laquelle une analyse détaillée des énoncés oraux (principalement en termes morphosyntaxiques) peut permettre un traitement efficace pour ce type de données, et s'avère incontournable dans une optique de développement d'applications génériques dans le domaine des technologies de la parole.

Dans le cadre de ce travail, nous proposons à la fois une étude linguistique détaillée et une stratégie d'analyse syntaxique automatique partielle des disfluences (en syntagmes minimaux non récursifs ou "chunks"). Le corpus final obtenu est ainsi segmenté en chunks non-disfluents d'une part, à côté des chunks disfluents d'autre part, après prise en compte des régularités observées dans notre corpus. Les résultats de l'analyse automatique sont finalement évalués de façon quantitative sur le corpus permettant ainsi de valider le modèle théorique de façon empirique.

Mot clés : Traitement Automatique des Langues, Analyse Syntaxique Automatique, Disfluences, Oral, Français Parlé, Étiquetage Automatique, Grammaire de Chunking, Linguistique de Corpus.

Abstract

The aim of this PhD is to study in a detailed way the impact of disfluencies in spoken french (repeats, self-repairs, word-fragments aso.) on speech parsing, and to provide an theoretical model being able to integrate them in this analysis. Our research orientation is based on the assumption that a detailed analysis of the speech uterrances (mainly in morphosyntaxic terms) can allow an efficient processing for this kind of data, and proves to be essential in the development of generic applications in the field of speech technologies.

In this work, we propose at the same time a detailed linguistic study and a strategy of disfluencies chunking. The final corpus obtained is thus segmented in non-disfluent chunks on the one hand, beside the disfluent chunks on the other hand, after taking into account the regularities observed in our corpus. The results of the automatic parsing are finally evaluated in a quantitative way on the corpus, thus allowing to validate the theoretical model in an empirical way.

Keywords : Natural Speech Processing, Robust Parsing, Disfluencies, Speech, Spoken French, Tagging, Chunking, Corpus Linguistics.