Analyse et détection automatique de disfluences dans la parole spontanée conversationnelle

Camille Dutrey

UNIVERSITÉ PARIS-SUD

ÉCOLE DOCTORALE 427 : INFORMATIQUE PARIS-SUD
LABORATOIRE D'INFORMATIQUE POUR LA MÉCANIQUE ET LES SCIENCES DE L'INGÉNIEUR

DISCIPLINE : INFORMATIQUE

THÈSE DE DOCTORAT

Soutenue le 16 décembre 2014 par

Camille DUTREY

Analyse et détection automatique de disfluences dans la parole spontanée conversationnelle

Directrice de thèse :	M^me^ Sophie ROSSET	DR2 (LIMSI, CNRS – Orsay)
Co-encadrante de thèse :	M^me^ Chloé CLAVEL	MCF (EDF & LTCI, CNRS & Télécom ParisTech – Paris)

Composition du jury

Présidente du jury :	M^me^ Anne VILNAT	Professeure (LIMSI, CNRS & Univ. Paris-Sud – Orsay)
Rapporteurs :	M. Frédéric BÉCHET	Professeur (LIF, CNRS & Aix Marseille Univ. – Marseille)
	M. Frédéric LANDRAGIN	CR1 (LaTTiCe, CNRS & ENS & Univ. Paris 3 – Paris)
Examinatrice :	M^me^ Katarina BARTKOVA	MCF (ATILF, CNRS & Univ. de Lorraine – Nancy)
Invitée :	M^me^ Martine ADDA-DECKER	DR2 (LPP, CNRS & Univ. Paris 3 – Paris)

Thèse préparée au LIMSI-CNRS UPR 3251
Université Paris-Sud
BP133 – 91403 Orsay CEDEX

Au grand brun.

Résumé

Extraire de l'information de données langagières est un sujet de plus en plus d'actualité compte tenu de la quantité toujours croissante d'information qui doit être régulièrement traitée et analysée, et nous assistons depuis les années 90 à l'essor des recherches sur des données de parole également. La parole pose des problèmes supplémentaires par rapport à l'écrit, notamment du fait de la présence de phénomènes propres à l'oral (hésitations, reprises, corrections) mais aussi parce que les données orales sont traitées par un système de reconnaissance automatique de la parole qui génère potentiellement des erreurs. Ainsi, extraire de l'information de données audio implique d'extraire de l'information tout en tenant compte du « bruit » intrinsèque à l'oral ou généré par le système de reconnaissance de la parole. Il ne peut donc s'agir d'une simple application de méthodes qui ont fait leurs preuves sur de l'écrit. L'utilisation de techniques adaptées au traitement des données issues de l'oral et prenant en compte à la fois leurs spécificités liées au signal de parole et à la transcription – manuelle comme automatique – de ce dernier représente un thème de recherche en plein développement et qui soulève de nouveaux défis scientifiques. Ces défis sont liés à la gestion de la variabilité dans la parole et des modes d'expressions spontanés. Par ailleurs, l'analyse robuste de conversations téléphoniques a également fait l'objet d'un certain nombre de travaux dans la continuité desquels s'inscrivent ces travaux de thèse.

Cette thèse porte plus spécifiquement sur l'analyse des disfluences et de leur réalisation dans des données conversationnelles issues des centres d'appels EDF, à partir du signal de parole et des transcriptions manuelle et automatique de ce dernier. Ce travail convoque différents domaines, de l'analyse robuste de données issues de la parole à l'analyse et la gestion des aspects liés à l'expression orale. L'objectif de la thèse est de proposer des méthodes adaptées à ces données, qui permettent d'améliorer les analyses de fouille de texte réalisées sur les transcriptions (traitement des disfluences). Pour répondre à ces problématiques, nous avons analysé finement le comportement de phénomènes caractéristiques de l'oral spontané (disfluences) dans des données orales conversationnelles issues de centres d'appels EDF, et nous avons mis au point une méthode automatique pour leur détection, en utilisant des indices linguistiques, acoustico-prosodiques, discursifs et para-linguistiques.

Les apports de cette thèse s'articulent donc selon trois axes de recherche. Premièrement, nous proposons une caractérisation des conversations en centres d'appels du point de vue de l'oral spontané et des phénomènes qui le caractérisent. Deuxièmement, nous avons mis au point *(i)* une chaîne d'enrichissement et de traitement des données orales effective sur plusieurs plans d'analyse (linguistique, prosodique, discursif, para-linguistique) ; *(ii)* un système de détection automatique des disfluences d'édition adapté aux données orales conversationnelles, utilisant le signal et les transcriptions (manuelles ou automatiques). Troisièmement, d'un point de vue « ressource », nous avons produit un corpus de transcriptions automatiques de conversations issues de centres d'appels annoté en disfluences d'édition (méthode semi-automatique).

Mots-clefs : Traitement Automatique des Langues, Traitement Automatique de la Parole, Oral spontané, Parole Conversationnelle, Disfluences, Analyse Robuste, Centres d'Appels.

Abstract

Title : Disfluency Analysis and Automatic Detection in Conversational Spontaneous Speech

Extracting information from linguistic data has gain more and more attention in the last decades in relation with the increasing amount of information that has to be processed on a daily basis in the world. Since the 90's, this interest for information extraction has converged to the development of researches on speech data. In fact, speech data involves extra problems to those encountered on written data. In particular, due to many phenomena specific to human speech (e.g. hesitations, corrections, etc.). But also, because automatic speech recognition systems applied on speech signal potentially generates errors. Thus, extracting information from audio data requires to extract information by taking into account the "noise" inherent to audio data and output of automatic systems. Thus, extracting information from speech data cannot be as simple as a combination of methods that have proven themselves to solve the extraction information task on written data. It comes that, the use of technics dedicated for speech/audio data processing is mandatory, and epsecially technics which take into account the specificites of such data in relation with the corresponding signal and transcriptions (manual and automatic). This problem has given birth to a new area of research and raised new scientific challenges related to the management of the variability of speech and its spontaneous modes of expressions. Furthermore, robust analysis of phone conversations is subject to a large number of works this thesis is in the continuity.

More specifically, this thesis focuses on edit disfluencies analysis and their realisation in conversational data from EDF call centres, using speech signal and both manual and automatic transcriptions. This work is linked to numerous domains, from robust analysis of speech data to analysis and management of aspects related to speech expression. The aim of the thesis is to propose appropriate methods to deal with speech data to improve text mining analyses of speech transcriptions (treatment of disfluencies). To address these issues, we have finely analysed the characteristic phenomena and behavior of spontaneous speech (disfluencies) in conversational data from EDF call centres and developed an automatic method for their detection using linguistic, prosodic, discursive and para-linguistic features.

The contributions of this thesis are structured in three areas of research. First, we proposed a specification of call centre conversations from the prespective of the spontaneous speech and from the phenomena that specify it. Second, we developed *(i)* an enrichment chain and effective processings of speech data on several levels of analysis (linguistic, acoustic-prosodic, discursive and para-linguistic) ; *(ii)* an system which detect automaticcaly the edit disfluencies suitable for conversational data and based on the speech signal and transcriptions (manual or automatic). Third, from a "resource" point of view, we produced a corpus of automatic transcriptions of conversations taken from call centres which has been annotated in edition disfluencies (using a semi-automatic method).

Keywords : Natural Language Processing, Speech Processing, Spontaneous Speech, Conversational Speech, Disfluency, Robust Analysis, Call Centre.

Remerciements

Les remerciements sont l'occasion de parcourir ce temps si particulier de la préparation d'un doctorat, sans se plier à un exercice d'ordre strictement scientifique et professionnel, mais plutôt en s'attachant à repenser à ce que l'on a vécu. La thèse n'est pas l'œuvre d'une seule personne. Son aboutissement est conditionné par une myriade d'interactions humaines, tour à tour stimulantes, réconfortantes, stressantes, motivantes, inspirantes, décevantes, douloureuses, passionnantes...

Les premières personnes auxquelles je pense sont les chercheuses impliquées de près ou d'un peu moins près dans mon encadrement : Martine Adda-Decker, Chloé Clavel, Sophie Rosset et Ioana Vasilescu, ainsi qu'Anne Peradotto et Delphine Lagarde. Je les remercie très sincèrement de m'avoir accompagnée durant ces trois années. Nos interactions n'ont pas toujours été faciles, loin de là, mais elles ont eu à cœur que ce projet aboutisse dans les meilleures conditions possibles. Je suis particulièrement reconnaissante à Sophie – que j'ai harcelée autant qu'elle m'a poussée et secouée – pour sa patience et sa combativité durant la deuxième moitié de ma thèse, sans quoi je doute que ce projet eut été mené à bien.

Je remercie également les membres de mon jury, qui ont accepté d'évaluer mon travail avec expertise et bienveillance : Martine Adda-Decker, Katarina Bartkova, Frédéric Béchet, Frédéric Landragin et bien sûr ma présidente de jury, Anne Vilnat. J'ai particulièrement apprécié leurs remarques et questions durant ma soutenance, un moment privilégié qui permet de discuter le travail de recherche comme on le fait trop peu durant ces années de thèse. Un grand merci en particulier à Frédéric et Frédéric pour leurs rapports très instructifs.

Un moment charnière pour l'avancée du travail de thèse a été pour moi la préparation d'une soutenance à mi-parcours, organisée avec clairvoyance par Sophie Rosset. Elle sait tout le bien que cela m'a fait. À cette occasion, je remercie chaleureusement Frédéric Béchet et Anne Vilnat, qui déjà à cette occasion ont eu la gentillesse d'évaluer mon travail et de le faire évoluer dans la bonne direction.

Le travail de thèse est loin de n'être que le fruit de l'apprenti docteur, de ses encadrants et de ses évaluateurs. Un ingrédient est également indispensable à ce travail : les collègues ! À ce titre, pour nos échanges scientifiques et ceux plus décontractés, je remercie l'ensemble de ceux avec qui j'ai eu la chance d'interagir au LIMSI et à EDF.

Au LIMSI, je remercie les très nombreuses personnes avec qui j'ai toujours eu beaucoup de plaisir à échanger, en particulier Anne-Laure et la petite Anne, Driss, Nicolas, Gabriel, Mathieu, Sami, Sylvain, Olivier, Vincent, Munshi, Cyril (miam miam les brownies au chocolat blanc), Martin, Thomas, Aurélie. Beaucoup ont quelque part contribué à l'aboutissement de ma thèse, que ce soit durant mes débuts en TAL et dans la recherche lors de mon stage de Master 1 (un grand merci à Delphine, Aurélien et Houda qui m'ont tout simplement donné le goût de la recherche) ou pendant le *sprint* final de mon doctorat (mention spéciale à Gabriel et Nicolas pour les répétitions de soutenance en conditions extrêmes !).

À EDF, je remercie l'ensemble des groupes E72 et E74, et en particulier, pour leur bonne humeur au café et tous les bons moments passés ensemble (ah, le Diamant) : dans le désordre bien sûr, Benjamin, Mathilde, Lucile, Nadia, Kévin, Anne de M., Christallan, Laurent, Leslie, Jérémy, Jonathan, Bérénice, Leeley, Silvia, Alexis, Alzennyr, Alina et tous les autres.

L'amitié prend une coloration particulière pendant la thèse : durant cette période se nouent des liens très forts alors même qu'il devient très difficile de consacrer du temps à ses proches. Je remercie intensément ceux qui ont accompagnés mes pensées, avec la perspective heureuse de passer davantage de temps ensemble : Maurice, Lolo, May, Camille, Viên, Marina, Valentin.

Qu'aurais-je fait sans le soutien indéfectible de ma petite Ghita & Miléna Minou(che) Mimi chérie ? Je les remercie pour cette extraordinaire amitié, elles sont les personnes dont on a toujours besoin.

Merci du fond du cœur à ma super famille pour son amour. Mon père Dominique, ma mère et mon beau-père Fabienne & Pascal, ma fratrie de choc Sylvain, Claire et Gabriel-Ange. Une pensée particulière pour ma grand-mère paternelle, dont les pensées aimantes m'accompagnent au quotidien, et pour ma grand-mère maternelle, dont le souvenir est toujours fervent. Merci aussi du fond du cœur à Janny, qui m'a accompagnée et soutenue durant de nombreuses années, et à qui je dois beaucoup. Merci de m'avoir aidée à me construire en tant que personne.

Je ne sais comment remercier la personne la plus importante de ma vie d'adulte, Pierre, pour son amour et son soutien, pour m'avoir donné confiance en moi et m'avoir prouvé que je pouvais réaliser des choses que je pensais hors d'atteinte, pour ces dix merveilleuses années passées à ses côtés. La vie suit son cours, et il fera toujours partie de la mienne.

Enfin, je remercie Colline, Françoise, Patoche et Karaba, qui *de facto* m'ont accompagnée et m'accompagnent toujours au quotidien. Six, le nombre parfait. Ainsi, pour l'atteindre, merci à celui qui m'accompagne depuis peu d'être là, tout simplement. La vie est plus douce.

Table des matières

Liste des tableaux

Table des figures

Table des exemples sonores

Conventions de lecture

Exemples sonores

EXEMPLE SONORE – Conventions de lecture

« Énoncé accompagné du signal de parole correspondant (lien hypertexte) »
Texte d'accompagnement.

Transcriptions de parole

Symbole	Description
[PAUSE]	Pauses silencieuses
x-	Amorces de morphèmes

Notation des disfluences d'édition

Segment	Notations	Étiquettes liées à la détection
Reparandum	reparandum, `rpd`	`B-rpd, I-rpd, rpd`
Interregnum	interregnum, `int`	`B-int, I-int, int`
Reparans	reparans, `rpr`	`B-rpr, I-rpr, rpr`

Chapitre 1

Introduction : De l'étude de l'oral spontané à la détection des disfluences en contexte conversationnel

Sommaire

1.1 Problématique de recherche

EXTRAIRE de l'information de données langagières est un sujet de plus en plus d'actualité compte tenu de la quantité toujours croissante d'information qui doit être régulièrement traitée et analysée. La majorité des travaux se sont portés sur l'écrit mais depuis les années 90, avec une accélération autour des années 1999-2000, nous assistons à l'essor des recherches sur des données de parole également. La parole pose des problèmes supplémentaires par rapport à l'écrit, notamment du fait de la présence de phénomènes propres à l'oral (hésitations, reprises, corrections) mais aussi parce que les données orales sont traitées par un système de reconnaissance automatique de la parole [1] qui génère potentiellement des erreurs. Ainsi, extraire de l'information de données audio implique d'extraire de l'information tout en tenant compte du « bruit » intrinsèque à l'oral ou généré par le système de reconnaissance automatique de la parole. Il ne peut donc s'agir d'une simple application de méthodes qui ont fait leurs preuves sur de l'écrit.

L'utilisation de techniques adaptées au traitement des données issues de l'oral et prenant en compte à la fois leurs spécificités liées au signal de parole et à la transcription – manuelle comme automatique – de ce dernier représente un thème de recherche en plein développement et qui soulève de nouveaux défis scientifiques. Ces défis sont liés à la gestion de la variabilité dans la parole et des modes d'expressions spontanés. Plus généralement, l'extraction d'information à partir de données audio est un sujet d'actualité pour la communauté « parole », qui cherche à travers le domaine du *speech understanding* à aller au-delà de la transformation du signal sonore en une suite de mots. Il s'agit en effet d'accéder au sens. La communauté « texte » aborde à son tour la problématique de l'extraction d'information sur des types de données textuelles de plus en plus variés et « bruités », par exemple à travers le traitement de données issues du Web (forum de discussions, réseaux sociaux, *etc.*).

En effet, si jusqu'à récemment l'extraction d'information ou la structuration de données langagières concernaient essentiellement l'écrit bien formé, depuis quelques années on assiste à l'essor des travaux sur des données orales. Cet essor est lié aux progrès décisifs faits ces deux dernières décennies en transcription automatique de la parole. L'analyse robuste des données de parole a été abordée selon différents points de vue : la compréhension de la parole (*spoken language understanding*) dans le cadre du dialogue oral homme-machine notamment (*cf.* par exemple le projet européen AMITIES (Hardy *et al.*, 2003, Rosset *et al.*, 2008), le projet français MEDIA (Bonneau-Maynard *et al.*, 2006) ou plus récemment le projet européen LUNA (Dinarelli *et al.*, 2010)) ou encore de la reconnaissance

1. La reconnaissance automatique de la parole, ou RAP, est une technique relevant du traitement de la parole visant à transformer automatiquement le signal vocal en chaîne textuelle de parole retranscrite.

d'entités nommées (*cf.* par exemple le projet ETAPE (Galibert *et al.*, 2014) pour le français).

Par ailleurs, l'analyse robuste de conversations téléphoniques a également fait l'objet d'un certain nombre de travaux dans la continuité desquels s'inscrivent ces travaux de thèse. On peut citer par exemple la thèse de N. Boufaden sur l'extraction d'informations dans des conversations téléphoniques dans le domaine de la recherche et du sauvetage maritime (Boufaden, 2004) ou encore le projet CallSurf (Garnier-Rizet *et al.*, 2008) sur l'indexation et la structuration de données de centres d'appels téléphoniques afin d'y appliquer des tâches d'extraction de connaissances. Le problème majeur rencontré par les systèmes développés dans ces travaux concerne la gestion des erreurs de reconnaissance automatique de la parole et les mots hors vocabulaire, ainsi que les phénomènes spécifiques de l'oral spontané tels que les hésitations, les reprises, les auto-corrections, *etc.* Des expériences ont été menées dans la perspective de gérer les spécificités de l'oral avant tout (Rosset *et al.*, 2008, Garcia-Fernandez *et al.*, 2010, Vasilescu *et al.*, 2010b). Ces expériences mettent en exergue l'importance qu'il y a à être capable de gérer efficacement les particularités du traitement de données issues de la parole.

Cette thèse porte plus spécifiquement sur l'analyse des *disfluences* et de leur réalisation dans des données conversationnelles issues des centres d'appels de l'entreprise EDF [2], à partir du signal de parole et des transcriptions manuelles et automatiques de ce dernier. Ce travail convoque différents domaines, de l'analyse robuste de données issues de la parole à l'analyse et la gestion des aspects liés à l'expression orale *via* l'analyse des erreurs de transcription automatique. Son principal objectif est donc de proposer, en s'inspirant des travaux antérieurs portant sur l'analyse et le traitement automatique de corpus d'oral spontané conversationnel, des approches et modèles visant à structurer les données conversationnelles téléphoniques pour l'extraction de l'information qu'elles contiennent. L'un des aspects cruciaux de ce travail est la prise en compte des spécificités propres à ce type de données et notamment du défi représenté par leur caractère de conversations transcrites automatiquement.

Les données conversationnelles d'EDF sont représentées essentiellement par des transcriptions manuelles et automatiques de conversations enregistrées dans des centres d'appel, et représentent un corpus riche fournissant un cadre de travail particulièrement motivant en matière de recherche fondamentale et d'enjeux applicatifs (Danesi et Clavel, 2010). La palette thématique et expressive est à la fois large et circonscrite par le cadre applicatif. Les sources de variabilités sont particulièrement présentes dans ces données : variantes de prononciation, présence de phénomènes disfluents, structures agrammaticales des énoncés, accents étrangers ou régionaux, bruit environnant, *etc.* La prise en compte des transcriptions automatiques de données orales représente également un défi scientifique

2. EDF (Électricité de France) est l'une des premières entreprises mondiales dans le secteur de l'énergie. Site Web : http://www.edf.com/.

et applicatif. Actuellement, les erreurs de RAP sur des données conversationnelles homme-homme, du type dialogues entre clients et conseillers, restent nombreuses et les transcriptions automatiques de ces données atteignent des taux d'erreurs qui vont de 20 % à plus de 30 %, taux d'erreur bien supérieur à celui de 10 % – voire moins – mesuré sur des données de types émissions radio-télé-diffusées. Les possibilités de compréhension en sont naturellement affectées et des traitements sont nécessaires à la fois pour rendre ces données plus lisibles, récupérer l'information manquante lorsque les erreurs concernent des zones de parole pertinentes pour circonscrire la thématique des conversations ou encore supprimer les insertions « causes de bruit ». Les défis scientifiques à relever sont donc nombreux et sont particulièrement liés à la variabilité inhérente à la parole et aux modes d'expression spontanés, qui rendent par ailleurs la tâche de transcription automatique plus difficile.

L'énoncé présenté *infra*, en exemple 1.1, illustre ces défis liés aux phénomènes inhérents à l'oral spontané. Nous en présentons ainsi trois versions[3] : *(i)* la transcription manuelle du signal de parole, *(ii)* une version de la transcription manuelle exempte de toute marque d'oralité, et *(iii)* la transcription automatique correspondante, issue d'un système de reconnaissance automatique de la parole.

EXEMPLE SONORE 1.1 – Différentes transcriptions d'un énoncé d'oral spontané

(i) transcription manuelle : « vous voyez il y a il y a une augmentation une forte augmentation sur le trois kilos c'est pour ça que maintenant euh en fonction de l'utilisation des personnes des clients on leur dit euh un un six kilos vous avez plus de confort et en prix vous voyez par exemple un six kilos [PAUSE] un six kilos un trois kilos vous coûte » ;

(ii) transcription manuelle « nettoyée » : « il y a une forte augmentation sur le trois kilos c'est pour ça que maintenant en fonction de l'utilisation des clients on leur dit un six kilos vous avez plus de confort et en prix par exemple un trois kilos vous coûte » ;

(iii) transcription automatique : « une augmentation forte augmentation sur 3 kilos {breath} c'est pour ça que maintenant en fonction de l'utilisation des personnes les clients on leur dit à 6 kilos vous avez plus de confondre et en prie oui pas un peu 6 kilos 6 kilos un 3 kilos beaucoup ».

3. Les « exemples sonores » présentés dans ce manuscrit de thèse, visant à illustrer au mieux notre propos, sont constitués de transcriptions accompagnées du signal de parole correspondant. Les modalités de lecture de ces exemples sont présentées dans les conventions de lecture, page xxi.

1.2 Contexte de réalisation des travaux de thèse

Cette thèse a été réalisée dans le cadre d'une convention CIFRE établie entre deux laboratoires de recherches académiques et l'entreprise EDF.

Dans la mesure où le cadre théorique de nos travaux se situe à la frontière entre Traitement Automatique de la Parole et Traitement Automatique des Langues naturelles (TAL), ce partenariat nous a permis de mener à bien nos travaux sur ces deux plans. D'une part, nous avons bénéficié de l'expertise du Laboratoire de Phonétique et de Phonologie [4] (LPP) dans la recherche en phonétique expérimentale et en phonologie : les recherches qui y sont menées concernent plus particulièrement les systèmes sonores et leur architecture dans les langues du monde, la production et la perception des sons de parole, la variation dans les voix normales et pathologiques à travers les différents styles de parole et les diverses situations de communication, ainsi que l'utilisation des nouvelles technologies pour la phonétique expérimentale à partir de grands corpus. D'autre part, nous avons également pu jouir de l'expertise du Laboratoire d'Informatique et de Mécanique pour les Sciences de l'Ingénieur [5] (LIMSI) dans le développement de systèmes de communication et d'interaction entre l'homme et la machine. En effet, une partie de ces systèmes est fondée sur l'analyse automatique de la langue écrite et orale, du traitement du signal à ses composantes sémantiques. Le traitement de la langue écrite se concentre plus particulièrement sur la recherche d'information dans de grands ensembles documentaires hétérogènes à l'aide de systèmes question-réponse. Le traitement de la langue parlée concerne la transcription automatique de dialogues oraux sur des supports variés incluant entre autres l'identification des locuteurs, la reconnaissance de la langue et l'identification des émotions.

De plus, nos travaux de thèse sont ancrés dans un cadre applicatif particulièrement stimulant pour l'analyse de données d'oral spontané, grâce à l'étude de conversations issues des centres d'appels EDF. L'entreprise emploie 8 000 téléconseillers et traite 25 millions d'appels par an pour le seul marché résidentiel. Compte tenu du volume croissant des appels téléphoniques, le développement de technologies capables d'analyser automatiquement les enregistrements de ces appels constitue un réel besoin pour des applications commerciales. Les centres d'appels constituent en effet une interface stratégique pour EDF : il s'agit non seulement de comprendre la relation des clients à l'entreprise (par exemple, par le biais de l'identification des motifs d'appels ou de l'analyse de la satisfaction) mais aussi de mesurer les impacts de certains événements (publicité, nouvelles offres, hausse de prix, situation de crise, *etc.*).

Cette thèse fait suite au projet Vox Factory, un projet collaboratif de Recherche et Développements démarré en 2009 et terminé en août 2011, réunissant plusieurs partenaires académiques et industriels, dont le LIMSI et EDF. Le but du

4. Site Web : http://lpp.in2p3.fr/.
5. Site Web : http://www.limsi.fr/.

projet était d'optimiser la qualité des services et la satisfaction des clients par une meilleure formation des téléconseillers *via* l'analyse et la modélisation de la qualité et de l'efficacité de l'interaction client/agent en s'appuyant sur les technologies du traitement automatique de la parole et du traitement automatique des langues naturelles. En effet, les informations contenues dans les conversations issues de centres d'appels sont cruciales pour l'amélioration d'une part de la connaissance client et d'autre part de la gestion de la relation client. Elles permettent d'identifier les différents dysfonctionnements et bonnes pratiques, comme par exemple une mauvaise stratégie de communication, et peuvent servir de base pour la proposition de pistes d'amélioration. Dans cet objectif, différentes études ont déjà été menées au sein de la R&D d'EDF, que nous énumérons ici :

- l'analyse des stratégies dialogiques mises en place par le téléconseiller en fonction des postures du client : posture de victime ou posture offensive. Sur cet aspect, des analyses statistiques sur la performance du conseiller ont été menées par le laboratoire Lille Économie et Management [6] dans le cadre du projet collaboratif Vox Factory. Une autre étude menée dans le cadre de ce projet par le Laboratoire d'Informatique et de Mécanique pour les Sciences de l'Ingénieur (LIMSI) porte sur les analyses acoustiques des interactions émotionnelles entre le conseiller et le client (Vaudable, 2012) ;
- l'optimisation de la relation client : il s'agit ici de combiner des analyses de fouille de texte et des analyses statistiques afin d'identifier les réappels [7], les sujets sur lesquels les conseillers passent le plus de temps, les profils de conversation les plus fréquents ;
- l'analyse des réclamations et/ou des motifs d'insatisfaction des clients : les réclamations sont actuellement analysées à partir des commentaires rentrés par les conseillers en ligne dans le système d'information d'EDF (études réalisées périodiquement par le département « Analyse et connaissance Client » de la Direction Commerce). L'accès direct aux conversations permet une analyse plus fine, comme cela a été fait dans le cadre d'analyses sociolinguistiques ;
- l'analyse des motifs d'appels : au-delà des réclamations, il peut être utile d'avoir une visibilité sur les thématiques abordées par les clients lors de leur appel (mise en service, mise à jour des données clients, relève de compteur, *etc.*).

6. Site Web : http://www.iae.univ-lille1.fr/.

7. Dans le cadre de centres d'appels, le réappel correspond au deuxième appel (et aux suivants) passé entre le client et l'entreprise, après un premier contact n'ayant pas suffi à satisfaire une demande par exemple. Ces réappels constituent un coût important pour l'entreprise.

1.3 Axes de recherche et positionnement

L'analyse automatique des données orales pose de nombreux défis, et ce quel que soit l'angle d'étude : syntaxique, sémantique, *etc.* En effet, une des plus grandes spécificités de l'oral réside dans les ruptures au sein de l'énoncé : si l'on se place dans une perspective de comparaison entre production orale et écrite, de nombreux éléments viennent « perturber » la fluidité du discours oral ; ces éléments sont souvent assimilés à des signes de mise en œuvre de la parole, à l'instar du rôle d'un brouillon pour la production écrite. Ces phénomènes caractéristiques de la langue orale spontanée ont été considérés comme de simples éléments nuisibles à la qualité du discours et à sa compréhension : l'oral a ainsi longtemps été négligé en TAL, souvent sur des jugements normatifs quant à son manque de noblesse par rapport à la langue écrite. À ce désamour pour la langue parlée s'est ajouté le manque de données réellement disponibles, bien que certains se soient assez tôt attachés à la constitution de corpus et d'exempliers de productions orales, à l'image de C. Blanche-Benveniste. De ce fait, les études réalisées sur la langue prennent appui sur une production écrite normée, et les avancées réalisées dans le domaine du TAL sont peu adaptées au traitement automatique de la parole spontanée (Maclay et Osgood, 1959).

L'oral a donc souffert d'un double constat : peu d'études sur ses particularités linguistiques, et peu de méthodes et d'outils de traitement automatique adaptés à ces mêmes particularités. En effet, l'étude de l'oral a été délaissée pendant plusieurs décennies :

- jusqu'à ce que l'on s'intéresse au traitement de l'oral spontané dans le cadre d'applications comme la Reconnaissance Automatique de la Parole (RAP). De plus, le glissement de l'analyse de données semi-préparées vers des données spontanées a provoqué une confrontation de la communauté scientifique à la version « déstructurée » – par rapport à l'écrit – de la langue ;
- jusqu'à ce que la linguistique dite de corpus prenne de l'ampleur. Cette évolution a été un élément important dans le développement des études sur la langue parlée, dans la mesure où l'utilisation de grands corpus de parole (émissions radiophoniques, conversations en centres d'appels...) a permis d'appréhender et de mesurer les particularités de la langue parlée de manière à la fois précise et généralisable, de faire émerger les caractères intrinsèques de l'oral spontané ou de lier certains phénomènes à leur contexte d'apparition (en fonction du domaine dans lequel se situe la production par exemple) ; la possibilité d'enregistrer plus de données a eu comme conséquence l'émergence de la question de l'extrapolation à l'oral des descriptions faites sur la langue écrite.

L'écart s'est ainsi manifesté entre langue écrite et parlée ; on parle même d'une grammaire de l'oral tant les spécificités de la langue parlée se font sentir :

> « De l'oral à l'écrit, il y a un monde. La différence est si grande que la description du français oral ressemble plus souvent à celle d'une langue exotique qu'à la grammaire du français écrit telle que nous la connaissons. »
>
> Morel et Danon-Boileau (1998)

Blanche-Benveniste (1997) propose d'ailleurs une grammaire de l'oral ; ces prises de positions ne sont cependant pas unanimes, comme en témoigne Gadet (1989), qui envisage la langue comme « un système unique à deux manifestations ».

La langue peut ainsi être considérée selon un double niveau : le niveau lexical classique de la langue écrite, celui du « message » proprement dit, et un deuxième niveau extra-lexical, où se situent de nombreux phénomènes spécifiques à l'oral. Ces items extra-lexicaux peuvent être envisagés selon deux angles. Ils peuvent être identifiés comme des sources de « bruit », notamment pour des outils de TAL, pour l'efficacité desquels ils constitueraient une gêne, un obstacle. Ce point de vue sur la langue parlée a longtemps prévalu (Maclay et Osgood, 1959). Plus récemment, les études sur les phénomènes de la langue parlée ont accordé davantage de considération à ces spécificités (*cf.* par exemple Shriberg (1994)), leur donnant un statut lexical et un rôle important dans la modélisation des conversations, en tant qu'indices dialogiques notamment (Swerts *et al.*, 1998, Clark et Fox Tree, 2002, O'Connell et Kowal, 2005).

La langue orale n'est cependant pas uniforme et il existe une réelle variété de styles (parole lue, préparée, semi-préparée, spontanée en situation de monologue, spontanée en situation d'interaction, *etc.*) : nous nous intéressons tout particulièrement aux productions orales spontanées conversationnelles, qui sont davantage éloignées de la norme de l'écrit que de la parole dite semi-préparée (par exemple des monologues de discours politiques prononcés de mémoire). Dans cette perspective, et dans le cadre de ce doctorat réalisé en partenariat avec l'entreprise EDF, nous étudions plus particulièrement un large corpus de conversations issues de centres d'appels EDF. Les caractéristiques intrinsèques de ces données en font un terrain particulièrement intéressant à étudier : d'une part la parole conversationnelle génère une information fragmentée et l'expression spontanée de spécificités langagières (les pauses silencieuses ou remplies, les phénomènes disfluents, *etc.*) ; d'autre part la production en domaine restreint (conversations liées à l'entreprise EDF, domaine de l'énergie) induit l'utilisation d'un vocabulaire spécialisé ; l'emploi d'un lexique technique peut à son tour perturber le flux de parole. Ce sont également des données langagières directement produites par des consommateurs, clients, *etc.* qui intéressent tout particulièrement de nombreuses tâches de fouille et d'extraction d'information : c'est le cas des données issues de centres d'appels comme des données issues du Web (forums de discussion, blogs, *etc.*). Ces sources ont en commun d'exister en très grande quantité – et donc de refléter les pratiques de fond et de forme du plus grand nombre – mais aussi et surtout de comporter

un grand nombre de spécificités encore peu prises en comptes en TAL par rapport à la langue normée sur laquelle se sont longtemps appuyés les systèmes d'analyse automatique de la langue.

Notre étude porte sur l'analyse des éléments propres à l'oral spontané qui correspondent à un moment de l'énoncé où « le déroulement syntagmatique est brisé » (Blanche-Benveniste *et al.*, 1990), soit le phénomène de disfluence. Les disfluences peuvent amoindrir l'efficacité d'outils de traitement automatique des langues naturelles adaptés soit à des données textuelles écrites normées soit à des données orales nettoyées de ces aspects ou n'en contenant que peu (comme la parole lue ou préparée par exemple), comme c'est le cas généralement. Elles peuvent également gêner la lecture – et donc la compréhension – de conversations issues de l'oral pour un être humain. Surtout, seules ou associées à d'autres indices, elles peuvent avoir de nombreuses fonctions dans la production d'un énoncé et le déroulement du dialogue. Savoir les appréhender, les décrire, les identifier et les analyser peut jouer un rôle primordial pour l'analyse et l'interprétation d'autres tâches d'extraction d'information comme la détection d'opinion ou l'identification de moments clefs dans le discours.

Dans cette optique, l'accent peut par exemple être mis :

- sur le lien que les phénomènes disfluents entretiennent avec l'expression d'un objet inconnu, comme celui d'un vocabulaire non-familier (Yoshida et Lickley, 2010), ou leur fonction de recherche de mot (voir notamment Clark et Fox Tree (2002)) ;
- sur le lien entre la position des disfluences dans l'énoncé et l'introduction d'une information nouvelle ;
- sur le lien entre l'expression de certains phénomènes d'oral spontané et le degré de familiarité entre les locuteurs.

Plus précisément, les caractéristiques des disfluences peuvent avoir un effet positif sur les problèmes liés à la prise de parole et l'établissement d'expressions référentielles communes (Yoshida et Lickley, 2010). De leur côté, les hésitations vocaliques en tant que type de pause ont été associées à tout un éventail de rôles de planification en lien avec l'effort du locuteur à construire son message verbal dans le temps (Clark, 2002, Vasilescu *et al.*, 2008). Dans les centres d'appels, les deux classes principales de locuteurs sont concernées : que ce soit le client, qui ne parvient pas à exprimer son problème, ou l'agent, qui peine à donner une explication claire au client.

1.4 Apports des travaux de thèse

Nous détaillons dans cette section les apports de nos travaux de thèse selon trois axes : les principes, méthodes et résultats d'analyse proposés, les outils conçus et mis en œuvre, les ressources produites et/ou enrichies.

Analyses de corpus Nous proposons une caractérisation des conversations en centres d'appels du point de vue de l'oral spontané et des phénomènes qui le caractérisent.

Outils de Traitement Automatique des Langues Nous proposons *(i)* une chaîne de traitement et d'enrichissement des données orales effective sur plusieurs plans d'analyse (linguistique, prosodique, discursif, para-linguistique) et *(ii)* un système de détection automatique des disfluences d'édition adapté aux données orales conversationnelles, utilisant le signal et les transcriptions (manuelles ou automatiques).

Ressources linguistiques Nous produisons un corpus de transcriptions automatiques de conversations issues de centres d'appels annoté en disfluences d'édition (méthode semi-automatique).

1.5 Organisation du manuscrit

Les données orales conversationnelles issues de centres d'appels sont à la fois très intéressantes pour le secteur industriel et très riches en phénomènes langagiers à analyser. Dans la perspective de détecter les caractéristiques de l'oral spontané afin d'améliorer des tâches d'extraction d'information sur des données issues de l'oral, nous présentons au sein de ce manuscrit les recherches et travaux suivants :

Chapitre 2 – Définir, caractériser et détecter l'oral spontané : ce chapitre présente un état de l'art axé sur la définition, la caractérisation et la détection des phénomènes caractéristiques de l'oral spontané.

Chapitre 3 – Corpus de données conversationnelles en français : au sein de ce chapitre, à la jonction entre état de l'art et analyses de corpus, est présentée une description des corpus sur lesquels s'appuient nos travaux.

Chapitre 4 – Caractérisation des disfluences dans la parole : ce chapitre, en lien étroit avec le suivant, met en avant les méthodes et résultats d'analyses menées sur les disfluences en contexte conversationnel.

Chapitre 5 – Détection automatique de disfluences : ce chapitre central, à partir de l'état de l'art et des analyses présentés au sein des chapitres précédents, décrit la conception et l'évaluation d'un système de détection automatique des

disfluences d'édition sur des données orales spontanées.

Cette thèse est donc articulée selon deux grandes parties : la partie I, qui regroupe les chapitres 2 et 3, est dédiée à une description des travaux et ressources existants, préalables à nos travaux ; la partie II, qui regroupe les chapitres 4 et 5, est dédiée à la présentation des travaux réalisés durant cette thèse.

En complément de la présente introduction et de ces deux parties, le chapitre 6 est dédié à une discussion d'ensemble sur nos travaux de thèse et la présentation de nos perspectives de recherche.

Première partie

Définition, typologie et détection des disfluences, faits théoriques et représentation en corpus

Chapitre 2

Définir, caractériser et détecter les phénomènes d'oral spontané

Sommaire

LA première partie de cette thèse est dédiée à l'étude théorique des *disfluences* dans l'oral spontané. Dans cette perspective, nous proposons au sein ce chapitre 2 une vue d'ensemble des travaux portant sur la définition, la caractérisation et la détection automatique des phénomènes d'oral spontané. Nous attachons une attention particulière aux *disfluences d'édition*, que nous étudions plus particulièrement dans ces travaux de thèse – en lien avec d'autres manifestations d'oral spontané – et aux méthodes conçues pour les détecter automatiquement dans des données principalement en langues anglaise et en française. Les conclusions tirées de ce chapitre guident non seulement les travaux décrits dans la partie II de ce manuscrit, mais permettent également de positionner nos travaux par rapport à la communauté scientifique, du point de vue du cadre conversationnel au sein duquel nous nous plaçons (les conversations menées en centres d'appels) et des réalisations menées sur un corpus fermé.

2.1 Introduction : les énonciations propres à l'oral spontané conversationnel

Les phénomènes caractéristiques de la langue parlée, et plus particulièrement les énonciations propres à l'oral spontané conversationnel, constituent l'objet d'étude de cette thèse. La mise en avant des rôles qu'ils jouent dans la compréhension du déroulement du dialogue et la modélisation des conversations participe de l'intérêt croissant qui leur est porté. Leur détection est dans cette perspective indispensable et nous nous intéressons dans ces travaux, à travers leur définition, aux caractéristiques qui permettent de les identifier. Une de leurs singularités est notamment leur incidence sur le déroulement de l'énoncé, plus précisément leur caractère de rupture de l'énoncé. Nous reprenons dans cette optique la définition proposée par Blanche-Benveniste *et al.* (1990), qui identifient ces phénomènes, souvent regroupés sous l'appellation de *disfluences*, comme « un endroit où le déroulement syntagmatique est brisé » (au sein de l'énoncé).

EXEMPLE SONORE 2.1 – Un énoncé caractéristique de l'oral conversationnel

« alors et bien donc euh hum j'ai euh je j'étais jusqu'au mois de novembre j'étais sur euh hum euh j'avais fait passer mon mon compte de d'électricité euh espaces communs sur euh hum copro et hum euh [...] »

L'exemple 2.1 *supra* matérialise cette notion de brisure du déroulement syntagmatique de l'énoncé – une lecture humaine peu familière de ce type d'énoncé « butera » à plusieurs reprises avant d'identifier l'information que tente de faire passer le locuteur – et convoque bien une impression de piétinement dans la construction

du discours. Dans la mesure où nous nous attachons à l'étude des phénomènes disfluents dans le cadre de l'étude de conversations issues de centres d'appels, les exemples fournis dans cette thèse pour illustrer nos propos sont à quelques exceptions près issus du corpus sur lequel nous menons nos travaux (corpus issu de centres d'appels de l'entreprise EDF). Ici, il s'agit de l'extrait d'un tour de parole très disfluent, prononcé par un agent EDF lors d'un échange avec un client à propos des tarifs de l'électricité.

Par ailleurs, dans une perspective davantage tournée vers le traitement automatique des langues (TAL), on imagine aisément les difficultés auxquelles doivent faire face des outils d'analyse comme les étiqueteurs morpho-syntaxiques. Pour éclaircir cet aspect, nous présentons deux versions de cet énoncé : l'exemple 2.1 *supra* correspond à la transcription exacte de l'énoncé tel que prononcé par l'agent EDF, alors que l'exemple 2.2 *infra* correspond à ce même énoncé dans une version au sein de laquelle les éléments textuels relevant de la notion de disfluence ont été supprimés ; on peut ainsi se rendre compte de ce que serait un énoncé « allant à l'essentiel », sans marque d'oralité, de discours ou de spontanéité.

EXEMPLE SONORE 2.2 – Un énoncé exempt des marques d'oralité le composant

« ~~alors~~ ~~et bien~~ ~~donc~~ ~~euh~~ ~~hum~~ ~~j'ai euh je j'étais~~ jusqu'au mois de novembre ~~j'étais~~ ~~sur~~ ~~euh~~ ~~hum~~ ~~euh~~ j'avais fait passer ~~mon~~ mon compte ~~de~~ d'électricité ~~euh~~ espaces communs sur ~~euh~~ ~~hum~~ copro et ~~hum~~ ~~euh~~ [...] »

⇓

« jusqu'au mois de novembre j'avais fait passer mon compte d'électricité espaces communs sur copro et [...] »

Cette première représentation de l'oral spontané et des phénomènes disfluents qui lui sont propres, à travers la définition proposée par Blanche-Benveniste *et al.* (1990) et l'énoncé qui l'exemplifie, met en avant la pluralité de ces phénomènes.

En accord avec notre intérêt porté à l'identification automatique des disfluences, nous précisons cette caractérisation générale de l'oral spontané et définissons les phénomènes qui y sont relatifs en section 2.2. La section 2.3 est consacrée à une présentation de la typologie des disfluences d'éditions, à travers une définition de leur structure interne et des classes d'événements qui en dépendent, orientée vers la détection de ce phénomène. Nous proposons en section 2.4 une vue d'ensemble des différentes méthodes de détection des disfluences, afin de permettre d'identifier les caractéristiques d'apparition des phénomènes disfluents dans le flux de parole (signal et transcriptions), et en section 2.5 une discussion générale sur cet état de l'art.

2.2 Définitions des phénomènes d'oral spontané

Nous nous attachons dans cette section à définir l'oral spontané à travers le prisme des phénomènes qui le caractérisent. Ainsi, nous présentons en section 2.2.1 une vue d'ensemble des définitions et terminologies existantes, souvent liées à des positionnements scientifiques quant au rôle joué par ces phénomènes dans l'étude de la langue parlée. La section 2.2.2 donne une définition et une caractérisation de ces phénomènes.

2.2.1 Choix terminologiques et positionnement

Dès lors que l'on s'attache à définir les phénomènes d'oral spontané, le premier constat auquel on se trouve confronté est leur manque d'hégémonie et de ce fait la nécessité faite de s'appliquer à les redéfinir en fonction des besoins des travaux ou du positionnement théorique. Nous nous garderons d'introduire dans ce champ d'étude une terminologie nouvelle et autres choix typologiques individuels. Nos choix définitoires sont ainsi guidés par deux axes principaux :

1. une volonté d'éviter, tant que faire se peut, tout lexique dénotant le caractère *anormal* des phénomènes d'oral spontané, afin d'opérer de manière claire la distinction entre une parole spontanée et une parole *dysfluente* au sens pathologique du terme (par exemple bégaiement, conséquences de la dyslexie, *etc.*) [1] ;
2. une adéquation entre la définition théorique de ces phénomènes et leur réalisation effective.

Bien que généraliste, la définition proposée par Blanche-Benveniste *et al.* (1990) permet une caractérisation factuelle des disfluences, sans parti pris si ce n'est celui de prendre comme plan de référence l'axe syntagmatique. Par ailleurs, les disfluences sont souvent désignées sous ce terme, mais pas uniquement. Bove (2008, chap. 2) a mené un travail de recensement des dénominations utilisées dans la littérature pour recouvrir ces phénomènes. Ce travail met en avant la grande variété de termes employés et fait apparaître soit des *a priori* sur leur place, soit des angles d'analyse privilégiés, *etc.* : la liste *infra* présente les principaux termes faisant référence aux phénomènes disfluents et illustrant cette variété d'approches :

- « Non-fluence » (*Non-fluency*) dans Hindle (1983) ;
- Bribes, turbulences, marques ou phénomènes de production de l'oral dans Blanche-Benveniste *et al.* (1987) ;

1. de manière générale, le terme *dysfluence* est réservé à la parole pathologique, alors que *disfluence* est utilisé pour parler de la production orale non pathologique. Cette frontière est cependant assez floue (*cf.* par exemple l'utilisation du terme *disfluence* par Amblard et Fort (2014) pour leur étude de la parole dans un contexte clinique (discours de schizophrènes).

- Disfluence (*Disfluency*) dans Lickley (1994), Shriberg (1994) ou encore Heeman (1997) ;
- Disfluence de parole (*Speech Disfluency*) dans Shriberg (1994) ,
- Achoppements à l'oral, scories, ratés dans Pallaud (1999) ;
- Distorsions dans Boufaden *et al.* (1998) ;
- Marque de travail de formulation dans Candea (2000a) ;
- Phénomènes dits d'hésitation, également dans Candea (2000a) ;

On peut ainsi remarquer le caractère de jugement normatif de certaines dénominations (par exemple « ratés »). Nous ne retiendrons pas non plus la notion d'*hésitation*, car les marques d'oral spontané ne sont pas toujours liées à un processus cognitif lié à l'hésitation. Bien que les termes mêmes de *disfluence* ou de *phénomènes disfluents* gardent encore trace d'une connotation négative, ils font largement consensus dans la littérature. Nous les conservons donc, pour référer à l'ensemble des phénomènes d'oral spontané (cette dénomination sera également utilisée), tout en gardant en tête que l'occurrence de tels phénomènes, bien que marquée lexicalement et induisant une certaine irrégularité sur le plan syntaxique, n'entraîne pas nécessairement de disfluence d'un point de vue discursif, dans la mesure où (dans un cadre conversationnel) ces derniers ne passent pas complètement inaperçus mais ne sont pas systématiquement perçus non plus par l'interlocuteur. De trop rares tests de perception ont été mis en place sur cet aspect, *cf.* notamment Candea (2000b, p. 107) pour une étude des pauses en français et Fox Tree (1995) pour une étude des répétitions et des faux-départs en anglais.

Exemple sonore 2.3 – Pluralité des phénomènes disfluents dans un énoncé d'oral spontané

« *Vous voyez* il y a il y a une augmentation une forte augmentation sur le trois kilos c'est pour ça que maintenant euh en fonction de l'utilisation des personnes des clients on leur dit euh un un six kilos vous avez plus de confort et en prix *vous voyez* par exemple un six kilos [PAUSE] un six kilos un trois kilos vous coûte 66 82 un six kilos coûte 77 0 8 à l'année »

Décrire les disfluences ne se limite pas à les nommer : il faut représenter ce qu'elles englobent. En effet, tout comme les terminologies qui les définissent, les classements des disfluences ne s'accordent pas tous. Leur périmètre est variable selon les études et certains éléments se situent à la frontière des phénomènes disfluents. C'est notamment le cas des marqueurs discursifs, présents dans l'exemple 2.3 *supra*. Dans cet énoncé, le syntagme verbal « vous voyez » (en italique dans le texte), présent à deux reprises au sein de ce tour de parole, a une fonction phatique qui l'inclue dans la catégorie des marqueurs discursifs. Ces derniers sont ponctuellement assimilés aux disfluences.

Par ailleurs, un certain nombre de travaux portant sur l'identification des phénomènes d'oral spontané mettent en exergue la notion de méta-données orales, pour référer à d'autres phénomènes complémentaires des disfluences qui permettent, tout comme elles, de modéliser la structure de la production orale, et plus spécifiquement du dialogue.

Le terme de « méta-donnée structurelle » *(structural metadata)* est introduit par Kim *et al.* (2004) sur la base des méta-données telles que présentées par le Linguistic Data Consortium dans les *Simple Metadata Annotation Specification* (dans le cadre du programme DARPA [2] EARS [3]). Ce terme réfère à un ensemble de phénomènes qui combinés les uns aux autres permettent d'identifier la structure dialogique des données orales. Dans ces travaux, leur repérage joue un rôle prépondérant dans l'analyse de la parole.

Outre les disfluences, les méta-données structurelles comprennent notamment les frontières de phrase *(Sentence-like Units ou Semantic Units (SU) boundaries)*. Le besoin de définir et de typer les disfluences en lien étroit avec d'autres types de méta-données, pour mieux représenter la structure dialogique des échanges conversationnels, intervient aussi et surtout dans des contextes de tâche d'annotation ou d'élaboration de systèmes de détection des phénomènes disfluents.

2.2.2 Pluralité des phénomènes d'oral spontané

À travers une perspective de détection automatique des disfluences, nous avons étudié comment les travaux portant sur leur analyse, leur annotation et leur détection classent ces phénomènes : il nous semble pertinent de distinguer deux types de phénomènes, eu égard à la complexité de leur forme et à la place qu'ils occupent sur l'axe syntagmatique.

Nous envisageons d'une part des phénomènes caractéristiques de l'oral spontané en soi, tels les marqueurs discursifs ou hésitations vocaliques (pauses remplies). Ces éléments sont la plupart du temps regroupés dans la littérature sous le terme de *fillers* ou *fillers words*, notamment dans les travaux portant sur leur identification et leur annotation en corpus (*cf.* par exemple. les travaux du Linguistic Data Consortium (Strassel, 2004)) ou dans les nombreux travaux en linguistique et en TAL (*cf.* par exemple Adda-Decker *et al.* (2003), Boula de Mareüil *et al.* (2005), Bove *et al.* (2006), Bazillon *et al.* (2012)). On trouve également la notion de disfluences ponctuelles (Bouraoui, 2008) pour identifier ces *fillers*.

2. *Defense Advanced Research Projects Agency* : agence du département de la Défense des États-Unis chargée de la recherche et du développement des nouvelles technologies destinées à un usage militaire. Site Web : http://www.darpa.mil/, consulté le 16/01/15.

3. *Effective Affordable Reusable Speech-to-Text* : programme de la DARPA visant à développer les technologies de transcription automatique de la parole avec le plus de précision et d'enrichissements des données possible.

D'autre part, nous mettons en avant les disfluences dites d'édition, qui ont une structure bien définie sur le plan syntaxique et peuvent faire intervenir les éléments évoqués *supra*.

EXEMPLE SONORE 2.4 – Amorces, hésitations vocaliques et marqueurs discursifs

« et moi je suis à **c- euh** à ce jour **hein donc euh** à aujourd'hui **donc** quand il est passé »

Amorces, hésitations vocaliques et marqueurs discursifs sont des phénomènes disfluents ne nécessitant pas une sous-structuration comme les disfluences d'édition. Ces éléments entretiennent une forte relation entre eux dans la mesure où ils sont souvent présents dans leur co-texte respectif, comme en témoigne l'exemple 2.4 *supra*. Cet extrait de tour de parole comporte en effet une amorce (« c- ») suivie d'une hésitation vocalique (« euh »), elle-même suivie de deux marqueurs discursifs (« hein » et « donc ») puis d'une seconde hésitation vocalique (« euh »).

2.2.2.1 Les hésitations vocaliques

Pour ce phénomène aussi les dénominations sont plurielles : ces pauses remplies (Henry *et al.*, 2004), aussi appelées pauses sonores ou pleines, sont inclues par Shriberg (1994) dans le phénomène plus générique de *fillers*, qui représente également les marqueurs du discours ou particules discursives.

EXEMPLE SONORE 2.5 – Les hésitations vocaliques

« **euh** ben en fait **euh hum** »

Définies comme l'« insertion dans le flux verbal d'une voyelle fortement allongée, telle que "euh" en français » par Vasilescu *et al.* (2010a), les pauses remplies constituent un cas un peu particulier. En effet, en appui aux travaux de Candea (2000a), elles peuvent également avoir pour rôle de garder la parole, auquel cas leur statut disfluent est remis en question. Vasilescu *et al.* (2008) insistent également sur l'importance de ne pas opérer de confusion entre les voyelles d'hésitation et les voyelles intra-lexicales allongées (qui ne présentent pas les mêmes propriétés acoustiques notamment). L'exemple 2.5 *supra* présente un énoncé comportant trois hésitations vocaliques (« euh », « euh », « hum »).

De manière générale, les pauses remplies ont été largement étudiées du point de vue de leurs propriétés acoustiques. Shriberg (1999) met également en exergue le lien entre pauses remplies et pauses silencieuses (qui leur succèdent dans le discours). Les pauses remplies présentent une fréquence fondamentale aux contours

plats, au delà de la médiane du locuteur ; leur énergie est stable et tombe en fin d'énoncé (Vasilescu *et al.*, 2008, Kaushik *et al.*, 2010).

Du point de vue de la distinction entre voyelles d'hésitation et voyelles intra-lexicales, Vasilescu *et al.* (2008) montrent que la hauteur des voyelles d'hésitation est en moyenne inférieure à celle des voyelles intra-lexicales.

En accord avec Shriberg (1994), de nombreux travaux (*cf.* par exemple Duez (2001), Candea *et al.* (2005), Vasilescu *et al.* (2008)) mettent en avant l'importance de la durée moyenne des hésitations vocaliques comme facteur discriminant (environ 300 millisecondes (ms) contre une estimation de 80 ms pour les voyelles intra-lexicales et 200 ms pour les syllabes accentuées).

2.2.2.2 Les marqueurs discursifs

EXEMPLE SONORE 2.6 – Les marqueurs discursifs

(i) « **bon** une une misère **quoi hein je veux dire** parce que **bon** euh **encore à la limite** »

(ii) « **bon ben** c'est pas grave **bon ben voilà bon ben** le **ben** nous on est tran- **bon enfin je veux dire** euh **bon** euh on est tranquille »

Ces éléments constituent une catégorie aux frontières floues constituée de petits mots de la parole. Sans briser complètement l'ordre syntagmatique de l'énoncé, il sont optionnels et ce dernier ne se trouve à l'inverse pas brisé par leur absence (c'est-à-dire que leur suppression au sein d'une phrase n'entraîne pas son agrammaticalité). Ils jouent un rôle important au niveau du discours et sont souvent représentés par des mots très courts, tout en intégrant des éléments variés allant de certains adjectifs à emploi interjectif (« bon ») à des syntagmes verbaux à fonction phatique (« tu vois »).

Bove (2008) dresse une brève typologie des marqueurs discursifs, que nous listons *infra*, liste à laquelle l'auteur ajoute tout élément non-régi syntaxiquement et n'entrant pas dans les catégories présentées.

- Connecteurs : « mais », « donc », « aussi », « parce que »… ;
- Particules discursives : « bon », « voilà », « ben », « quoi »… ;
- Phatiques : « tu sais », « tu vois », « hein »… ;
- Régulateurs : « oui », « d'accord », « je vois »… ;
- Certaines locutions : « de toute façon », « en définitive »… ;
- Verbes parenthétiques : « je veux dire », « je précise ».

L'extrait d'énoncé présenté dans l'exemple 2.6 (i) *supra*, au sein duquel 75 % des mots interviennent dans un marqueur discursif, comporte des marqueurs discursifs très variés : des particules discursives (« quoi », « bon ») ; une interjection à fonction phatique (« hein ») ; un syntagme verbal parenthétique (« je veux dire ») ;

un connecteur (« parce que ») ; une locution (« à la limite ») ; enfin un adverbe dont le rôle est clairement discursif ici (« encore »).

Ces deux exemples mettent également en exergue une difficulté liée à l'identification automatique des marqueurs discursifs. En effet, certains syntagmes n'endossent pas toujours le rôle de marqueurs discursifs au sein de l'énoncé : c'est ici notamment le cas, dans l'exemple (i), pour la locution adverbiale « à la limite », qui pourrait tout aussi bien dans un autre contexte avoir le sens d'*arriver au bout de quelque chose* par exemple. C'est aussi et surtout le cas pour l'adverbe « encore », qui en tant que marqueur discursif n'exprime pas un indice de temporalité mais ne garde éventuellement qu'un trait de gradation intensive sans objet.

2.2.2.3 Les amorces

Les amorces ou fragments de mots sont un phénomène très fréquent dans la langue parlée. Pallaud et Henry (2004) proposent de distinguer trois types d'amorces, découlant d'analyses opérées sur des annotations manuelles en corpus (corpus de travail constitué de plus d'un million de mots, cependant avec globalement peu de productions spontanées), selon la continuité de l'énoncé suite à l'énonciation d'une amorce :

- les amorces complétées ;
- les amorces modifiées ;
- les amorces inachevées.

Exemple sonore 2.7 – Les amorces

Amorce complétée :
« c'est pas **im-** euh **imp-** important »

Amorce modifiée :
« tous les numéros que vous pourrez **raccor-** euh rappeler »

Amorce abandonnée :
« et je reçois régulièrement des **prél-** parce que c'était moi »

Cette distinction entre trois types d'amorces prend bien en compte les différents types d'énonciation possibles ; ces derniers sont illustrés dans l'exemple 2.7 *supra*.

Les amorces complétées sont la succession de deux tentatives d'énonciation d'un même mot : en premier lieu ce mot n'est pas énoncé entièrement, puis le locuteur reprend son énonciation pour produire un mot complet. Plusieurs cas de figure sont cependant à envisager, et la reprise complète du mot peut ne pas être immédiate (insertions de pauses silencieuses, d'hésitations vocaliques ou autres mots d'édition entre l'amorce et sa complétion). Ainsi, dans l'amorce complétée présentée dans l'exemple 2.7 *supra*, on repère deux fragments différents du même

mot, « important », ce que l'on peut déduire par l'énonciation du mot entier dans la continuité immédiate de la prononciation de ces fragments « im- » et « imp- ». Non seulement l'amorce complétée est double, mais en plus une hésitation vocalique est susceptible d'être insérée entre les deux amorces.

Dans le cas des amorces modifiées, le locuteur abandonne le mot qu'il est en train de produire pour en produire un autre.

Dans le cas d'amorces abandonnées, l'énoncé est interrompu.

2.2.2.4 Les disfluences d'édition

Les disfluences d'édition (*edit disfluencies*, *speech repairs* ou *speech disfluencies* dans la littérature anglophone), ont un impact lourd sur le déroulement syntagmatique de l'énoncé. Leur fréquence d'apparition, de même que la complexité de leur structure en font des phénomènes à la fois :

- très intéressants à étudier d'un point de vue discursif et dialogique ;
- difficiles à détecter automatiquement ;
- ayant une incidence forte sur le déroulement de l'énoncé.

Les disfluences d'édition sont reconnaissables notamment par la présence de deux éléments plus ou moins longs qui s'entassent sur l'ordre syntagmatique : l'*ébauche* et l'*achèvement* de l'énoncé. De plus, elles embrassent l'ensemble des phénomènes présentés dans cette section (amorces, hésitations vocaliques et marqueurs discursifs), qui sont tous susceptibles d'intervenir dans le cadre d'une disfluence d'édition.

Compte tenu de ces spécificités, nous consacrons la section suivante (section 2.3) à une étude de la structure interne de ces phénomènes, ainsi qu'aux différentes classes de disfluences d'édition que l'on peut distinguer.

2.3 Typologie des disfluences d'édition

Nous l'avons esquissé en section 2.2, les disfluences intègrent de nombreux phénomènes hétérogènes. Cependant, bien que les typologies proposées dans la littérature diffèrent quelque peu, la majorité des études réalisées sur les disfluences prennent pour objet les éléments que nous avons choisis de distinguer – toujours en référence à la définition des disfluences de Blanche-Benveniste *et al.* (1990) – selon leurs effets sur le déroulement syntagmatique de l'énoncé. En nous appuyant essentiellement sur les travaux d'E. Shriberg et sur les différentes typologies factuelles réalisées dans des objectifs d'annotation (travaux de l'équipe DELIC [4] (Piu et Bove, 2007), spécifications du LDC (Strassel, 2004)) nous détaillons

4. Équipe Description Linguistique Informatisée sur Corpus : équipe de recherche de l'Université de Provence à présent rattachée au Laboratoire d'Informatique Fondamentale de Marseille.

ici le schéma de construction des disfluences d'édition (section 2.3.1) ainsi que les classes de disfluences d'édition (section 2.3.2).

2.3.1 Structure interne des disfluences d'édition

Parmi l'ensemble de phénomènes que recouvre la définition générale des disfluences, les disfluences d'édition telles que définies en section 2.3.2 présentent la structure interne la plus complexe. En effet, autant les hésitations vocaliques (« euh ») sont de simples lexèmes, autant les disfluences d'édition font intervenir des segments lexicaux beaucoup plus longs. Cette catégorie de disfluences nécessite donc pour sa compréhension d'une part et son identification d'autre part une modélisation fine de sa structure interne.

Dans le cadre de ses travaux sur les phénomènes d'auto-correction dans le discours, Levelt (1983) analyse leur construction selon trois phases successives :

1. l'interruption de l'énoncé lorsque le locuteur souhaite s'auto-corriger et le segment d'énoncé que le locuteur souhaite corriger ;
2. la construction de la correction, caractérisée par ce que l'auteur nomme de manière générale des termes d'édition *(edition terms)* ;
3. l'auto-correction en soi, c'est-à-dire l'énoncé venant corriger la première partie de cette structure.

La modélisation proposée par Levelt (1983) a ensuite été schématisée par Shriberg (1994) lors de ses travaux sur les disfluences en parole (*speech disfluencies*) : nous adoptons cette schématisation dans ces travaux de thèse et la reproduisons en figure 2.1 *infra*.

Outre le contexte précédent la zone disfluente (contexte amont ou *prior context*) et la continuation de l'énoncé une fois la correction effectuée (continuation de l'énoncé ou *continuation*), les trois éléments constitutifs des disfluences d'édition apparaissent clairement [5] :

- le *reparandum*, soit la zone à corriger ;
- l'*interregnum*, soit la zone de préparation de la correction, initiée ici par le point d'interruption du discours (représenté par le symbole $*$), soit le moment ou le locuteur décide de modifier son énoncé ;
- le *reparans*, soit la zone de correction.

Dans la structure proposée par E. Shriberg, le *reparandum* correspond à la phase 1. de la construction d'une auto-correction telle que définie par W. Levelt, l'*interregnum* à la phase 2. et le *reparans* à la phase 3 (*cf.* énumération des phases *supra*).

5. À partir de ce point et pour tout notre document de thèse, selon les conventions exposées préliminairement à ces travaux (*cf.* conventions de lecture, page xxi), les trois éléments constitutifs de la structure des disfluences d'édition seront mis en exergue par le jeu de couleurs présenté en figure 2.1.

contexte amont / **reparandum** / ∗ **interregnum** / **reparans** / **continuité de l'énoncé**

FIGURE 2.1 – Structure interne des disfluences d'édition selon Shriberg (1999), sur le modèle de Levelt (1983)

Afin d'illustrer la réalisation de cette structure dans un contexte énonciatif réel, nous reprenons dans l'exemple 2.8 *infra* l'énoncé présenté dans la section 2.2.1 (exemple 2.3) pour illustrer la pluralité des phénomènes disfluents (cet énoncé présente cinq disfluences d'édition). On remarque l'absence d'*interregnum* dans cet énoncé. En effet, ce dernier constitue une phase optionnelle.

EXEMPLE SONORE 2.8 – Mise en relief des éléments structurant les disfluences d'édition

« Vous voyez il y a il y a une augmentation une forte augmentation sur le trois kilos. C'est pour ça que maintenant euh en fonction de l'utilisation des personnes des clients on leur dit euh un un six kilos vous avez plus de confort et en prix vous voyez par exemple un six kilos [PAUSE] un six kilos un trois kilos vous coûte 66 82 un six kilos coûte 77 0 8 à l'année »

L'*interregnum* peut donc être construit de quatre manières différentes, illustrées au sein de l'exemple 2.9 *infra*. Celui-ci peut être vide, représenté par une pause silencieuse, par une pause remplie ou encore par un segment d'édition. Les items composant l'*interregnum* sont également appelés *filler words*, auquel cas ils comprennent – en partie ou en totalité selon les études – les marqueurs discursifs, les pauses remplies ayant fonction d'hésitation, et des mots d'édition autres que ces deux premières catégories.

EXEMPLE SONORE 2.9 – Les constructions possibles de l'*interregnum*

Vide : « et ça c'est payé ∅ c'est compris dans les locations de compteurs »

Pause silencieuse : « euh ben le le [PAUSE] le mois »

Pause remplie : « j'ai cons- euh j'ai consulté mon fils »

Segment d'édition : « je vous ai en ben en 10 secondes »

Par ailleurs, nous avons défini en section 2.2.2.4 la présence de deux segments constitutifs des disfluences d'édition : l'ébauche et l'achèvement. En effet, nous définissons la structure des disfluences d'éditions selon deux plans :

- un plan discursif, où l'ébauche représente l'ensemble du segment d'énoncé au sein duquel le locuteur est dans la modification (ce segment peut être mis en relation avec différentes raisons pour lesquelles le locuteur effectue cette modification) et l'achèvement représente la « correction » effective de l'ébauche ; cette distinction est justifiée par les enjeux liés à la détection de disfluences d'édition dans le discours : dans un objectif de nettoyage des données, alors on souhaitera ôter tout élément relevant de l'ébauche des disfluences, afin de ne conserver que l'achèvement.
- un plan lexico-syntaxique, reprenant la structure proposée par E. Shriberg, où le *reparandum* et l'*interregnum* constituent l'ébauche et le *reparans* constitue l'achèvement.

Cette représentation des disfluences est schématisé en figure 2.2 *infra*. Il s'agit de la représentation d'une disfluence d'édition complète : en effet, dans certains cas l'*interregnum* peut être vide, auquel cas l'ébauche n'est constituée que du *reparandum*, et dans d'autres cas l'achèvement peut également être vide.

DISFLUENCE D'ÉDITION		
Ébauche		Achèvement
Reparandum	Interregnum	Reparans
j'ai cons-	*euh*	*j'ai consulté*

FIGURE 2.2 – Représentation de la structure des disfluences d'édition aux niveaux lexico-syntaxique et discursif.

En opposition à cette modélisation linéaire des disfluences d'édition, syntagmatique, Blanche-Benveniste (1997) propose au contraire une représentation de ces phénomènes à la fois sur l'axe syntagmatique et sur l'axe paradigmatique (*cf.* figure 2.3 *infra*, selon un exemple tiré de l'ouvrage de l'auteur), pour mieux montrer la place de chaque élément dans l'élaboration du discours et plus particulièrement les phénomènes de substitutions lexicales :

> « Les énoncés écrits sont produits selon un déroulement linéaire orienté, qui figure l'enchaînement des syntagmes qui se suivent, comme se suivent par exemple sujet, verbe et complément [...] À cet axe syntagmatique (dit aussi axe des *contiguïtés* ou des *proximités*), représenté, selon une ligne horizontale qui se déchiffre dans le sens de la lecture, F. de Saussure, puis R. Jakobson, en ont opposé un autre, l'axe des paradigmes (dit aussi axe des *similarités*). Cet axe des paradigmes, représenté à la verticale, ne correspond à rien qui soit écrit, dans la pratique ordinaire d'un texte écrit. Pour Saussure et Jakobson, les éléments d'un même paradigme ne peuvent jamais survenir en même temps [...] Cette analyse s'applique bien aux productions écrites une fois qu'elles

> ont été corrigées et qu'elles se présentent comme des produits finis. Mais elle convient beaucoup moins aux brouillons de l'écrit et encore moins bien aux productions de la langue parlée [...] Dans l'usage de la conversation, la langue parlée laisse voir les étapes de sa confection : on y trouve des entassements d'éléments paradigmatiques et des allers et retours sur l'axe des syntagmes. »
>
> Blanche-Benveniste (1997)

Plutôt que mises en relation sur une opposition, nous considérons que les deux manières de modéliser la structure interne des disfluences sont complémentaires. En effet, la modélisation d'E. Shriberg (figure 2.1 *supra*) nous semble adéquate dans une perspective de détection et de typage des disfluences (perspective de traitement automatique des langues) alors que la représentation de C. Blanche-Benveniste (figure 2.3 *infra*) nous paraît davantage adaptée à une lecture humaine du texte ; la représentation des textes issus de l'oral et de leurs disfluences à des fins de visualisation pour analyse humaine.

« il fallait avoir le euh
ah zut ah j'arrive plus à trouver le nom
le brevet
le
le diplôme là de secouriste
le brevet de secouriste »

FIGURE 2.3 – Structure interne des disfluences d'édition selon Blanche-Benveniste (1997)

2.3.2 Classes d'événements

Les disfluences d'éditions sont divisées en quatre grandes classes de phénomènes, qui ont elles-mêmes des modalités d'apparition multiples. Les disfluences d'édition consistant en une perturbation de l'ordre syntagmatique, nous les considérons de manière dissociée des phénomènes propres à l'oral spontané, comme les amorces, les hésitations vocaliques et les marqueurs discursifs. En effet, ces phénomènes marqués lexicalement peuvent tout à fait apparaître au sein d'une disfluence d'édition.

Les **répétitions** consistent en la répétition exacte (duplication) d'un morphème, d'un mot complet ou de plusieurs mots, comme illustré dans l'exemple 2.10 *infra*. En effet, cet exemple met en avant la différence de taille lexicale à laquelle peut être soumise une répétition (de la répétition d'une amorce à celle d'un syntagme verbal relativement long).

EXEMPLE SONORE 2.10 – Disfluences d'édition : les répétitions

Répétition d'un morphème : « po- po- »

Répétition d'un mot : « aujourd'hui euh aujourd'hui »

Répétition d'un groupe de mots : « on repart on repart », « qu'ils me rétablissent euh qu'ils me rétablissent »

Contrairement aux répétitions, les **auto-corrections**, qui par ailleurs présentent une structure proche de ces dernières, impliquent soit un changement de sens (et donc de formule) entre l'élément à corriger (morphème, mot ou plusieurs mots) et sa correction, soit une modification d'ordre morphologique pour assurer la bonne formation de la phrase. Des énoncés prototypiques des auto-corrections sont présentés au sein de l'exemple 2.11 *infra*.

EXEMPLE SONORE 2.11 – Disfluences d'édition : les auto-corrections

Correction d'un morphème : « d'o- d'opérateur »

Correction d'un mot : « le euh les », « mais donc »

Correction d'un groupe de mots : « vous allez avoir des factures euh vous allez avoir un prélèvement », « j'habite pas dans la maison j'habitais pas dans la maison »

Les **faux-départs** ont une structure formelle différente des répétitions et des auto-corrections, dans la mesure où il s'agit de l'abandon d'une portion d'énoncé, avec une absence de lien sémantique (à aucun moment le segment de départ n'est partiellement répété pour être corrigé par exemple). Dans ce cas, la disfluence n'est constituée que de l'ébauche, et ne présente pas d'achèvement (le *reparans* est donc vide). Le contenu de cette classe est très variable (*cf.* exemple 2.12 *infra*) et pose des problèmes d'identification. En effet, déjà pour un être humain expert de cette tâche, la distinction entre auto-correction et faux-départ notamment n'est pas toujours évidente, même en ayant à disposition non seulement les transcriptions de la parole mais aussi le signal qui leur est associé.

EXEMPLE SONORE 2.12 – Disfluences d'édition : les faux-départs

« c'est vous êtes assuré »

« il m'a j' avais pas le papier »

« ils m'ont dit que l'électricien avait mis un j'ai un espèce de truc digital »

« ça com- euh c'est la mairie »

« qu'est-ce que c'est pas une nouvelle facture »

Enfin, les **disfluences combinées** (aussi appelées disfluences complexes, *complex disfluencies* dans la littérature anglo-saxonne[6]) sont une succession ou une imbrication des précédentes classes, comme attesté dans les énoncés présentés dans l'exemple 2.13 *infra*.

EXEMPLE SONORE 2.13 – Disfluences d'édition : les disfluences combinées

« je je je »

« vous me donnez vous m- donnez-moi »

« il y a il y a une erreur de euh il y a eu du- il doit y avoir une erreur »

« qui a eu qui ont son qui a son numéro de porta- qui a qui a pris votre numéro de portable »

« il est il a ils ont changé non non j'ai changé attends j'ai changé de ils ont changé le fournisseur »

2.4 Identification des disfluences

Ces travaux de thèse portent sur la détection des disfluences d'édition. Afin d'étudier les différentes méthodes mises en œuvre dans les systèmes de détection existants dans la littérature, nous nous sommes en premier lieu intéressée aux principes généraux de ces systèmes ainsi qu'à la portée de leur détection : les systèmes de détection des disfluences sont-ils tous fondés sur le même principe ? Quels indices utilisent-ils ? Détectent-ils tous les mêmes phénomènes ou certains sont-ils spécialisés dans le repérage de certaines classes de disfluences ? Nous nous

6. Nous préférons au terme de *disfluence complexe* celui de *disfluence combinée*, qui selon nous non seulement fournit une meilleure représentation de ce phénomène mais surtout évite de considérer les autres classes de disfluences d'édition comme des classes *simples*.

focalisons sur les types de disfluences auxquels s'intéressent les systèmes de détection ainsi qu'aux indices qu'ils utilisent. Une bonne connaissance de ces indices nous permet notamment de mieux aiguiller les analyses que nous menons pour caractériser les disfluences dans nos données d'oral spontané conversationnel.

Nous nous concentrons notamment ici sur le choix des descripteurs utilisés par les systèmes de détection en fonction du type de phénomènes qu'ils extraient (quelles disfluences ?). En effet, les différentes recherches menées sur la détection automatique de disfluences ne sélectionnent pas les mêmes phénomènes à identifier. Alors que certains sont très spécialisés, Hokkanen (2001) pour l'autocorrection par exemple, d'autres cherchent à être les plus couvrants et génériques possibles (Liu *et al.*, 2006).

La plupart des études portant sur la détection de disfluences ont pour but d'identifier le plus précisément possible les zones disfluentes pour les retirer avant d'appliquer des tâches de traitement de la parole, et utilisent pour cela des méthodes d'apprentissage automatique. Les systèmes de détection automatique des disfluences sont élaborés selon deux principes :

- ceux qui interviennent en amont des systèmes de reconnaissance automatique de la parole, dans le but notamment d'améliorer ces derniers. Ils sont de fait fondés sur des indices acoustiques exclusivement (Kaushik *et al.*, 2010) ;
- ceux qui interviennent en aval des systèmes de reconnaissance automatique de la parole, pour améliorer la performance des systèmes de TAL en leur fournissant en entrée un texte « classique » et/ou pour faciliter la lecture et la compréhension des documents pour un humain (Kim *et al.*, 2004, Liu *et al.*, 2006).

2.4.1 Tâches de détection

L'identification des disfluences, comme toute tâche visant à identifier un phénomène apparaissant dans des données langagières, peut servir des intérêts scientifiques variés, allant de la compréhension du phénomène d'un point de vue linguistique, théorique (quelles sont les propriétés des disfluences ? quel est leur impact sur la fluidité du discours ?) à l'identification de la localisation exacte des disfluences dans le signal ou les transcriptions orales.

En ce qui concerne la suppression des zones disfluentes, la plupart des études se focalise exclusivement sur la détection du *reparandum* (*cf.* par exemple Qian et Liu (2013)). En effet, nous constatons que même si une très large majorité des travaux cités utilise des annotations de références incluant une identification du *reparans* dans ces données, les méthodes de détection ne sont évaluées que sur la détection des bornes du *reparandum* et parfois de l'ébauche complète (*reparandum* et *interregnum*). Dans notre approche de la tâche de détection, nous envisageons une tâche de détection légèrement différente : en plus de supprimer l'ébauche de

la séquence disfluente, ce qui améliore certainement la lisibilité des transcriptions, nous nous intéressons également à l'effort opéré par le locuteur pour modifier son énoncé : selon ces différences de perspective, la tâche de détection diffère et nécessite la mise en œuvre de méthodes adaptées pour la détection de l'achèvement de la disfluence en plus de celle de l'ébauche.

Il est également notable qu'une très large majorité des études recensées, qui travaillent sur des données en langue anglaise (Kim *et al.*, 2004, Johnson et Charniak, 2004, Georgila, 2009, Georgila *et al.*, 2010, Zwarts et Johnson, 2011, Rasooli et Tetreault, 2013, Qian et Liu, 2013), a été conduite sur le `switchboard` corpus (Godfrey *et al.*, 1997). Constitué de conversations téléphoniques en anglais, ce corpus offre en effet l'accès à une très grande quantité de données enrichies (comprenant des annotations manuelles de référence, en disfluences notamment) [7]. L'omniprésence de ce corpus pour mener des travaux de détection des disfluences d'édition a très certainement un grand impact sur la manière dont ces travaux envisagent de détecter les disfluences, en se basant sur les annotations de références qui y sont présentes.

Ainsi, d'un point de vue des tâches de détection, trois grandes tâches sont en général envisagées et traitées soit de manière conjointe, soit successivement :

- la détection de frontières de phrases (plus ou moins, il s'agit de la détection des frontières de *Sentence-like Units ou Semantic Units*) ;
- la détection des *filler words* ;
- la détection des mots d'édition : est communément entendu par mots d'édition les mots inclus dans l'ébauche de la disfluence (sans l'*interregnum* en général, dont la composition place sa détection dans la tâche précédente de détection des *filler words*).

C'est cette dernière tâche qui concerne directement la détection de disfluences d'édition : en effet, tout en prenant pour acquis la structuration des disfluences d'édition telle que proposée par E.Shriberg, et que nous adoptons également, l'objectif de la détection communément opérée dans la littérature réside en l'identification (et la plupart du temps la suppression) de ce que nous appelons l'ébauche de la disfluence ; les mots inclus dans cette phase, et plus particulièrement au sein du *reparandum* sont regroupés sous le terme de mots d'édition (*edited words*). Dans la mesure où les hésitations vocaliques et marqueurs discursifs, qui composent la plupart du temps l'*interregnum*, sont généralement détectés en amont de la détection des mots d'édition, ils sont eux regroupés sous l'appellation de *fillers*.

L'exemple le plus utilisé pour illustrer les disfluences d'édition est le suivant : « *I want a flight to Boston uh I mean to Denver* », ce que nous pouvons aisément transposer au français par « je veux un vol pour Bordeaux euh je veux dire pour Biar-

7. Large corpus de conversations téléphoniques en langue anglaise (environ 2 400 appels) issues de centres d'appels. Les données ont été recueillies par l'entreprise Texas Instruments, sous le parrainage de la DARPA (*cf. page 18*). Les versions actuellement disponibles sont assemblées et publiées par le Linguistic Data Consortium. Site Web : https://catalog.ldc.upenn.edu/LDC97S62, consulté le 16/01/15.

ritz » (transposition permettant de conserver une structure identique à l'exemple en anglais). Qian et Liu (2013) illustrent bien, à l'aide de cet exemple, comment la tâche de détection est envisagée dans la littérature, en présentant cet exemple de la manière suivante : « je veux un vol {pour Bordeaux}$_{edited}$ {euh je veux dire}$_{filler}$ pour Biarritz ». Nous avons utilisé cet exemple pour dresser un tableau récapitulatif des principaux travaux portant sur la détection de disfluences d'édition dans des données d'oral spontané en français et en anglais (*cf.* tableau 2.1 *infra*).

Travaux	**Étiquetage de référence**								**Corpus**	**Langue**
	pour	Bordeaux	euh	je	veux	dire	pour	Biarritz		
Johnson et Charniak (2004)	rpd	rpd	rpd	rpd	rpd	rpd	rpr	rpr	SWB	EN
Kim *et al.* (2004)	rpd	rpd	F	F	F	F	O	O	SWB	EN
Liu *et al.* (2006)	B-RPD	I-RPD+IP	O	O	O	O	O	O	[LIU]	EN
Bouraoui (2008)	rpd	rpd	O	O	O	O	O	O	[BOU]	FR
Georgila (2009)	B-RPD	IP	O	O	O	O	O	O	SWB	EN
Kaushik *et al.* (2010)	rpd	rpd	O	O	O	O	O	O	[KAU]	EN
Zwarts et Johnson (2011)	rpd	rpd	O	O	O	O	O	O	SWB	EN
Constant et Dister (2010)	rpd	rpd	int	int	int	int	O	O	[CON]	FR
Qian et Liu (2013)	B-rpd	E-RPD	F	F	F	F	O	O	SWB	EN
Rasooli et Tetreault (2013)	rpd	rpd	F	F	F	F	O	O	SWB	EN

TABLEAU 2.1 – Comparaison des principaux travaux menés sur la détection de disfluences d'édition en fonction de la tâche de détection (segment identifié). Corpus : SWB = Switchboard corpus, les codes entre crochets renvoient à des corpus spécifiques décrits dans les références bibliographiques associées aux travaux décrits.

Comparaison des séquences de détection Afin d'établir une comparaison entre les tâches de détections effectuées dans ces travaux, nous avons retenu comme séquence de détection (c'est-à-dire comme jeu d'étiquette associé aux mots composant la disfluence d'édition) uniquement les segments sur lesquels est menée l'évaluation des différentes méthodes. Par exemple, Kim *et al.* (2004) proposent une méthode permettant de détecter la frontière de plusieurs événements : les *Sentence-like Units ou Semantic Units*, les *fillers* et les disfluences d'édition. Dans la mesure où ils présentent des résultats d'évaluation par type de détection et où l'objectif de nos travaux est de comparer des recherches menées sur la tâche précise de détection des disfluences d'édition, nous n'avons retenu dans notre exemple que les étiquettes relevant de cette détection, et ne reportons pas sur le segment « euh je veux dire » les étiquettes relevant de leur seconde tâche de détection, celle des *fillers*.

Nous avons également choisi un jeu d'étiquette commun à ces travaux : `rpd` pour l'identification des mots d'édition présents dans le *reparandum*, `int` pour l'*interregnum* (les mots de l'*interregnum* sont toutefois étiquetés `rpd` lorsqu'ils sont

inclus dans la détection du *reparandum* sans en être distingués, comme c'est le cas pour Johnson et Charniak (2004)), `rpr` pour le *reparans*.

Lorsque cela était précisé dans la description des travaux, nous avons par ailleurs ajouté des préfixes indiquant une étiquette différenciée pour les premiers et/ou derniers mots présents dans une séquence, soit : `B-` pour le premier mot de la séquence (*Begin*), `I-` pour un mot à l'intérieur d'une séquence (*Inside*), `E-` pour le dernier mot de la séquence (*End*). On assigne aux mots hors séquence l'étiquette `O` (*Outside*). Nous avons également reporté une étiquette `F` pour *fillers* lorsque les méthodes de détection s'appuient sur la détection/suppression de ces éléments pour détecter les disfluences d'édition.

Ainsi, Qian et Liu (2013) ont défini des labels très précis pour la détection des régions d'édition :

- `BE` pour le premier mot d'une séquence composée de plusieurs mots ;
- `IE` pour les mots à l'intérieur de la séquence ;
- `EE` pour le dernier mot de la séquence ;
- `SE` pour le mot d'une séquence n'en comprenant qu'un.

Enfin, cette comparaison met en avant l'absence totale d'évaluation menée sur la détection de la totalité du segment composant la disfluence d'édition (son ébauche et son achèvement) et, de ce fait, sur une évaluation de la tâche de détection structurée qui permettrait d'ancrer dans une perspective de traitement automatique la définition et la structuration théorique des disfluences d'édition à laquelle, pourtant, tous se réfèrent.

Comparaison de la portée de la détection Au-delà de la séquence impliquée dans la tâche de détection, l'autre aspect primordial à prendre en compte pour positionner et évaluer les tâches de détection menées dans la littérature concerne la portée de la détection. En effet, nous avons défini quatre classes de disfluences d'édition et souhaitons comparer les principaux travaux menés en détection de ce point de vue.

Le tableau 2.2 *infra* présente le résultat de cette comparaison et permet de se rendre compte que plus la classe de disfluence d'édition est complexe (nous en avons eu un aperçu lorsque nous avons défini ces classes en section 2.3), moins elle est prise en compte dans les tâches de détection. En effet, alors que la totalité des études s'attache à identifier les répétitions, seuls Johnson et Charniak (2004), Liu *et al.* (2006), Georgila (2009), Zwarts et Johnson (2011) et Qian et Liu (2013) se sont attachés à évaluer l'identification de disfluences combinées.

Et même au sein de ces travaux la prise en compte des disfluences combinées est parfois à relativiser. Par exemple, Georgila (2009) considère parmi les disfluences combinées uniquement celles qui sont constituées d'une *succession* de répétitions ou d'auto-corrections, et pas d'*imbrications* comme cela peut pourtant être le cas.

	Répétitions	Auto-corr.	Faux-départs	Dis. combinées
Johnson et Charniak (2004)	✓	✓	✓	✓
Kim *et al.* (2004)	✓	✓	✓	
Liu *et al.* (2006)	✓	✓	✓	✓
Bouraoui (2008)	✓	✓		
Bove (2008)	✓	✓		
Georgila (2009)	✓	✓	✓	✓
Kaushik *et al.* (2010)	✓			
Zwarts et Johnson (2011)	✓	✓	✓	✓
Constant et Dister (2010)	✓	✓		
Qian et Liu (2013)	✓	✓	✓	✓

TABLEAU 2.2 – Comparaison d'études sur la détection de disfluences en fonction des classes de disfluences d'édition prises en compte (portée de la détection).

2.4.2 Méthodes de détection

De nombreuses approches ont été testées pour détecter automatiquement les événements disfluents. Johnson et Charniak (2004) et Zwarts et Johnson (2011) ont utilisé une approche basée sur l'étiquetage du canal bruité pour modéliser les réparations vocales et identifier les mots d'édition. Liu *et al.* (2006) ont de leur côté montré que les champs aléatoires conditionnels (CRF) donnent de meilleures performances que d'autres approches comme les modèles de Markov cachés ou l'entropie maximale.

Les CRF ont également été utilisés par Georgila (2009), avec une étape de post-traitement appliquant des patrons communs à la formation des disfluences sur les sorties de leur système pour améliorer les résultats. Les résultats à l'état de l'art ont récemment été atteints en opérant une tâche conjointe de *parsing* et de détection des disfluences par Rasooli et Tetreault (2013), qui ont utilisé un analyseur déterministe basé sur les états de transition, et pour la détection des *reparandum* par Qian et Liu (2013), qui ont utilisé un classifieur (réseaux de Markov à Maximum de Marge ou *Max-Margin Markov Networks*, (Taskar *et al.*, 2004)) utilisant une combinaison d'indices acoustiques et linguistiques (méthode avec laquelle ils obtiennent une f-mesure [8]) de 84,1 %.

Liu *et al.* (2006) mènent une étude comparative d'approches pour la détection de disfluence d'édition en étudiant une approche générative basée sur les modèles de Markov cachés (ou HMM, pour *Hidden Markov Model*, modèle statistique largement utilisé en TAL) et deux modèles conditionnels : un modèle à maximum d'entropie (Maxent) et une approche à base de champs aléatoires conditionnels (CRF). L'objectif final de cette étude est l'amélioration des résultats obtenus lors de l'étape

8. Pour une définition de cette mesure, voir la section 5.5.1.2 (page 111) dédiée aux mesures d'évaluation, au sein du chapitre 5.

de reconnaissance automatique de la parole. Ils se sont cependant concentrés exclusivement sur l'identification du *reparandum*, comme une grande majorité des autres études, précisant que l'information concernant le *reparans* n'était pas disponible dans leurs données d'application (absence d'annotation). Leurs résultats montrent qu'une approche à base de CRF donne les meilleurs résultats.

2.4.3 Indices utilisés pour la détection de disfluences

Les systèmes développés pour la détection de disfluences peuvent faire usage d'indices acoustiques (Audhkhasi *et al.*, 2009, Kaushik *et al.*, 2010), d'indices lexicaux (Snover *et al.*, 2004) ou d'une combinaison des deux types d'indices (Kim *et al.*, 2004, Liu *et al.*, 2006). Nous nous sommes particulièrement intéressée aux systèmes utilisant des indices mixtes, dans la mesure où nous faisons l'hypothèse qu'il s'agit de la méthode la plus riche et la plus précise pour identifier les disfluences, ce phénomène étant caractérisé à la fois du point de vue du signal de parole et de la chaîne lexicale qui lui est associée.

Depuis les travaux de Labov (1966) mettant en avant la présence d'un signal acoustique pour marquer le lieu de l'auto-correction, aussi appelé signal d'édition acoustique au moment de l'interruption selon Shriberg (1999) citant Hindle (1983), de nombreuses méthodes s'appuient sur des paramètres acoustiques exclusivement.

Shriberg (1999) analyse les caractéristiques acoustiques découlant des disfluences (en anglais américain) dans la perspective de faire émerger des descripteurs pertinents pour le traitement automatique de ces phénomènes. Cette étude montre notamment que les syllabes qui précèdent immédiatement le point d'interruption sont particulièrement longues pour certaines disfluences, tout en mettant en avant que même dans ce cas l'intonation du mot est préservée. Audhkhasi *et al.* (2009) proposent un algorithme de détection des pauses remplies fondé sur les fréquences des deux premiers formants. Kaushik *et al.* (2010) étendent les travaux de Audhkhasi *et al.* (2009) à la détection des répétitions en proposant un nouvel algorithme utilisant d'autres descripteurs acoustiques ; en appui aux résultats de Shriberg et Stolcke (2002), ils utilisent des indices prosodiques uniquement, et plus précisément : la durée des voyelles, la fréquence fondamentale, des paramètres basés sur le spectre et les quatre premiers formants. En effet, contrairement à Audhkhasi *et al.* (2009), ils affirment qu'une analyse basée sur les deux premiers formants uniquement n'est pas assez robuste.

Les travaux présentés dans Liu *et al.* (2006) décrivent davantage leur méthode de détection des Sentence-like Units ou Semantic Units. Concernant les descripteurs acoustiques, les auteurs utilisent des informations s'appuyant sur la durée (des mots, des pauses), la fréquence fondamentale, l'énergie et les pauses. Ils utilisent également des informations non-prosodiques, que nous qualifions d'indices discursifs, comme le genre du locuteur et le changement de locuteur. Concernant

les descripteurs lexicaux, outre l'utilisation des formes lexicales et des parties du discours [9], les auteurs prennent en compte les indicateurs suivants : présence de marqueurs du discours, hésitations vocaliques et *backchannel*, classe sémantique des mots et cooccurrences (entre mots et avec les événements voisins).

Contrairement à la plupart des recherches sur l'identification des répétitions, Kaushik *et al.* (2010) utilisent uniquement des indices acoustiques, basés sur les formants et des informations spectrales. Après extraction des segments concernés (même méthode que pour les pauses remplies), ils identifient les fréquences des formants et anti-formants [10] puis calculent la distance euclidienne entre les logarithmes du spectre du LPC pour chaque paire de segment.

Kim *et al.* (2004) et Liu *et al.* (2006) présentent un système de détection conçu dans le cadre de la campagne d'évaluation NIST RT-04F [11]. Ils développent un système de détection en deux étapes : après avoir prédit certains événements (comme les frontières de phrases (*sentence unit boundaries*) aux frontières des mots (avec un arbre de décision appliqué aux transcriptions), ils détectent les pauses remplies et les disfluences d'édition à partir de ces résultats (enrichis d'autres informations lexicales). Un post-traitement permet d'éliminer les segments présentant une prédiction de point d'interruption sans détection de disfluences et inversement. Les paramètres lexicaux utilisés sont les parties du discours (obtenus grâce à l'étiqueteur morpho-syntaxique décrit dans Ratnaparkhi (1996)) et l'indication de la présence de marqueurs du discours à la frontière droite du mot. Les auteurs montrent que les paramètres lexicaux sont de moindre utilité que les paramètres prosodiques dans la mesure où ils appliquent leur système de détection à des transcriptions automatiques. Les auteurs travaillent également sur des transcriptions manuelles pour supprimer les disfluences en aval du système de reconnaissance automatique de la parole dans le but d'améliorer la lisibilité humaine et l'entrée des systèmes de TAL. En complément des indices lexicaux, ils utilisent également des indices discursifs comme les marques de début et de fin d'un tour de parole et de parole superposée, la position du mot au sein du tour de parole.

9. *Part-Of-Speech* (POS) ou Parties du discours : catégories grammaticales assignées à chaque mot de la langue. Selon le détail donné par les outils les assignant de manière automatique (communément appelés outils d'étiquetage morpho-syntaxique), elles peuvent inclure en sus de la catégorie grammaticale les informations de genre, de nombre, de personne, *etc.*.

10. Les anti-formants sont définis comme suit dans le lexique linguistique de l'Université d'Utrecht : « L'articulation des sons nasaux crée des anti-résonances dans le tractus vocal. Ces anti-résonances ou anti-formants sont des régions de fréquence dans lesquelles les amplitudes du signal source sont atténuées car les cavités nasales absorbent l'énergie de l'onde sonore. Les effets de ces anti-formants sont davantage marqués pour les consonnes nasales que pour les voyelles nasales ou nasalisées parce que les consonnes sont articulées avec une occlusion complète de la cavité buccale » (notre traduction). Source Web : http://www2.let.uu.nl/UiLOTS/Lexicon/zoek.pl?lemma=Antiformant&lemmacode=1142 .

11. Site Web : http://www.itl.nist.gov/iad/mig//tests/rt/2004-fall/index.html ;
Plan d'évaluation : http://www.itl.nist.gov/iad/mig/tests/rt/2004-fall/docs/rt04f-eval-plan-v14.pdf.

Liu *et al.* (2006) utilisent un certain nombre de descripteurs acoustiques pour l'apprentissage de l'arbre de décision permettant l'identification de mots d'édition : durée, fréquence fondamentale (statistiques sur un mot et différences entre les statistiques de deux mots voisins), énergie sur un mot, durée du silence suivant un mot. Les paramètres de fréquence fondamentale et d'énergie sont normalisés en fonction de chaque locuteur.

2.4.4 Détection des disfluences pour le français

Concernant les recherches portant sur des données en français, l'état de l'art n'est pas aussi riche et développé que celui concernant l'anglais ; cet état de fait est essentiellement dû à la quasi absence de corpus annoté disponible pour entraîner et tester des systèmes de détection. Les études portant sur les disfluences sont de fait largement descriptives, et davantage tournées vers les caractéristiques inhérentes à ces phénomènes que vers des tâches de détection automatique.

Il existe toutefois quelques études portant sur la détection automatique de disfluences, proposant des modèles à base de règles syntaxiques, comme le `Distagger`, décrit dans Constant et Dister (2010), qui s'attache au repérage des amorces, des hésitations vocaliques « euh », des auto-corrections et des répétitions, ou le module d'analyse syntaxique des disfluences (répétitions et auto-corrections) proposé par Bove (2008). D'autres études sur le français présentent des méthodes de détection appliquées conjointement à et dans le flux d'une autre tâche principale de classification (Peshkov *et al.*, 2013) ou appliquées à des domaines très spécialisés (Bouraoui et Vigouroux, 2009). Dans cette dernière étude, la détection automatique est faite en utilisant un patron de modélisation des disfluences (répétitions et auto-corrections), appliqué à des données de communication aéronautiques. Peshkov *et al.* (2013) présentent quant à eux une approche basée sur les transcriptions pour localiser des événements disfluents (pauses remplies et fragments de mots) sur des données conversationnelles. Toutes ces études ont été menées sur des transcriptions manuelles.

Nous détaillons ici plus précisément les méthodes incluant une détection d'événements relevant des disfluences d'édition, dans la mesure où nous nous intéressons dans nos travaux à cette tâche en particulier. Nous discutons notamment les choix de détection en lien avec la définition et la typologie des disfluences d'édition (définies en sections 2.2 et 2.3), afin de mieux éclairer les résultats avancés dans l'évaluation de ces systèmes.

2.4.4.1 Bove (2008) : détection non-structurée de répétitions et d'auto-corrections immédiates par analyse syntaxique de surface

Description de l'approche et de la méthode Bove (2008) propose un module d'analyse syntaxique de surface de certains phénomènes disfluents, basé sur

l'identification de séquences de n-grammes et utilisant des indices linguistiques uniquement. Deux types de patrons ont été développés dans cette étude : des patrons « simples » utilisant la forme des mots et des patrons « mixtes » utilisant en plus de cette information les parties du discours et lemmes (produits par le `Treetagger`). La méthode proposée consiste en l'identification de répétitions et d'auto-corrections immédiates (détection de segments « répétable – répété » ou « séquence d'origine – séquence modifiée ») sans prise en compte réelle de l'*interregnum* : en effet, afin de pouvoir identifier des séquences continues, certains phénomènes d'oral spontané sont supprimés dans une étape antérieure afin de rapprocher dans l'énoncé les séquences à détecter (tout en étant réinjectés par la suite afin de reconstituer l'énoncé) [12]. Les faux-départs et disfluences combinées ne sont pas considérés.

Dans la mesure où l'objectif de la détection des disfluences dans cette étude est de catégoriser des *chunk* [13] en segments disfluents et non-disfluents, les zones de disfluences ne sont pas structurées lors de la tâche de détection : la présence d'une disfluence est indiquée dans le segment, mais ni l'ébauche ni l'achèvement de la disfluence ne sont identifiés. Ce n'est pas non plus le cas de l'*interregnum* qui, comme nous l'avons décrit, doit être absent de l'énoncé pour que la répétition ou l'auto-correction puisse être détectée.

En évaluant sa méthode sur un corpus de test issu du CRFP, Bove (2008) obtient une f-mesure de 94 % pour le repérage des répétitions et de 67 % pour le repérage des auto-corrections. La baisse du rappel concernant la détection des auto-corrections (-40 points) est expliquée par l'auteur par la variété dans la structure syntaxique des auto-corrections ; par exemple, un segment tel « nous avions nous n'étions pas très tranquilles » n'est pas repéré par le module proposé ici. L'évaluation de la méthode proposée par Bove (2008) prend en compte les erreurs de détection de répétitions intensives ou emphatiques.

Résumé de la tâche de détection Identification hybride des segments de disfluences d'édition (*reparandum* et *reparans* avec suppression préalable et non-contrôlée de l'*interregnum*) ne prenant pas en compte toutes les classes de disfluences d'édition (répétitions et auto-corrections immédiates seulement). Aucune structuration du segment disfluent n'est définie (l'ébauche et l'achèvement de la disfluence sont confondus). Les amorces ne sont pas intégrées au processus de détection. La tâche est évaluée sur des transcriptions manuelles uniquement.

12. L'efficacité de la méthode de suppression des hésitations vocaliques et des marqueurs discursifs dans la totalité de l'énoncé afin de simplifier les schémas d'apparition des disfluences d'édition n'a pas fait l'objet d'une évaluation.

13. Les *chunks* sont des segments intra-phrastiques résultant d'une analyse syntaxique de surface qui vise à identifier les différents constituants de la phrase (étape dite de *chunking*).

2.4.4.2 Constant et Dister (2010) : détection semi-supervisée et non-structurée de répétitions et d'auto-corrections immédiates par application itérative de patrons lexicaux

Description de l'approche et de la méthode Le `Distagger` a été développé et testé sur un jeu de données issu du corpus VALIBEL et composé de quatre transcriptions (totalisant 34 624 mots), disjoint du corpus d'évaluation. L'objectif de la détection est d'identifier l'ébauche des disfluences d'édition afin de nettoyer les énoncés de parole spontanée.

La méthode mise en place par les auteurs consiste en l'application itérative d'un patron simple, basé sur la forme des mots et leur appartenance à une même classe, selon les trois étapes suivantes : *(i)* identification d'une séquence *w* de mots, *(ii)* identification d'une séquence *I* d'insertions (pauses silencieuses, marqueurs discursifs) et *(iii)* identification d'une séquence optionnelle *c* de correction de *w*. Cette dernière séquence est validée soit par la présence du mot ou de la séquence de mots *w* (répétition exacte), soit par l'appartenance de *c* à la même « classe d'équivalence » que *w*, soit par le fait que *c* soit un mot dont *w* est le préfixe.

Nous qualifierons la méthode proposée par Constant et Dister (2010) de méthode semi-supervisée dans le sens où elle nécessite deux ressources externes construites par l'utilisateur de l'outil. En effet, afin d'identifier les hésitations vocaliques, l'utilisateur doit renseigner leurs différentes réalisations possibles dans son corpus : cette étape n'est pas très contraignante dans la mesure où les hésitations vocaliques en français sont peu nombreuses. En revanche, l'utilisateur doit également définir les classes d'appartenances des mots (par exemple {le, la, les}), ce qui nous semble beaucoup plus coûteux en temps tout en ayant *a priori* beaucoup d'effet sur les performances effectivement obtenues par l'outil.

Concernant l'identification des différentes classes de disfluences d'édition considérées, le patron d'application est générique (il détecte de manière indifférenciée les répétitions et les auto-corrections) et le typage est effectué dans un second temps. Dans leur description des auto-corrections, les auteurs précisent que l'un des traits morphologiques de l'élément répété varie (par exemple « le journalisme et puis euh le les études de journalisme »). Qu'en est-il des auto-corrections plus complexes, avec davantage de variation ? En effet, les auto-corrections immédiates contiennent souvent beaucoup plus de variation entre le *reparandum* et le *reparans* que la seule variation d'un trait morphologique, comme en attestent les types d'auto-corrections présentés dans l'exemple 2.14 *infra*. Ces exemples, issus d'un corpus d'oral spontané conversationnel, sont-ils pris en compte par le *Distagger* ?

Les performances avancées lors de l'évaluation du `Distagger` sont toutefois très élevées : 98 % de f-mesure pour l'identification des répétitions et 73 % pour celle des auto-corrections, et laissent penser à une bonne couverture des phénomènes envisagées, ce qui nous parait assez contradictoire avec la complexité effective de certaines auto-corrections et la rigidité du patron appliqué. L'évaluation a été faite sur un corpus de deux transcriptions (22 476 mots) issues de la base de

données VALIBEL, pour lequel la référence a été construite semi-automatiquement (pré-annotation avec le `Distagger` et validation manuelle : correction d'erreurs et annotation des disfluences manquantes). Concernant les répétitions intensives, la détection (erronée donc) d'une répétition intensive n'est pas considérée comme une erreur.

EXEMPLE SONORE 2.14 – Différents types d'auto-correction

« je sais pas du tout j'ai pas j'ai même pas de document »

« non il y a pas de normalement il y a pas de »

« on paie nos moi je paie mes factures »

« j'ai payé fini de payer »

« j'ai jamais euh euh je suis pas »

Résumé de la tâche de détection Identification complète des disfluences d'édition (*reparandum*, *interregnum* et *reparans*) ne prenant pas en compte toutes les classes de disfluences d'édition (répétitions et certains types d'auto-corrections immédiates seulement). Aucune structuration du segment disfluent n'est définie (l'ébauche et l'achèvement de la disfluence sont confondus). Les amorces ne sont pas intégrées au processus de détection (elles sont détectées par ailleurs). La tâche est évaluée sur des transcriptions manuelles uniquement.

2.5 Discussion

Nous avons dans cet état de l'art défini les disfluences en prenant en compte la littérature. Nos choix définitoires et typologiques sont fonction de l'objectif poursuivi : savoir caractériser les disfluences dans des données orales conversationnelles pour mieux modéliser le dialogue et en extraire de l'information. Nous avons ainsi choisi de distinguer les disfluences d'édition des autres phénomènes d'oral spontané compte tenu de la complexité de leur structure interne : alors que les hésitations vocaliques, marqueurs discursifs, et amorces (généralement regroupés sous l'appellation de *fillers*) ne nécessitent pas de sous-structuration car ils correspondent à des éléments lexicaux bien identifiés, souvent des lexies simples ou des unités phraséologiques relativement figées, les disfluences d'édition présentent une structure beaucoup plus complexe et prennent davantage de place sur les axes syntagmatique et paradigmatique. Notre étude des disfluences a glissé vers le cadre analytique plus large des méta-données structurelles, auxquelles appartiennent les éléments qu'il est souvent nécessaire de repérer pour mieux modéliser

les données orales (et le dialogue), dans une perspective de traitement automatique.

Dans la suite logique de cette caractérisation, nous avons synthétisé l'état de la recherche portant sur les systèmes de détection des disfluences : quels phénomènes détectent-ils ? quel est leur lien avec l'étape de Reconnaissance Automatique de la Parole (RAP) ? quels indices utilisent-ils pour détecter des événements disfluents ? Cette analyse, notamment sur les indices utilisés, permet de mieux orienter les analyses descriptives sur les disfluences. Nous avons également étudié les méthodes utilisées par ces systèmes, en nous interrogeant notamment sur les aspects suivants : quelle est la performance de ces systèmes ? cherchent-ils à établir des frontières d'événements disfluents ou opèrent-ils une classification de ces disfluences (dans le but de les utiliser pour d'autres tâches) ?

Cette vue d'ensemble des principaux travaux menés sur la détection automatique de disfluences d'édition permet de dresser un constat de l'état de l'art sur plusieurs points :

1. Tout d'abord, malgré un large consensus établi autour de la définition des disfluences d'édition, des phénomènes qui lui sont attachés (répétitions, auto-corrections, faux-départs et disfluences combinées), et de la reconnaissance de leur structure interne (*reparandum*, *interregnum* et *reparans*), les terminologies associées à ces caractéristiques sont encore plurielles et entraînent un manque de clarté sur les phénomènes réellement étudiés dans la littérature (que ce soit du point de vue de leur analyse comme de leur détection) ;
2. Concernant les travaux menés sur la détection automatique des disfluences d'édition, il n'y a à l'heure actuelle pas de système qui prenne en compte ce phénomène dans sa globalité :
   - premièrement, les travaux sont focalisés sur la détection de l'ébauche des disfluences, ce qui montre, malgré le positionnement de nombreuses études qui reconnaissent le rôle de ces phénomènes au sein du discours, que l'intérêt est toujours mis sur leur suppression ;
   - deuxièmement, la détection de ce phénomène n'est pas arrivée à maturité : en témoigne la prépondérance de travaux évaluant leur méthode sur des transcriptions manuelles exclusivement ;
   - troisièmement, autant la détection des répétitions fait totalement partie des tâches de détection, autant ce n'est pas le cas pour des classes plus complexes à traiter, comme les faux-départs ou les disfluences combinées. Rares sont les études qui justifient cette prise de position ;
3. Du point de vue des méthodes mises en place pour la détection automatique des disfluences d'édition, il y a une grande différence entre les travaux menés sur l'anglais et ceux menés sur le français :
   - sur l'anglais, le SWITCHBOARD corpus offre une grande quantité de données d'oral spontané conversationnel annotées en disfluences d'édition :

ce corpus constitue non seulement une solide référence pour évaluer les systèmes de détection, mais permet le développement de méthodes d'apprentissage automatique du fait de sa grande taille et surtout la confrontation des résultats obtenus par les différentes méthodes développées. Ces dernières sont donc majoritairement tournées vers des systèmes de type classifieurs (qui nécessitent une grande quantité de données annotées) et utilisent plus ou moins les mêmes corpus d'entraînement, de développement et de test pour leur évaluation ;

- sur le français, *a contrario*, les études sont très peu nombreuses, tout comme les corpus. Il n'existe pour l'instant pas d'équivalent au SWITCHBOARD corpus qui permettrait *(i)* le développement de méthodes nécessitant un apprentissage sur beaucoup de données et *(ii)* la confrontation des résultats obtenus par les différents systèmes développés. Ces derniers fonctionnent donc essentiellement sur l'application de patrons d'analyse morpho-syntaxique. De plus, l'absence de consensus sur l'annotation des disfluences d'édition implique un manque d'homogénéité théorique sur les tâches de détection, ce qui accroît d'autant plus la difficulté de comparer les travaux entre eux.

4. À propos des indices utilisés pour détecter les phénomènes d'oral spontané et plus particulièrement les disfluences d'édition : les travaux présentant des méthodes à base de règles utilisent des indices lexicaux uniquement ; ceux proposant des méthodes de type classifieurs ou étiquetage séquentiel s'accordent à dire que l'utilisation d'indices mixtes (linguistiques, discursifs et acoustico-prosodiques) donne les meilleurs résultats.
5. Sur la tâche de détection des *reparandum* toutes classes de disfluences confondues les résutlats à l'état de l'art sont atteints par Qian et Liu (2013), avec une f-mesure de 84,1 % (performance obtenue sur des données de test issues du SWITCHBOARD corpus ; sur le français, la méthode à base de règles proposée par Constant et Dister (2010) donne les meilleurs résultats, sur une tâche de détection des *reparandum* prenant en compte une partie seulement des classes de disfluences, avec une f-mesure à 95,5 % (performance obtenue sur un corpus de test composé de deux transcriptions issues de la base de données VALIBEL).

Chapitre 3

Corpus de données conversationnelles en français

Sommaire

NOUS avons discuté au sein du chapitre 2 les aspects théoriques et définitoires des *disfluences* dans la langue parlée. Nous avons également dressé un état de l'art des méthodes et outils visant la détection automatique de ces tâches, en portant un intérêt particulier à la définition précise des tâches de détection évaluées. En complément de cette étude, nous souhaitons avec ce chapitre 3 mettre en valeur la construction de corpus de langue parlée en français, en présentant d'une part les corpus de grande envergure utilisés par des travaux portant sur l'oral spontané, et d'autre part les deux corpus composés de conversations menées en centres d'appels, ces dernières étant le matériau sur lequel s'appuient ces travaux de thèse. Au-delà de la présentation de notre corpus de travail, nous souhaitons en effet mettre en avant un souci de comparaison de ces données et d'évaluation de nos travaux par rapport à l'existant.

3.1 Traitement Automatique des Langues naturelles, linguistique de corpus et positionnement académique : comparer les données pour évaluer les travaux de recherche

Étudier et traiter un objet linguistique de manière empirique suppose l'utilisation de corpus sélectionnés en adéquation avec les phénomènes analysés et les objectifs de cette étude. Avec l'avènement de la linguistique dite de corpus et du Traitement Automatique des Langues naturelles (TAL) à grande échelle, la nécessité se fait de disposer de ressources répondant à des critères précis et dans l'idéal construites dans le respect de nombreuses exigences, de manière à ce que les analyses menées sur ces ressources touchent au plus près les faits de langues tels qu'ils se produisent en contexte réel. Ainsi, l'heure n'est plus à la présentation de théories linguistiques ou de méthodes automatiques associées à une petite poignée d'exemples artificiellement construits par le chercheur, mais à la génération et à la validation d'hypothèses solidement étayées par de nombreux cas de figures.

Concernant les travaux menés en linguistique et en TAL sur la langue parlée, les bases de données sont nombreuses et illustrent des styles de parole et des situations de communication variées. Cependant, comme souligné notamment par Nguyen et Adda-Decker (2013), elles présentent un caractère encore trop hétérogène et épars. L'étude citée ci-dessus ainsi que Cappeau et Seijido (2005) pour la Délégation Générale à la Langue Française et aux Langues de France [1] fournissent un bilan de ces nombreuses ressources orales. Pour nos travaux sur l'analyse et la détection de phénomènes d'oral spontané – et plus particulièrement de disfluences d'édition – nous nous concentrons sur un cadre énonciatif bien particulier : les

1. Ou DGLFLF. Site Web : http://www.dglf.culture.gouv.fr/, consulté le 16/01/15.

conversations téléphoniques menées en français et en centres d'appels, par le biais d'un corpus élaboré au sein de l'entreprise EDF. Mais parce que nous ne souhaitons pas limiter la portée de nos travaux à ce corpus et d'autant plus car ce dernier n'est pas mis à disposition de la communauté scientifique, nous avons pour objectif, dans la mesure du possible, de nous confronter à la communauté scientifique en comparant ces données à des corpus similaires et/ou contenant des annotations, faits langagiers, *etc.* proches des analyses menées pour cette thèse et/ou sur lesquels ont été menés des travaux sur l'oral spontané.

Dans cette perspective, nous décrivons au sein de la section 3.2 les principaux corpus utilisés pour des travaux portant sur l'analyse de la langue parlée ; en section 3.3, une attention particulière est portée aux corpus enregistrés dans le contexte des centres d'appels : le corpus issu du projet DECODA (conversations issues des centres d'appels de la RATP) et le corpus issu du projet Infom@gic (conversations issues des centres d'appels d'EDF) sont en effet les seuls corpus de grande taille constitués sur les conversations en centres d'appels ; le corpus EDF est l'objet de nos travaux sur l'analyse et la détection de disfluences d'éditions en contexte conversationnel. La section 3.4 est dédiée à une discussion sur ces corpus et leurs spécificités.

3.2 Corpus d'oral spontané en langue française

Nous listons ci-dessous un nombre de corpus proches des données que nous étudions, tant en caractéristiques qu'en objectifs d'analyse. Il s'agit de corpus de référence homme-homme et homme-machine en français.

Les corpus présentés dans cette section ont été sélectionnés car ils ont fait l'objet de travaux récents en analyse de la langue parlée. Nous les décrivons ici de manière succincte, en y liant les études portant sur *(i)* de manière globale, la caractérisation de l'oral spontané, souvent en comparaison avec une parole préparée (quelles sont les caractéristiques acoustico-prosodiques et/ou lexicales de l'oral spontané ?) et *(ii)*, plus précisément, la définition, la caractérisation ou l'identification de phénomènes d'oral spontané (en lien étroit avec les phénomènes disfluents : marqueurs discursifs, hésitations vocaliques, disfluences d'édition, *etc.*).

3.2.1 Le CRFP : Corpus de Référence du Français Parlé

Le « Corpus de référence du Français parlé » (CRFP), décrit dans DELIC (2004), est le résultat d'un vaste projet de constitution d'un corpus de référence pour le français qui s'est déroulé entre 1998 et 2002, financé par la délégation générale à la langue française et aux langues de France. Ce corpus contient 36 heures de parole pour des données orales hétérogènes, regroupées selon trois situations d'enregistrement : parole privée, publique et professionnelle.

Du fait de sa mise à disposition de la communauté scientifique, ce corpus a été utilisé dans de nombreuses recherches, notamment sur les disfluences. Campione (2001) en a par exemple tiré un corpus de parole spontanée de 54 minutes pour son étude sur les relations entretenues entre syntaxe et prosodie notamment. Henry *et al.* (2004) ont également utilisé ce corpus pour leurs travaux sur les répétitions et les pauses silencieuses et remplies, en sélectionnant cependant uniquement des segments monologiques – même lorsqu'il s'agit de réponses aux questions d'un enquêteur. Il s'agit également du corpus d'étude de Piu et Bove (2007) pour leurs travaux sur l'annotation des disfluences (répétitions, auto-corrections, amorces, inachèvements et disfluences combinées). Les auteurs proposent un schéma d'annotation XML[2] de ces phénomènes fondé sur un principe de « mise en grille » basé sur le modèle de représentation des disfluences de Blanche-Benveniste (1997) (présenté en section 2.3.1). Ces annotations ont permis le développement d'un module d'identification des zones de répétitions et d'auto-corrections immédiates proposé par Bove (2008).

Bien que Campione (2001) ait utilisé le CRFP pour la réalisation d'études prosodiques, Nguyen et Adda-Decker (2013) pointent la difficulté à produire des analyses phonétiques et phonologiques fiables sur ces données du fait de leur mauvaise qualité d'enregistrement.

3.2.2 RITEL : un corpus de dialogues homme-machine dans un système de question/réponse

Le corpus RITEL (« Recherche d'information par téléphone »), décrit dans Rosset et Petel (2006), est un corpus de dialogues homme-machine orienté question/réponse en domaine ouvert. Sa collecte a constitué la première étape du projet RITEL[3]. L'objectif de ce projet, tel que décrit dans Rosset *et al.* (2005), était de « réaliser un système de dialogue homme-machine permettant à un utilisateur de poser oralement des questions, et de dialoguer avec un système de recherche d'information généraliste ».

À partir d'une liste de 300 exemples de questions, treize personnes ont été invitées à interroger le système RITEL (par le biais de la première plate-forme développée pour le système, décrite dans Galibert *et al.* (2005)) : cette collecte a généré 6 heures de parole, soit 5 360 énoncés répartis dans 582 dialogues. Bien que RITEL soit un corpus de dialogues homme-machine, il nous intéresse tout particulièrement sur deux aspects :

1. l'enrichissement du corpus avec des annotations en thématiques (au niveau de l'énoncé) et en entités (notamment en entités nommées) ;

2. *Extensible Markup Langage*, format standard de structuration.
3. Site Web : http://ritel.limsi.fr/realisation.html/, consulté le 16/01/15.

2. les travaux qui ont été réalisés en appui sur ces données sur le rôle joué par les marqueurs discursifs (Vasilescu *et al.*, 2010b) et les hésitations vocaliques (Vasilescu *et al.*, 2010a).

3.2.3 ESTER : enregistrements d'émissions radiophoniques

Le corpus ESTER (« Évaluation des systèmes de transcription enrichie d'émissions radiophoniques »), décrit dans Galliano *et al.* (2006), fait partie de la campagne d'évaluation éponyme [4]. Il est composé d'environ 90 heures d'émission radiophoniques et a bénéficié d'une annotation en entités nommées.

Adda-Decker *et al.* (2009) ont mené sur une partie du corpus ESTER (13h de parole, 165 000 mots graphiques) une étude visant à localiser les frontières des mots à l'oral, *via* une analyse précise d'indices prosodiques. Les méthodes mises en place pour faire émerger des régularités sur le contour prosodique sont très intéressantes. Nos propres travaux s'inscrivent dans la lignée de ce type d'analyse, mêlant pour l'analyse indices prosodiques et traits morpho-syntaxiques.

3.2.4 Le NCCFr : Nijmegen corpus of casual french

Le « Nijmegen corpus of casual french » [5] (NCCFR) est constitué de conversations informelles entre amis et bénéficie d'un double enregistrement audio/vidéo. Il est décrit dans Torreira *et al.* (2010). Ce corpus est divisé en trois parties : les deux premières parties sont des conversations libres et pour la troisième partie les participants étaient invités à fournir une réponse commune sur des questions de culture générale.

Ce corpus, annoté manuellement, a bénéficié d'annotations en « mots disfluents », ce qui correspond pour les auteurs aux pauses remplies et hésitations « ben », « euh » et « hum ». Le corpus NCCFR a notamment été décrit en comparaison avec le corpus ESTER (*cf.* section 3.2.3).

3.2.5 Rhapsodie : un corpus de français parlé annoté pour la prosodie et la syntaxe

Le projet Rhapsodie [6] interroge les rapports entre syntaxe et prosodie dans la langue parlée. Dans ce contexte, un corpus annoté en syntaxe et en prosodie a

4. Campagne d'Évaluation des Systèmes de Transcription Enrichie d'Émissions Radiophoniques. Site Web : http://www.afcp-parole.org/camp_eval_systemes_transcription/, consulté le 16/01/15.
5. Site Web : http://www.mirjamernestus.nl/Ernestus/NCCFr/, consulté le 16/01/15.
6. Site Web : http://projet-rhapsodie.fr/, consulté le 16/01/15.

été développé ; ce *treebank* intonosyntaxique de 33 000 mots alignés au niveau du phonème est actuellement mis à disposition de la communauté.

Beliao et Lacheret (2013) ont mené sur le corpus RHAPSODIE des travaux portant sur les phénomènes d'oral spontané, et plus particulièrement sur le lien entretenu entre présence de marqueurs discursifs et d'hésitations vocaliques comme « hm », que les auteures assimilent aux marqueurs discursifs dans leur étude, et caractérisation de la spontanéité du discours. Elles valident notamment sur ces données la forte corrélation entre le taux de disfluences (à leur sens : hésitations, allongements syllabiques, interruptions et segments répétés) et le taux de marqueurs discursifs.

Beliao *et al.* (2014) étudient une approche intéressante visant à valider des traits linguistiques par le biais de tâches de catégorisation effectuées avec des méthodes d'apprentissage supervisé. Elle utilisent en effet la richesse d'annotation du corpus RHAPSODIE (ainsi que des traits issus des données directement, comme le nombre de mots graphiques des énoncés considérés) afin de classer les énoncés de ce corpus selon des catégories liées à la caractérisation de corpus oraux (type de parole, degré de planification du discours, interactivité, genre discursif).

3.3 Corpus de conversations menées dans des centres d'appels

Deux projets de recherche en partenariat privé-public ont permis la construction de larges corpus d'oral spontané conversationnel en français, par le biais d'enregistrements de conversations en centres d'appels. Il s'agit d'une part du projet DECODA (*cf.* section 3.3.1), qui a produit le corpus éponyme composé de dialogues entre des agents et des usagers de la RATP, et le projet Infom@gic, qui a produit deux corpus de dialogues entre des agents et des clients d'EDF (le premier, CALLSURF, avec des clients professionnels et le second, VOXFACTORY, avec des particuliers). Le corpus issu des centres d'appels d'EDF, décrit en section 3.3.2, constitue notre corpus d'étude.

3.3.1 DECODA : conversations issues de centres d'appels RATP

Le corpus DECODA (« Dépouillement automatique de conversations provenant de centres d'appels »), développé dans le cadre du projet éponyme [7], est un corpus de conversations de centres d'appels de la RATP [8]. Les objectifs du projet ainsi que les principales caractéristiques du corpus DECODA sont décrits dans Béchet *et al.*

7. Site Web : http://decoda.univ-avignon.fr/, consulté le 16/01/15.

8. Régie autonome des transports parisiens : entreprise en charge des transports publics de Paris et de sa proche banlieue.

(2012). Placé dans le cadre applicatif du suivi et de l'évaluation des centres d'appels, le projet a pour objectif de développer des méthodes et des outils d'analyse de la parole spontanée robustes et le moins supervisés possible afin d'extraire l'information pertinente de ces données en limitant le coût de l'annotation manuelle nécessaire au développement des modèles.

Le corpus DECODA est constitué des 1 514 conversations (pour environ 74 heures de signal), enregistrées sur deux journées différentes. Les conversations du centre d'appels de la RATP sont relativement courtes : 3 minutes de dialogue en moyenne ; seulement 12 % des conversations durent plus de 5 minutes. Entièrement anonymisé, le corpus a été constitué grâce à une segmentation et une transcription manuelles et a été enrichi de diverses annotations, en trois étapes successives : l'anonymisation des données personnelles sur le flux audio, une segmentation manuelle en actes de dialogue (ouverture, présentation du problème, résolution du problème, clôture) et en tours de parole [9], et une transcription manuelle réalisée avec l'outil `Transcriber` (Barras *et al.*, 1998), selon les conventions fournies par la campagne d'évaluation ESTER (Galliano *et al.*, 2009). Diverses annotations sont venues enrichir les données brutes :

- en thèmes, à l'échelle de la conversation, selon une ontologie fournie par la RATP (par exemple « information sur le trafic », « objet perdu/trouvé », *etc.*) ;
- en parties du discours, en `chunk` et en dépendances syntaxiques au moyen d'une méthode semi-supervisée. L'annotation en parties du discours et en *chunk* a été réalisée à l'aide de l'outil `Macaon` [10], et l'annotation syntaxique est détaillée dans Bazillon *et al.* (2012) ;
- en disfluences (répétitions, marqueurs discursifs et faux-départs) et en entités nommées, de manière manuelle.

Le projet DECODA a permis la création d'un corpus de conversations issues de centres d'appels riche en annotations.

3.3.2 Infom@gic : conversations issues de centres d'appels EDF

Les données conversationnelles dont nous disposons sont le résultat de deux campagnes d'enregistrements menées au sein des centres d'appels d'EDF. Elles couvrent de ce fait un large panel de sujets concernant les services proposés par l'entreprise, comme l'ouverture d'un contrat, des questions à propos des factures, des problèmes techniques rencontrés par les clients, *etc*. Ces corpus ont été constitués au sein du projet Infom@gic, dont une partie a été utilisée au sein du projet Vox Factory qui lui fait suite.

9. Tout comme les données issues des centres d'appels d'EDF sur lesquelles portent ces travaux de thèse, les données DECODA ont été enregistrées sur un seul canal, ce qui a impliqué une étape indispensable de segmentation en locuteurs.

10. Site Web : http://macaon.lif.univ-mrs.fr/, consulté le 16/01/15.

Un tableau représentatif de l'ensemble de ces données, et que nous reproduisons dans le tableau 3.1 *infra*, est présenté dans Clavel *et al.* (2013) et donne un aperçu global de l'ensemble des données et de leur composition. Par ailleurs, l'enregistrement et l'exploitation de ces conversations requiert une protection des données personnelles fournies par le client au conseiller (les noms des personnes, leur numéro de téléphone, leurs informations bancaires, des informations relatives à l'état de santé ou à des problèmes personnels). L'anonymisation des données (Grouin, 2013) s'appuie sur des méthodes de fouille de texte.

Nom corpus	**Durée**	**Nb appels**	**Trans. fine**	**Trans. rapide**	**Disfluences**
EDF (professionnels)					
Call20-fine	20	98	✓		
Call150-fast	150	1 268		✓	
Call200-fast	200	1 548		✓	
Call10-fine	10	90	✓		
EDF (particuliers)					
Vox14-fine : Vox5-neu-fine	5	33	✓		✓
Vox14-fine : Vox5-ang-fine	5	27	✓		✓
Vox14-fine : Vox4-joy-fine	4	17	✓		
Vox50-fast	50	319		✓	
Vox1000-auto	1 000	8 556			

TABLEAU 3.1 – Vue d'ensemble des corpus EDF CALLSURF et VOXFACTORY (Clavel *et al.*, 2013).

Les deux campagnes d'enregistrement menées au sein du projet Infom@gic sont les suivantes :

1. une première campagne concerne des appels de clients professionnels. Ces enregistrements ont été effectués dans un centre d'appels de Montpellier durant quatre mois et ont produit le corpus CALLSURF, décrit dans Garnier-Rizet *et al.* (2008) ;
2. une deuxième campagne concerne les particuliers. Elle a été effectuée dans un centre d'appels d'Aix-en-Provence, entre décembre 2009 et février 2010. Elle a donné lieu à un ensemble de données utilisé dans le projet Vox Factory et décrit dans Clavel *et al.* (2013).

Les données ont été collectées, transcrites et annotées par la société Vecsys [11], partenaire d'EDF. Notre étude constitue une suite logique aux travaux entrepris

11. Vecsys est une entreprise française de traitement automatique de la parole, proposant des services et logiciels de technologies vocales. Site Web : www.vecsys-technologies.fr/, consulté le 16/01/15.

dans les projets Infom@gic – CallSurf et Vox Factory. En effet, elle porte également sur l'analyse de données issues de centres d'appels, tout en constituant un travail complémentaire à celui réalisé dans ces précédents projets : nos travaux se concentrent sur l'analyse et de la détection des spécificités inhérentes à ce type de données d'un point de vue des phénomènes d'oral spontané, venant notamment porter un regard nouveau sur ces données déjà largement étudiées du point de vue des émotions (Vaudable, 2012).

CALLSURF : les clients professionnels EDF a effectué la campagne d'enregistrement suivante dans le cadre du projet Infom@gic – CallSurf (Garnier-Rizet *et al.*, 2008) : dix agents volontaires ont été enregistrés durant quatre mois lors de leurs appels avec des clients professionnels. Cette campagne a permis de constituer un large corpus de parole spontanée (5 755 appels, soit 620 heures de conversations entre clients et agents) dont les caractéristiques sont détaillées dans Danesi et Clavel (2010). Les auteures ont relevé les spécificités propres à l'oral de ce corpus (par exemple les phénomènes liés au travail de mise en mots et les effets disfluents).

VOX FACTORY : les clients particuliers Dans le cadre du projet Vox Factory, les analyses effectuées se sont basées sur les données produites à partir de la deuxième campagne d'enregistrement du projet Infom@gic. Cette campagne a permis la production d'un corpus segmenté en trois parties :

- 8 856 conversations (soit plus de 1 000 heures de signal) ont été transcrites automatiquement (corpus VOX1000-AUTO) ;
- 319 conversations (50 heures environ) ont bénéficié d'une transcription manuelle rapide (corpus VOX50-FAST) ;
- 77 conversations (14 heures environ) ont été transcrites manuellement de manière détaillée (corpus VOX14-FINE).

Nous présentons ici une description statistique du corpus VOX14-FINE. En effet, dans la mesure où nous utilisons cet ensemble de données comme corpus de référence, il convient d'en établir le profil afin de déterminer, d'une part, quel autre jeu de données parmi ceux accessibles librement sera le plus à même de soutenir une comparaison et d'autre part quels sont les paramètres, les annotations dont nous disposons pour nos travaux. Le tableau 3.1 *supra* présente une subdivision du corpus VOX14-FINE en trois sous-ensembles : VOX5-NEU-FINE, VOX5-ANG-FINE et VOX5-JOY-FINE ; elle correspond à un classement du corpus en émotions : « neutre » (*neu*), « joie » (*joy*) et « colère » (*ang*). Nous pouvons utiliser ces premières informations pour faire émerger des corrélations avec d'autres paramètres, en particulier la présence de disfluences, en utilisant les annotations de références présentes dans les deux échantillons VOX5-NEU-FINE et VOX5-ANG-FINE.

La première caractérisation de ces données concerne la durée des conversations (*cf.* figure 3.1 *supra* et tableau 3.2 *infra*). On constate que des conversations « neutre » sont plus courtes et plus homogènes que celles issues des deux autres

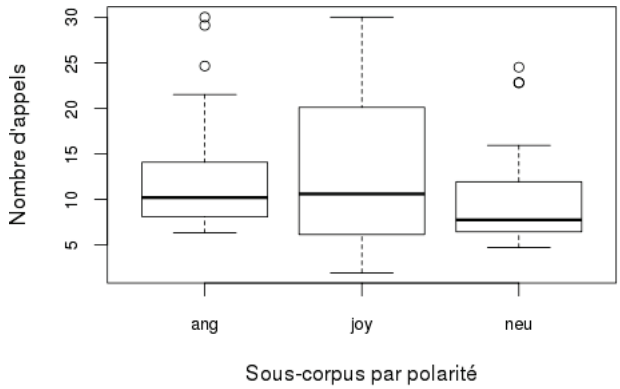


FIGURE 3.1 – Répartition des conversations par polarité dans le corpus EDF VOXFACTORY (*ang* = colère, *joy* = joie, *neu* = neutre).

échantillons ; les conversations « joie » sont les plus longues (elles ont une durée moyenne supérieure aux autres) tout en étant relativement proches de la durée moyenne des conversations « colère ». De plus, les conversations « joie » sont davantage dispersées en matière de durée (sur un intervalle d'environ 14 minutes contre environ 6 minutes pour « neutre » et « colère »).

	Neutre	**Colère**	**Joie**
Durée min.	4'70	6'31	1'56
Durée Q1	6'44	8'09	6'14
Médiane	7'75	10'20	10'60
Moyenne	9'89	12'68	13'90
Durée Q3	11'94	14'08	20'11
Écart inter-quartiles	5'50	5'59	13'57
Durée max.	24'52	30'02	30'02

TABLEAU 3.2 – Distribution des appels en fonction de leur durée (calculée en secondes) sur les sous-ensemble « Neutre », « Colère » et « Joie » du corpus EDF VOXFACTORY.

Enfin, on peut établir une représentativité des données en matière de durée : 68,8 % des conversations durent entre 5 et 15 minutes. Il est également intéressant de regarder la composition de ces données quant aux locuteurs ; le nombre de locuteurs dans une conversation, s'il dévie trop de la conversation canonique, peut par exemple être un indicateur de la singularité de la conversation (soit elle se passe globalement mal, soit le motif d'appel est complexe, *etc.*). Le corpus VOX14-FINE est composé de trois classes de locuteurs différentes :

- la classe Agents : il y a un agent dans chaque conversation du corpus, il peut y en avoir jusqu'à trois dans certaines conversations ; il y en a cependant un seul dans 93,5 % des conversations ;
- la classe Clients : on observe entre un et deux clients par conversation (un seul dans 92,2 % des conversations) ;
- la classe Répondeur : certains tours de parole correspondent en fait à un répondeur téléphonique ; 93,5 % des conversations en sont exemptes.

Ce corpus implique 134 locuteurs divisés en deux classes principales : 48,5 % des locuteurs sont des agents EDF (classe Agents) et 47,8 % sont des clients EDF (classe Clients). De plus, la conversation type en matière de locuteurs comporte deux locuteurs, un client et un agent, où l'agent prend la parole en premier (86,9 % des conversations). Le nombre maximum de locuteurs impliqués dans une conversation est six. On n'observe aucune corrélation entre la durée d'une conversation et le nombre de locuteurs qui la compose (ceci s'explique par la rareté et donc la singularité des conversations comportant plus de deux locuteurs). Le dernier des principaux paramètres descriptifs intrinsèques aux données (c'est-à-dire hors annotation injectée *a posteriori*) concerne les tours de parole, dont la distribution est représentée en figure 3.2 *infra*.

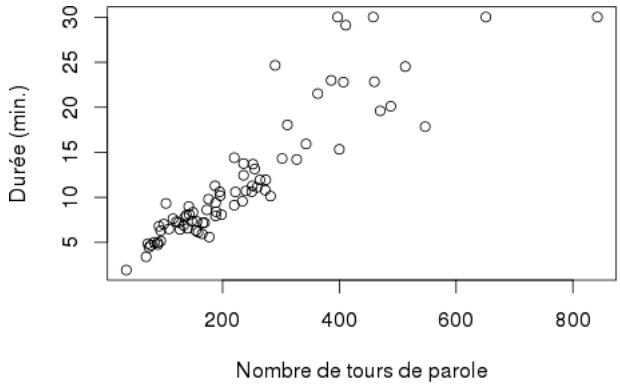


FIGURE 3.2 – Répartition des conversations par nombre de tours de parole (TP) dans le corpus EDF VOXFACTORY.

61 % des conversations sont constituées de 100 à 300 tours de parole. Cependant, comme le montrent les mesures exprimées dans le tableau 3.3 *infra*, les données sont assez dispersées de ce point de vue. On notera qu'il y a une forte corrélation entre la durée d'une conversation et le nombre de tours de parole qui la composent.

Nous avons par ailleurs observé la distribution en tours de parole des classes de locuteurs : la répartition est relativement équivalente entre le premier agent à prendre la parole et le premier client à prendre la parole (qui sont les deux classes de locuteurs les plus importantes). On peut toutefois noter que la distribution est davantage homogène pour la classe Clients que pour la classe Agents.

Mesures	Valeurs
Nombre de TP min.	35
Nb de TP dans Q1	135
Médiane	189
Nb moyen de TP	233.3
Nb de TP dans Q3	282
Écart inter-quartiles	147
Nombre de TP max.	842

TABLEAU 3.3 – Mesures sur la distribution des tours de parole (TP) par conversation dans le corpus EDF VOXFACTORY.

Ces mesures, bien qu'ayant un caractère préliminaire, nous permettent d'ores et déjà de mieux appréhender les données issues des centres d'appels d'EDF (notamment pour l'étape d'échantillonnage indispensable au traitement de corpus) mais aussi de pouvoir nous positionner par rapport aux corpus librement accessibles à la communauté scientifique, et donc fréquemment utilisés lors des campagnes d'évaluations et pour de nombreux travaux académiques.

VOXDISS : un échantillon pour étudier les disfluences d'édition Dans Clavel *et al.* (2013) nous trouvons une description en émotions de ce corpus. Nous nous appuyons sur ces descriptions afin de rendre compte des spécificités du sous-corpus que nous utilisons pour étudier plus finement les phénomènes disfluents. Ainsi, le corpus VOXDISS comprend les parties « joie » (5 heures) et « colère » (5 heures) du corpus VOXFACTORY. Clavel *et al.* (2013) détaillent également les informations sur l'annotation en émotions du corpus VOXFACTORY et sur la constitution du sous-corpus VOXDISS. Le sous-corpus VOXDISS est composé de 60 conversations entre des agents EDF et des clients particuliers. Il comprend 12 735 tours de parole (212,2 tours de parole et 2,2 locuteurs en moyenne par conversation) pour environ 10 heures de conversations enregistrées, incluant quelques superpositions de parole (36 minutes). Certaines conversations font intervenir plus de deux locuteurs lorsque, par exemple, un client fait appel à un proche pour l'aider à expliquer sa situation, ou lorsque l'intervention de plusieurs agents est nécessaire pour répondre à la demande du client.

EXEMPLE SONORE 3.1 – Les quatre classes de disfluences d'édition annotées dans VOXDISS

- Répétition : « je voulais payer par euh ben par carte peut-être »
- Auto-correction : « réponse immédiate enfin pas immédiate mais je vous »
- Faux-départ : « euh oui j'ai ø la première personne que j'ai eue »
- Disfluence combinée : « EDF dans mon circuit enfin de à à euh à Béziers »

Ce corpus a été manuellement annoté en disfluences d'édition selon une stratégie d'annotation s'appuyant sur le guide du Linguistic Data Consortium (LDC), version 6.2 (Strassel, 2004). Cette annotation a été faite de manière très précise, dans la mesure où ont été annotées à la fois les parties théoriquement distinctes des disfluences d'édition (soit l'ébauche et l'achèvement des disfluences d'édition, *cf.* la définition de ce phénomène dans le chapitre 2, section 2.2.2.4) et les différentes classes de disfluences d'édition (également décrites au sein du chapitre 2, section 2.3.2).

Concernant l'identification de la structure des disfluences d'édition, seuls le *reparandum* et le *reparans* ont été localisés. Nous avons cependant pu reconstruire automatiquement, à partir de ces annotations, la localisation de l'*interregnum*. Nous illustrons les quatre classes annotées dans l'exemple 3.1 *supra* avec pour chacune d'entre elles (répétitions, auto-corrections, faux-départs et disfluences combinées) un énoncé issu du corpus VOXDISS.

3.4 Discussion

Au sein de ce chapitre, nous avons sélectionné et présenté en section 3.2 des corpus de langue parlée en français selon quatre critères (combinés ou non) :

- leur utilisation comme corpus de référence dans la communauté scientifique pour des recherches menées sur l'oral spontané ;
- leur proximité de genre, de style de parole et/ou de domaine ;
- les annotations avec lesquelles ils ont été enrichis ;
- leur utilisation dans des travaux explorant les caractéristiques de l'oral spontané par rapport à des données de parole préparée et surtout les principaux travaux effectués sur l'analyse et la détection de phénomènes d'oral spontané (phénomènes disfluents) sur de grands corpus oraux en français.

À travers cette étude, nous nous sommes donc focalisée sur les corpus en langue française et les corpus largement étudiés dans la communauté scientifique, afin de mieux comparer et positionner nos propres travaux, effectués sur un corpus de centres d'appels non disponible à la communauté scientifique (le corpus EDF VOXDISS). Bien entendu, de nombreux travaux sur l'identification de phénomènes disfluents ont par ailleurs été menés dans d'autres configurations : soit

sur des données en langue autre que le français, et notamment sur l'anglais à travers l'étude du SWITCHBOARD corpus, soit sur des corpus de langue française appartenant à des domaines fermés. Par exemple, Bouraoui (2008) a également travaillé sur les disfluences dans un corpus conversationnel en domaine restreint (le contrôle aérien). Ces travaux sont discutés au sein du chapitre 2.

Nous avons attaché une attention particulière à la description des deux corpus de langue française constitués de conversations en centres d'appels : le corpus issu des centres d'appels d'EDF (corpus VOXDISS) et le corpus issu des centres d'appels de la RATP (corpus DECODA). Ces corpus ont tout deux été annotés en disfluences, et nous souhaitons à terme évaluer nos travaux en analyse et détection des disfluences d'édition sur ces deux ensembles. Pour les travaux décrits dans ce document, nous avons utilisé le corpus VOXDISS décrit *supra*, qui grâce aux annotations en disfluences qui y ont été réalisées dans le cadre du projet Vox Factory nous a permis de produire une analyse du comportement de ces événements dans un corpus de dialogues homme/homme en contexte fermé. Le tableau 3.4 *infra* résume les caractéristiques quantitatives des corpus décrits dans cette section.

Corpus	**Modalité**	**Durée**	**NbE**	**Dur. moy.**	**NbL**	**NbM**	**Lexique**
EDF	conv. form.	1 620 h	14 311	7 min.	–	–	–
DECODA	conv. form.	74 h	1 514	3 min.	–	483 k	8 806
CRFP	divers	36 h	134	16 min.	10	440 k	–
RITEL	requêtes	6 h 40	652	–	13	71 k	3 434
NCCFr	conv. inform.	36 h	23	1 h	46	469 k	15 574
ESTER	divers	90 h	–	–	–	–	–

TABLEAU 3.4 – Description comparative des corpus d'oral en langue française, où NbE = nombre d'enregistrements, NbL = nombre de locuteurs et NbM = nombre de mots du corpus.

Deuxième partie

Actualisation des disfluences en contexte conversationnel : des indices statistiques de localisation à l'identification automatique

Chapitre 4

Caractérisation des disfluences dans la parole conversationnelle

Sommaire

APRÈS avoir présenté, dans la première partie de cette thèse, les aspects théoriques et définitoires des *disfluences* (*cf.* chapitre 2), nous nous attachons dans cette deuxième partie à leur analyse en corpus et à leur identification automatique dans des données d'oral conversationnel. Plus précisément, en lien avec les corpus d'oral spontané également décrits dans la partie I, au sein du chapitre 3, nous nous attachons dans ce chapitre 4 à étudier les caractéristiques que présentent les *disfluences* dans des conversations issues de centres d'appels.

4.1 Introduction : caractériser la parole spontanée et les phénomènes disfluents dans les dialogues homme-homme

La dernière décennie a vu se développer un intérêt grandissant pour l'analyse de la parole sur différents types de données orales, et plus particulièrement sur des données de centres d'appels, dans des perspectives d'application marketing notamment. Cette attention portée aux données spontanées d'interaction téléphonique élève au premier plan des défis scientifiques connus ; un problème clef est le développement de systèmes d'extraction d'information capables de traiter des indices d'oral spontané à différents niveaux : acoustico-phonétique, lexical, syntaxique, sémantique, dialogique, *etc.* Pendant de nombreuses années, les phénomènes caractéristiques de l'oral spontané ont été envisagés comme des éléments dégradant la qualité du discours et nuisant à sa compréhension.

En particulier, les cassures dans l'énoncé, avec des éléments venant « rompre » la progression syntagmatique du message linguistique, sont toujours considérés comme le signal d'un message en cours d'élaboration – comme le serait un brouillon par rapport au texte écrit finalisé. L'étude de la langue orale spontanée a ainsi longtemps été négligée dans les études linguistiques.

Cependant, ces dernières années, un intérêt plus important à été porté à l'étude des phénomènes d'oral spontané (Shriberg, 1994), tant en analyse qu'en identification automatique, et sur des données hétérogènes en différentes langues. Dans le but de détecter les disfluences et de les catégoriser afin de mieux les prendre en compte pour l'amélioration de tâches d'extraction d'information, nous présentons ici une caractérisation des disfluences dans un corpus d'oral conversationnel. Notre hypothèse de travail est qu'une analyse fine des phénomènes disfluents doit permettre de mieux les identifier et donc de mieux les détecter : les disfluences présentes dans le corpus de conversations en centres d'appels étudié dans cette thèse présentent-elles des caractéristiques particulières par rapport à celles identifiées dans la littérature ? Si oui, cela peut-il avoir un impact sur leur détec-

tion ? Présentent-elles des traits discriminants aux niveaux lexical et acoustico-prosodique qui feraient de bons candidats d'indices pour leur détection ?

Pour répondre à ces questions et mener à bien ces analyses, nous avons conçu et mis en œuvre une chaîne de traitement des données orales conversationnelles, ayant pour but d'enrichir ce type de données orales (à partir du signal de parole et des transcriptions associées) de nombreuses informations (acoustiques, lexicales, discursives, annotations, méta-données). Les principes et détails de cette chaîne de traitement sont décrits en section 4.2. Nous présentons ensuite les analyses menées en caractérisation de la parole spontanée conversationnelle et en caractérisant des disfluences d'édition (*cf.* section 4.3), avant de discuter ces approches et apports, qui seront mis en perspective avec notre objectif de détection automatique des disfluences dans la parole conversationnelle (*cf.* section 4.4).

4.2 Méthodologie : traitements pour l'analyse linguistique et acoustique des disfluences

4.2.1 Introduction : objectifs et principes de génération de données orales enrichies pour l'analyse des disfluences

Nous détaillons dans cette section la méthodologie mise en place pour le traitement et l'enrichissement d'un corpus d'oral spontané conversationnel (ici le corpus EDF VOXDISS), afin d'étudier les propriétés acoustiques, prosodiques, lexicales et discursives de ce type de données. En effet, l'objectif poursuivi est de mieux appréhender l'étude et l'analyse des données orales grâce à une meilleure connaissance d'un phénomène très caractéristique de la parole spontanée, les disfluences. Par le biais d'une analyse de ces dernières en contexte, nous souhaitons mieux comprendre leurs spécificités de réalisation en corpus afin de pouvoir les modéliser et les détecter dans des données orales. Pour cela, nous avons besoin d'enrichir automatiquement les corpus oraux d'un certain nombre de traits relevant de plusieurs niveaux d'analyse de la langue (prosodique, lexique, discours, *etc.*) afin d'en mener une analyse statistique détaillée.

Cette étude permet donc de caractériser les phénomènes disfluents dans un corpus d'oral spontané conversationnel pour répondre notamment aux questions suivantes : les traits acoustiques et lexicaux des phonèmes et des mots composants le corpus se réalisent-ils différemment dans un mot apparaissant à l'intérieur d'une disfluence ? Si oui, existe-t-il des variations entre les différentes classes de disfluences ? Ce mot est-il précédé et/ou suivi d'une pause silencieuse ou remplie ? d'un mot tronqué (amorce...) ? *etc.*

Nous avons pour cela mis au point une chaîne de traitement, pour laquelle la figure 4.1 *infra* propose une vue d'ensemble. Cette chaîne, qui prend en entrée

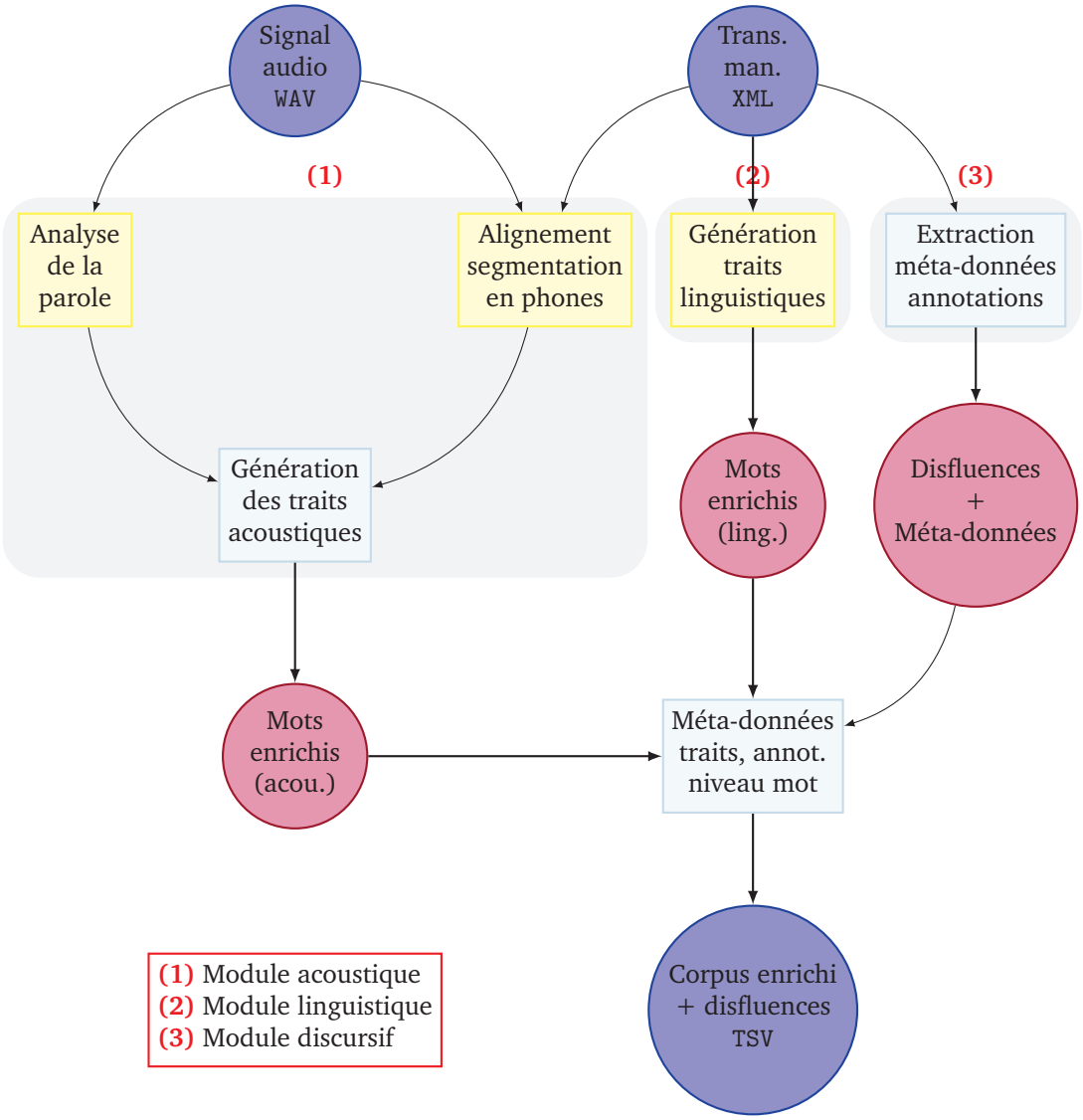


FIGURE 4.1 – Vue d'ensemble de la chaîne d'enrichissement des données orales conversationnelles.

des données issues de l'oral (d'une part le signal de parole et d'autre part les transcriptions manuelles de ce signal), est composée de trois modules principaux :

- un module « acoustique », détaillé en section 4.2.2 ;
- un module « linguistique », détaillé en section 4.2.3 ;
- un module « discursif », détaillé en section 4.2.4.

La section 4.2.5 est dédiée à des remarques générales sur la généricité de notre approche et de nos choix techniques, et nous discutons l'ensemble de ces travaux en section 4.2.6.

4.2.2 Module d'extraction de traits acoustiques

Le module d'analyse acoustique de la chaîne de traitement, schématisé en figure 4.2 *infra*, est composé de deux traitements principaux, à partir desquels sont enrichies les données orales. Il prend en entrée le signal de parole au format `audio/wav` et les transcriptions manuelles qui lui sont associées au format `text/xml`. Le module produit en sortie les mots du corpus enrichis avec des traits acoustiques, au format `text/csv`.

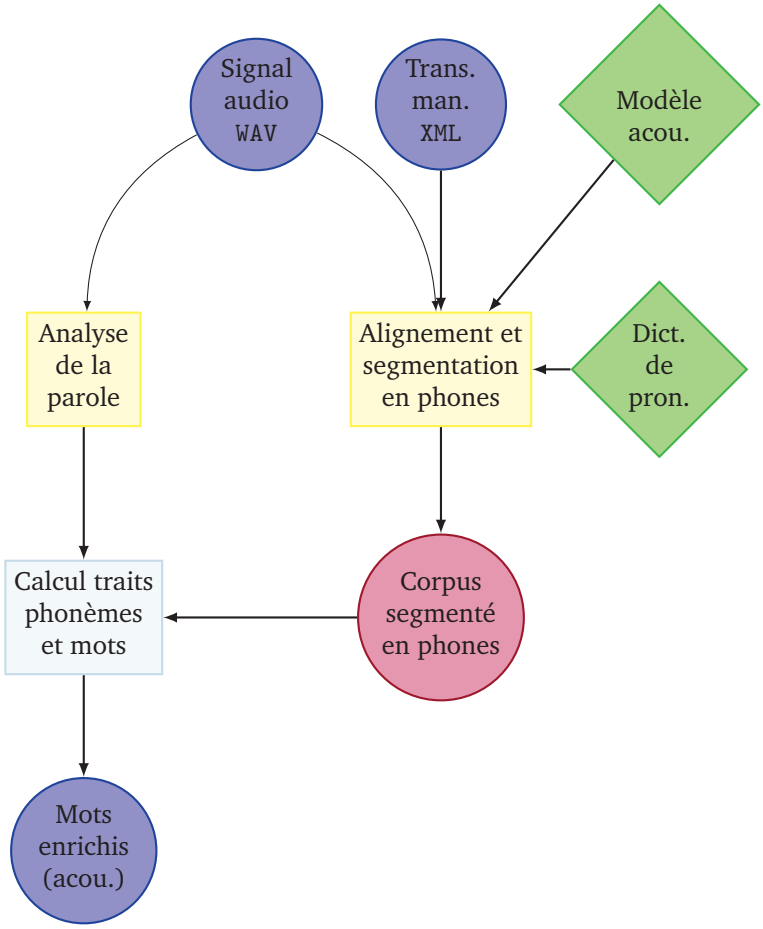


FIGURE 4.2 – Module acoustique de la chaîne d'enrichissement de données orales conversationnelles.

D'une part, le corpus audio a été automatiquement aligné par le système d'alignement automatique de la parole du Laboratoire d'Informatique et de Mécanique pour les Sciences de l'Ingénieur (LIMSI), basé sur le système de reconnaissance de la parole décrit dans Gauvain *et al.* (2005). Deux ressources externes sont utilisées pour cette étape : le dictionnaire de prononciation du système d'alignement auto-

matique du LIMSI dans une version légèrement enrichie et le modèle acoustique développé pour le système de dialogue en domaine ouvert RITEL (Van Schooten *et al.*, 2007). Cette procédure d'alignement forcé permet l'obtention d'une segmentation en mots et en phones, avec indication des durées et pauses silencieuses. Le système de reconnaissance de la parole utilise l'alphabet phonétique du LIMSI, dont l'inventaire des symboles comparé à l'Alphabet Phonétique International (API) est reproduit dans le tableau 4.1 *infra*.

API	i	e	ɛ	y	ø	ə	a	ɔ	o	u	ɛ̃	ɑ̃	ɔ̃	ɥ	w	j
LIMSI	i	e	E	y	@	x	a	c	o	u	I	A	O	h	w	j

(a) Voyelles, voyelles nasales et semi-voyelles.

API	p	b	t	d	k	g	f	v	s	z	ʃ	ʒ	m	n	ɲ	l	ʁ
LIMSI	p	b	t	d	k	g	f	v	s	z	S	Z	m	n	N	l	r

(b) Consonnes.

	Silence	Respiration	Hésitation
LIMSI	.	H	&

(c) Autres.

TABLEAU 4.1 – Correspondance des phonèmes du système d'alignement automatique du LIMSI avec les symboles de l'Alphabet Phonétique International.

Quatre segments n'ont pas pu être alignés par le système (équivalent à sept *tokens*), en raison de la qualité acoustique des enregistrements : l'alignement forcé entre la fenêtre temporelle et le nombre de phonèmes à associer n'a pas pu être effectué. Par ailleurs, une vérification manuelle a également permis de repérer quelques défauts d'alignement. Pour cette phase de vérification manuelle, nous nous sommes concentrée sur les phonèmes dont la durée paraissait exceptionnellement longue ; l'erreur suivante, illustrée dans l'exemple 4.1 *infra*, exemplifie le type d'erreurs rencontré.

EXEMPLE SONORE 4.1 – Illustration des défauts d'alignement automatique : erreurs de durées.

« au revoir » [orəvwar]

Dans cet exemple, le système génère une pause silencieuse de 2.7 secondes entre [o] et [rəvwar]. De plus, la réalisation du phonème /w/, qui apparaît en syllabe finale du mot « au revoir » [orəvwar], présente en sortie d'alignement une

durée de 14.9 secondes. Les durées assignées à chacun des phonèmes sont présentées dans le tableau 4.2 *infra*. L'erreur du système d'alignement est double, comme nous pouvons le constater à l'écoute de cet exemple (*cf.* exemple sonore 4.1 *supra*) : elle porte à la fois sur l'insertion erronée d'une pause silencieuse et sur la durée du phonème /w/. Cette vérification manuelle pointe les difficultés posées par une tâche d'alignement automatique forcé sur des données orales.

Phonème	o	[PAUSE]	r	ə	v	w	a	r
Durée (s.)	0.340	2.700	0.060	0.700	0.270	14.900	0.030	0.030

TABLEAU 4.2 – Exemple d'erreurs de durées assignées par le système d'alignement automatique.

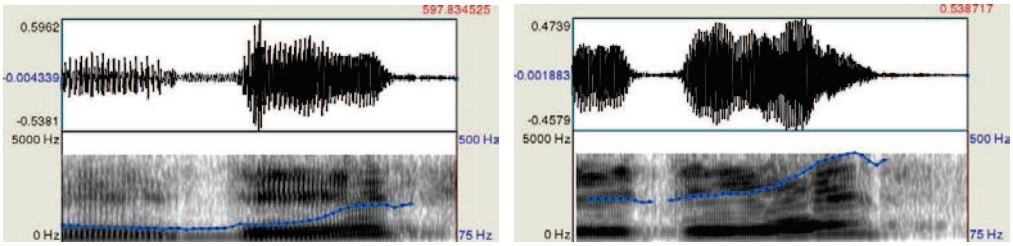
Nous avons par ailleurs extrait à l'aide du logiciel `Praat` (Boersma et Weenink, 2001) les contours de fréquence fondamentale (F_0) et des trois premiers formants (F_1, F_2 et F_3) à partir du signal de parole, en utilisant les paramètres standards proposés. Les mesures ont été faites toutes les 5 millisecondes, selon la méthode utilisée par Adda-Decker *et al.* (2008). La fréquence fondamentale, calculée en Hertz (Hz), encode la hauteur de la voix (notions de voix aiguë ou grave) et relève de l'intonation prosodique du discours ; elle se situe en général en dessous de 300 Hz. L'exemple 4.2 *infra* présente deux énonciations du mot « EDF » par la même locutrice, prenant une voix grave et une voix plus aiguë. Le tracé de F_0 correspondant est illustré en figure 4.3 *infra*.

EXEMPLE SONORE 4.2 – Exemple de différence de F_0 basse et élevée sur une même locutrice.

- « EDF » prononcé avec une voix grave (F_0 basse) : [edeɛf]
- « EDF » prononcé avec une voix aiguë (F_0 élevée) : [ədeɛf]

Les trois premiers formants relèvent quant à eux de la caractérisation vocalique : F_1 encode les degrés d'aperture, F_2 la position antérieure ou postérieure et F_3 la position des lèvres (arrondie ou étirée). L'ensemble de ces traits acoustiques, de même que les durées, ont été extraits suivant un pas de 5 millisecondes (ms), puis calculés pour chaque phonème du corpus grâce à la segmentation issue de l'alignement automatique.

Ces extractions acoustiques permettent de calculer les traits acoustiques et prosodiques associés à chaque mot du corpus de parole. Ces traits sont dérivés des valeurs de durée des phonèmes, de leur fréquence fondamentale, F_1 et F_2, et des présences et durées de pauses silencieuses autour des mots.



(a) [edɛɛf] voix grave (F_0 basse) (b) [ədɛɛf] voix aiguë (F_0 élevée)

FIGURE 4.3 – Variations de F_0 sur la prononciation d'un même mot par une même locutrice.

Chaîne phonémique Nous avons associé à chaque mot du corpus sa chaîne phonémique, le nombre de phonèmes le composant et sa longueur syllabique. Cette dernière a été calculée en fonction du nombre de voyelles le composant, en excluant le schwa final le cas échéant : par exemple, le mot « commune », prononcé [komynə], contient 6 phonèmes et a une longueur syllabique de 2.

Durées de phonèmes À partir de la durée de chaque phonème, nous avons calculé pour chaque mot la durée moyenne des phonèmes qui le composent, ainsi que la durée moyenne des voyelles et des consonnes qui le composent. La durée de chaque mot a également été calculée.

Durées des pauses silencieuses La durée des pauses silencieuses, fournies par le système d'alignement automatique, permet d'associer pour chaque mot, le cas échéant, la durée des pauses silencieuses qui le précèdent et le suivent.

Moyennes de F_i Les moyennes de la fréquence fondamentale (F_0) et des deux premiers formants (F_1 et F_2) de chaque mot ont été calculées à partir des valeurs centrales associées à chaque voyelle composant le mot. En effet, nous avons extrait avec `Praat` les valeurs de F_i pour chaque segment vocalique en mesurant cette valeur en trois points (segment de début, segment central, segment de fin). Les mesures de segments centraux présentent davantage de stabilité acoustique.

Taux de voisement des voyelles Nous avons également déterminé avec un script `Praat` le taux de voisement de chaque segment phonémique : ce taux de voisement est calculé par le ratio entre le nombre de mesures de F_0 non nulles et le nombre total de mesures. Le taux de voisement atteint 100 % si la voyelle est totalement voisée, et 0 % s'il n'y a aucun voisement dans le segment.

Nous considérons, sur l'exemple de Nemoto (2011), qu'avec un taux de voisement supérieur à 70 %, nous écartons tout problème potentiel résultant de la

prononciation (par exemple phénomène de voix craquée), de l'alignement et de la segmentation automatique, ou de l'extraction de la fréquence fondamentale. À partir de ces informations, nous avons généré pour chaque mot les informations suivantes, qui permettent d'évaluer la qualité théorique des mesures acoustiques de F_0 associées à chaque mot (avec plus ou moins de sévérité selon le mode de calcul de ces traits et filtres) :

- le nombre de voyelles appartenant au mot et présentant un taux de voisement strictement inférieur à 70 % ;
- un filtre relatif à cette information : si le mot présente au moins une voyelle avec un taux de voisement inférieur à 70 %, alors le filtre est négatif ; sinon le filtre est positif ;
- le taux de voisement moyen des voyelles composant le mot ;
- un filtre relatif à cette information : si le taux de voisement moyen des voyelles est supérieur à 70 %, alors le filtre est positif ; sinon, le filtre est négatif.

Deltas et contours de F_0 Des différences entre mesures de F_0 sur les voyelles composant chaque mot ont également été calculées : d'une part le $\Delta_{max-min}$, qui calcule la différence entre la F_0 la plus élevée sur les voyelles et la F_0 la plus basse, d'autre part le $\Delta_{fin-deb}$, qui calcule la différence entre la F_0 de la première voyelle du mot et la F_0 de la dernière voyelle ; cette deuxième mesure prend donc en compte la position de la voyelle au sein du mot. Nous avons ensuite évalué le contour de F_0 global du mot, à partir du $\Delta_{fin-deb}$: si ce dernier est strictement supérieur à zéro, alors le contour est montant ; si ce dernier est égal à zéro alors le contour est plat ; si ce dernier est strictement inférieur à zéro alors le contour est descendant. Les deux formules *infra* formalisent les méthodes de calcul des deux delta.

$$\Delta_{max-min} = \frac{F_0 \text{ la plus élevée sur les voyelles du mot}}{F_0 \text{ la plus basse sur les voyelles du mot}}$$

$$\Delta_{fin-deb} = \frac{F_0 \text{ de la première voyelle du mot}}{F_0 \text{ de la dernière voyelle du mot}}$$

4.2.3 Module d'extraction de traits linguistiques

Le module d'analyse linguistique de la chaîne de traitement, schématisé en figure 4.4 *infra*, consiste en une extraction du texte à partir des transcriptions manuelles du corpus VOXDISS, de manière à enrichir les mots *via* un étiquetage morpho-syntaxique.

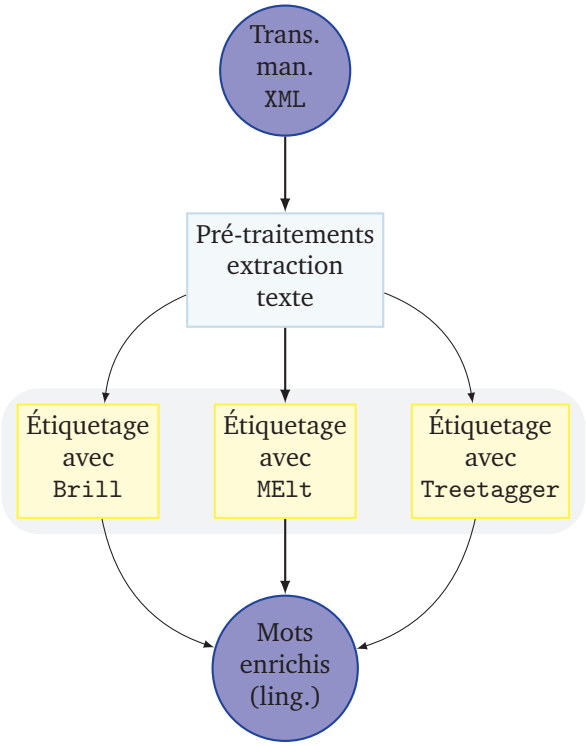


FIGURE 4.4 – Module linguistique de la chaîne d'enrichissement de données orales conversationnelles.

Nous avons combiné plusieurs outils d'étiquetage : une version adaptée au français de l'étiqueteur `Brill` (Allauzen et Bonneau-Maynard, 2008), `MElt` (Denis et Sagot, 2009) et le `Treetagger` (Schmid, 1994). Ces trois outils d'étiquetage sont assez disparates en matière de jeu d'étiquettes. En effet, sur l'ensemble du corpus de travail VOXDISS, l'étiquetage avec `Brill` produit 239 étiquettes uniques, alors que `MElt` et le `Treetagger` en produisent respectivement 28 et 30 (ces deux derniers présentent un jeu d'étiquettes très proche). Le jeu d'étiquette de `MElt` et celui du `Treetagger` sont référencés en annexe A. Concernant `Brill`, dont le jeu d'étiquettes est beaucoup plus conséquent, il est décrit dans Paroubek (2000). Nous avons conservé pour l'ensemble des tâches d'étiquetage la segmentation en mots d'origine des données, de manière à rendre possibles *(i)* l'union des sorties des trois outils d'étiquetage pour enrichir nos données à l'échelle des mots et *(ii)* l'évaluation des trois outils sur un extrait de nos données annoté manuellement. Par ailleurs, le `Treetagger` mène également une tâche de lemmatisation, ce qui n'est pas le cas des deux autres étiqueteurs. Afin de pallier cette disparité et d'exploiter au mieux les sorties de ces outils, nous avons opéré une étape d'éclatement des étiquettes de `Brill` et de `MElt` produisant huit traits supplémentaires aux étiquettes d'origine (soit neuf traits en tout pour chacun des deux étiqueteurs).

Nous avons mené une évaluation succincte de ces trois étiqueteurs morpho-syntaxiques, en comparant les étiquetages produits à un étiquetage de référence réalisé manuellement sur un extrait du corpus VOXDISS, composé de 11 tours de parole consécutifs pour 90 mots. Le tableau 4.3 *infra* résume les performances obtenues par `Brill`, `MElt` et le `Treetagger` sur cet extrait, mesurées en précision. L'extrait manuellement étiqueté en parties du discours comporte de nombreuses difficultés pour la réalisation automatique de cette tâche : il comporte notamment des amorces de mots, des segments non transcrits du fait de superposition de parole, et deux disfluences d'édition. La confusion la plus fréquente, tout étiqueteurs confondus, est opérée entre prépositions et adverbes. Dans la mesure où le `Treetagger` donne les meilleurs résultats sur cet extrait, nous nous sommes basée sur les parties du discours qu'il a générées pour réaliser les analyses de corpus et caractérisation des disfluences d'un point de vue lexical.

Étiqueteur	`Brill`	`MElt`	`Treetagger`
Score	0.8089	0.8523	0.8764

TABLEAU 4.3 – Évaluation des trois étiqueteurs morpho-syntaxiques `Brill`, `MElt` et `Treetagger` sur un extrait du corpus VOXDISS manuellement annoté en parties du discours.

4.2.4 Module d'extraction de traits discursifs : méta-données et annotations

Le module d'extraction des traits discursifs, schématisé en figure 4.5 *infra*, est un module dédié à l'extraction d'informations déjà présentes dans les données ; aucune information supplémentaire n'a été générée. Il s'agit ici, à partir des transcriptions manuelles manuellement annotées en disfluences (*cf.* section 3.3.2 pour la description des annotations dans le corpus VOXDISS) de formater les données pour en extraire *(i)* des « méta-données » relatives au discours et au dialogue et *(ii)* les annotations en disfluences d'édition.

Concernant les traits discursifs, nous avons extrait le genre (homme/femme) et la classe (principalement Agents *versus* Clients) des locuteurs. Ces informations ont été reportées sur chaque mot. À partir du nombre de mots de chaque tour de parole et du nombre de tours de parole de chaque conversation, deux traits relatifs à la progression discursive ont été générés. Ces deux indices sont similaires et représentent, pour chaque mot, la place que ce dernier occupe au sein du tour de parole et la place qu'occupe le tour de parole auquel il appartient à l'échelle de la conversation ; ils ont été calculés comme présenté dans les formules *infra*.

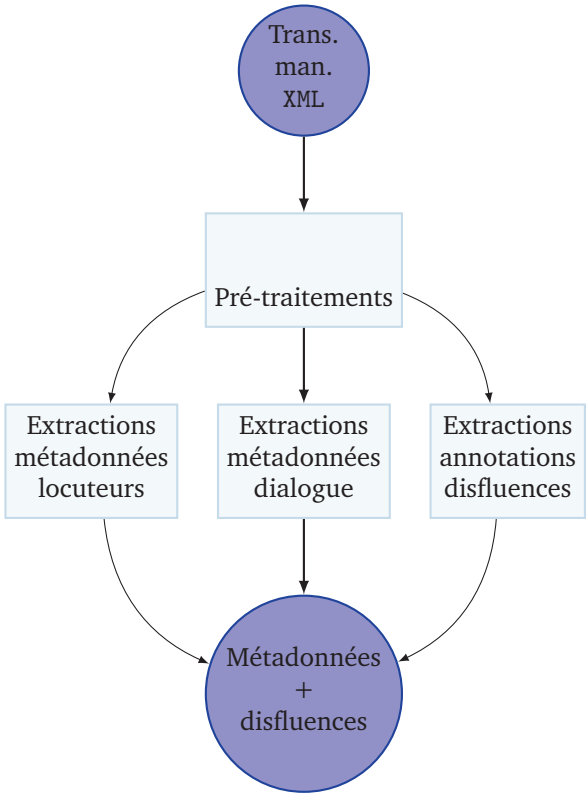


FIGURE 4.5 – Module discursif de la chaîne d'enrichissement de données orales conversationnelles.

$$\text{Prog}_{TP} = \frac{\text{rang du mot dans le TP}}{\text{nombre de mots dans le TP}} \times 100$$

$$\text{Prog}_{CONV} = \frac{\text{rang du TP dans la conversation}}{\text{nombre de TP dans la conversation}} \times 100$$

Concernant les annotations en disfluences, nous avons reporté ces dernières sur chaque mot du corpus pour indiquer son appartenance à un segment de structuration des disfluences (*reparandum*, *interregnum*, *reparans*, *cf.* section 2.3.1) et à une classe de disfluence (répétition, auto-correction, faux-départ ou disfluence combinée, *cf.* section 2.3.2). Ces annotations sont cruciales pour l'étude des contextes d'apparition des disfluences dans ce corpus d'oral conversationnel.

4.2.5 Remarques sur la généricité de la chaîne de traitement

Cette chaîne de traitement, dédiée à l'extraction et à la génération de traits acoustico-prosodiques, linguistiques et discursifs pour l'analyse des données orales, a été construite pour l'analyse du corpus VOXDISS. Elle est toutefois en grande partie générique, dans la mesure où les seules modifications à apporter pour l'adapter à un autre corpus résident dans la modification des scripts suivants :

- au sein du module « linguistique » : prise en compte d'un schéma `XML` de structuration des transcriptions manuelles différent de celui actuellement traité (soit le vocabulaire `XML` standard de l'outil `Transcriber`) ;
- au sein du module « discursif » : également la prise en compte d'un autre schéma `XML` que celui de `Transcriber` ; également une adaptation aux méta-données et éventuelles annotations présentes dans les transcriptions manuelles.

De même, nous nous attachons ici à l'étude des phénomènes disfluents (*via* l'extraction d'annotations manuelles en disfluences d'édition) mais la chaîne de traitement est parfaitement extensible et/ou adaptable à tout autre phénomène de langue que l'on souhaiterait analyser, que celui-ci soit présent dans les méta-données des documents ou sous la forme d'annotations manuelles et quelle que soit son unité d'attache (phonème, mot, tour de parole, *etc.*).

4.2.6 Discussion

Nous avons présenté dans cette section la méthodologie que nous avons mise en place afin de traiter et d'enrichir un corpus d'oral conversationnel. La chaîne de traitement développée permet d'extraire les informations présentes dans les données (lexique, méta-données, annotations manuelles, *etc.*) et d'en générer d'autres, selon plusieurs dimensions : enrichissement acoustique à partir du signal de parole et d'un alignement automatique entre celui-ci et les transcriptions manuelles, enrichissement linguistique et enrichissement discursif à partir des transcriptions manuelles.

Cette chaîne de traitement est dans le carde de ces travaux orientée vers l'analyse de phénomènes d'oral spontané *via* l'extraction d'annotations manuelles en disfluences d'édition, mais peut également être adaptée à l'analyse d'autres phénomènes qu'il serait intéressant d'analyser en parole conversationnelle. Les choix opérés pour sélectionner le type d'enrichissement des données sont motivés par une analyse de la littérature sur les caractéristiques acoustico-prosodiques et linguistiques des disfluences et autres phénomènes d'oral spontanés (par exemple, les marqueurs discursifs ou les hésitations vocaliques) et sont notamment destinés à permettre de confirmer ou d'invalider nos hypothèses sur la caractérisation des disfluences dans un corpus d'oral conversationnel issu de centres d'appel.

4.3 Analyses lexicales et acoustico-prosodiques des disfluences

Nous nous attachons ici à décrire les disfluences présentes dans le corpus VOXDISS, de manière globale et par classe de disfluences lorsque cela s'y prête. L'enjeu principal et l'apport de ces travaux sont de fournir une analyse détaillée des données conversationnelles et de faire émerger des traits caractéristiques des disfluences en corpus, en prenant en compte des paramètres acoustiques et lexicaux. Les disfluences sont également considérées en fonction de certains paramètres conversationnels, comme les tours de parole ou le profil du locuteur (Agent *vs* Client). Nous souhaitons ainsi mieux identifier ce type de données à travers le prisme de l'analyse de l'oral conversationnel, tout en mettant en évidence l'analyse les disfluences de manière dissociée selon un certain nombre de paramètres (différenciation selon les locuteurs par exemple). Nous nous efforçons par ailleurs de comparer le corpus analysé avec d'autres corpus de parole spontanée en français, notamment en référence aux corpus de parole présentés en section 3.2 et aux études qui s'y rapportent.

Après une caractérisation générale des disfluences dans le corpus EDF VOXDISS (*cf.* section 4.3.1), nous présentons les résultats d'analyses avec une mise en perspective acoustique (*cf.* section 4.3.2) et linguistique (*cf.* section 4.3.3). La section 4.3.4 est consacrée à une étude transverse mettant en avant les profils de locuteurs dans le corpus VOXDISS.

4.3.1 Actualisation des disfluences dans la parole spontanée conversationnelle

Il paraît difficile d'évaluer le taux global de disfluences présentes dans le corpus VOXDISS par rapport à d'autres corpus comparables, par manque de référentiel commun. Par exemple, Adda-Decker *et al.* (2004) précisent que 8 % des mots du corpus qu'ils étudient (10 heures d'interview entre journalistes et personnalités) correspondent à des disfluences, et plus particulièrement que 3,2 % des mots sont des répétitions et 2,3 % des faux-départs. Nous pouvons avancer que les répétitions et faux-départs couvrent davantage de mots dans leurs données (5,5 %) que dans le corpus VOXDISS (4,4 %), mais cela revient à une caractérisation partielle des données qui soulève d'autres questions : de manière globale, les disfluences rencontrées dans le corpus VOXDISS sont-elles particulièrement couvrantes (quant au lexique) par rapport à des corpus similaires ? Peut-être sont-elles moins couvrantes mais plus nombreuses ? Si oui, pourquoi ?

Le tableau 4.4 *supra* montre la distribution dans le corpus VOXDISS des quatre classes de disfluences d'édition annotées (soit : répétitions, auto-corrections, faux-départs, disfluences combinées). Nous notons d'ores et déjà des similarités avec

	Nb de disfluences	Moyenne / appel	Moyenne /TP
Répétitions	1 166	19.4	0.10
Auto-corrections	1 056	17.6	0.09
Faux-départs	569	9.5	0.05
Disfluences combinées	742	12.4	0.06
Disfluences d'édition	3 533	58.9	0.32

TABLEAU 4.4 – Caractéristiques principales des quatre classes de disfluences d'édition dans le corpus VOXDISS (où TP = tour de parole).

d'autres données en matière de distribution de certaines disfluences. Ainsi les répétitions sont particulièrement nombreuses dans le corpus VOXDISS (*cf.* notamment Adda-Decker *et al.* (2003), Béchet *et al.* (2012), Piu et Bove (2007)). Par exemple, un échantillon de 54 minutes issu du Corpus de Référence du Français Parlé (CRFP), décrit en section 3.2.1 et qui a également été annoté selon le guide du Linguistic Data Consortium, présente la distribution de disfluences suivante : 50 % de répétitions, 19 % de disfluences combinées, 16 % de faux-départs et 15 % d'auto-corrections (Piu et Bove, 2007) (*cf.* tableau 4.5 *infra*) [1].

	VOXDISS	CRFP
Répétitions	33 %	50 %
Auto-corrections	30 %	15 %
Faux-départs	16 %	16 %
Disfluences combinées	21 %	19 %

TABLEAU 4.5 – Distribution des disfluences d'édition dans le corpus EDF VOXDISS et le CRFP.

À travers un autre grand projet d'étude des conversations issues de centres d'appels en français, le projet DECODA (*cf.* section 3.3.1), Béchet *et al.* (2012) présentent notamment une description quantitative d'événements disfluents. Dans cette étude, les disfluences considérées sont les répétitions, les faux-départs et les marqueurs discursifs. Il est notable que les deux corpus présentent des taux de répétitions assez proches : les répétitions sont présentes dans 8 % des tours de parole du corpus DECODA et dans 8,8 % des tours de parole du corpus VOXDISS. Il

1. La répartition des classes de disfluences présentée ici pour le CRFP diffère quelque peu des chiffres annoncés dans (Piu et Bove, 2007). En effet, ces derniers incluent dans les disfluences (et donc dans leurs calculs) les amorces. Nous avons recalculé la répartition des quatre classes que nous intégrons aux disfluences d'édition sans prendre en compte les amorces dans la valeur de référence (total de phénomènes disfluents pris en compte).

y a cependant davantage de faux-départ dans VOXDISS (5 % des tours de parole) que dans le corpus DECODA (1,1 % des tours de parole).

Plus spécifiquement dans les données VOXDISS, nous avons également analysé la corrélation entre les disfluences et la longueur des conversations : les deux variables sont hautement corrélées, avec un coefficient de 0,84[2]. Ce résultat est principalement dû à la corrélation entre la longueur des conversations et la présence d'auto-corrections (c=0,87). Les répétitions sont elles moins corrélées à la longueur des conversations bien que ces événements disfluents sont logiquement plus présents dans les conversations les plus longues. Les répétitions constituent le type de disfluences le plus fréquent et ont une distribution relativement homogène par rapport aux autres disfluences entre conversations longues et courtes.

En complément, le tableau 4.6 *infra* met en avant la distribution des tours de parole contenant au moins une disfluence *versus* ceux exempts de disfluences : même si les tours de parole disfluents ne représentent que 20,9 % de la totalité des tours de parole, ils couvrent 49,2 % des mots prononcés. Ils ont également un taux élevé de nombre moyen de mots par tour de parole et un débit moyen élevé (4,4 mots par seconde contre 3,5 mots par seconde dans les tour de parole sans disfluence).

	TP disfluents	**TP non-disfluents**
Répartition	20,9 %	79,1 %
Durée	04h16	05h02
Durée moyenne (sec.)	06'80	02'12
Couverture lexicale	49,2 %	50,8 %
Nombre moyen de mots	29,1	7,9
Débit moyen (mots/sec.)	4,4	3,5

TABLEAU 4.6 – Caractéristiques principales des tours de parole (TP) contenant au moins une disfluence *versus* les tours de parole exempts de disfluences dans le corpus VOXDISS.

Nous sous sommes également intéressée à la position des disfluences dans le discours et dans les conversations. En effet, nous faisons l'hypothèse qu'elles n'apparaissent pas nécessairement au même endroit dans le flux de parole. Pour étudier cette composante, nous avons étudié la répartition des disfluences d'édition selon deux paramètres, chacun informatif sur deux niveaux discursifs : en premier lieu, la progression globale de la conversation (*cf.* section 4.8) et en second lieu la progression locale des tours de parole (*cf.* section 4.7). Le mode de calcul de ces deux informations est décrit en section 4.2.4.

2. Méthode utilisée : coefficient de corrélation de PEARSON. Corrélation calculée avec le logiciel de traitement statistique R (Site Web : http://www.r-project.org/), utilisant le langage de programmation S.

Il est intéressant de noter que la distribution des disfluences au sein des tours de parole est très homogène et ne semble pas vraiment déterminante (*cf.* tableau 4.7 *infra*). En revanche, si l'on s'attache à l'analyse des résultats présentés au sein du tableau 4.8 *infra*, on peut observer que le lien entre présence de disfluences et progression du dialogue est beaucoup plus intéressant : en effet, on peut remarquer une nette décroissance du nombre de disfluences au fur et à mesure que la conversation progresse. Le premier quart contient 37 % des disfluences alors que le dernier quart n'en contient que 13,8 %. On pourrait émettre l'hypothèse que cette décroissance est liée à une adaptabilité du locuteur au contexte d'énonciation et à son interlocuteur. À l'inverse, il semblerait probable, du fait de ces résultats, que certaines conversations puissent afficher une tendance inverse avec un accroissement du nombre de disfluences en fin de conversation, si l'enjeu de la conversation n'a pas abouti par exemple (nous nous plaçons pour cette hypothèse dans le cadre strict de conversations clients/agents dans un centre d'appel).

Progression des tours de parole	0 %-25 %	25 %-50 %	50 %-75 %	75 %-100 %
Proportion de disfluences	24,9 %	25,1 %	25,3 %	24,7 %

TABLEAU 4.7 – Répartition des disfluences d'édition en fonction de la progression du discours au sein des tours de parole. Progression à 0 % : le tour de parole débute ; progression à 100 % : le tour de parole est achevé.

Progression des conversations	0 %-25 %	25 %-50 %	50 %-75 %	75 %-100 %
Proportion de disfluences	37,0 %	26,3 %	22,9 %	13,8 %

TABLEAU 4.8 – Répartition des disfluences d'édition en fonction de la progression de la conversation. Progression à 0 % : la conversation débute ; progression à 100 % : la conversation est achevée.

À l'échelle des tours de parole, il est intéressant de constater que 16 % des disfluences d'édition sont positionnées sur le premier mot d'un tour de parole, ce qui nous a amenée à étudier les caractéristiques des tours de parole commençant par une disfluence. En effet, le positionnement des disfluences au sein des tours de parole semble également déterminant et fortement lié à la longueur de ces derniers. Nous avons comparé, au sein du corpus VOXDISS, les tours de parole (TP) qui débutent par une disfluence d'édition (5,37 % des TP) des autres tours de parole (94,63 % des TP). Lorsqu'un tour de parole débute par une disfluence, il contient en moyenne 26 mots ; il n'en contient que 13,2, en moyenne, lorsque ce n'est pas le cas. De plus, comme nous pouvons le voir dans le tableau 4.9 *infra*,

la proportion des TP très courts (moins de 3 mots) par rapport à ceux beaucoup plus long (au moins 32 mots) est très contrastée.

Longueur	TP dis-deb	TP null-deb
$\leq$ 3 mots	1,9 %	29,1 %
$\geq$ 32 mots	26,4 %	9,3 %

TABLEAU 4.9 – Répartition des tours de parole commençant par une disfluence (TP dis-deb) et des tours de parole ne commençant pas par une disfluence (TP null-deb) en fonction de leur longueur exprimée en nombre de mots.

4.3.2 Caractérisation acoustico-prosodique des disfluences

Dans cette section nous nous penchons sur les spécificités acoustico-prosodiques des contextes contenant une disfluence *versus* les contextes non-disfluents. L'hypothèse de travail est la suivante : les disfluences peuvent-elles être vues comme des « ruptures » discursives, des éléments perturbateurs au niveau local, qu'il s'agisse de disfluences d'édition ou de *fillers* comme les hésitations vocaliques ou les marqueurs discursifs ? Existe-t-il des marqueurs acoustico-prosodiques nous permettant d'identifier la zone disfluente et sa relation avec le contexte lexical environnant ? Des études antérieures (Vasilescu *et al.*, 2010a, Shriberg, 1999) ont par exemple pointé les différences entre les caractéristiques acoustiques des voyelles d'hésitation et celles des voyelles intra-lexicales. En partant de ces constatations nous proposons de fournir dans un premier temps une analyse descriptive de nos données (durées des phonèmes) et dans un second temps une analyse des spécificités acoustiques et prosodiques des disfluences par rapport au contexte lexical proche et au corpus en général.

Ces travaux permettent de caractériser le corpus VOXDISS selon la distribution des phonèmes qui le composent, pour ce qui est de leur durée. La figure 4.6 *infra* présente ainsi la distribution des phonèmes ayant une durée comprise entre 30 ms et 300 ms ; 30 ms étant la durée minimale pour laquelle le système peut localiser un phone dans le signal acoustique (modèles de Markov cachés à trois états avec transition entre états correspondant à 10 ms).

La courbe de distribution fait clairement apparaître un pic de valeurs à 30 ms ; environ 20 % des phonèmes de notre corpus présentent cette durée. En accord avec Adda-Decker (2006) et Nguyen et Adda-Decker (2013), ce pic est très caractéristique des données de parole conversationnelle. Avec les mêmes mesures, Adda-Decker (2006) a calculé pour le français que cette proportion passait à environ 8 % pour de la parole préparée d'émissions journalistiques *versus* plus de 18 % pour de l'oral spontané de conversations téléphoniques. En effet, ce pic dans la

distribution rend compte d'un fort décalage entre prononciation attendue par le système et prononciation effective, mis en avant de la manière suivante :

> « Plus le maximum de la distribution se trouve décalé vers la gauche (*i.e.* vers les segments courts) plus la courbe indique un risque de désaccord entre prononciations standard attendues et prononciations effectivement réalisées par les locuteurs, ces réalisations présentant alors potentiellement des réductions temporelles. Pour ces dernières on peut chercher des explications, la première articulatoire : le phonème, facile et rapide à réaliser, a une durée intrinsèque courte. Un débit rapide avec une articulation incomplète réduit alors encore cette durée. Une deuxième explication peut venir des fréquences d'occurrence : une observation très fréquente est une observation à contenu d'information faible et risque donc d'être négligée dans le signal acoustique. Il est fort plausible que dans une parole spontanée ces deux facteurs se trouvent combinés. » Adda-Decker (2006)

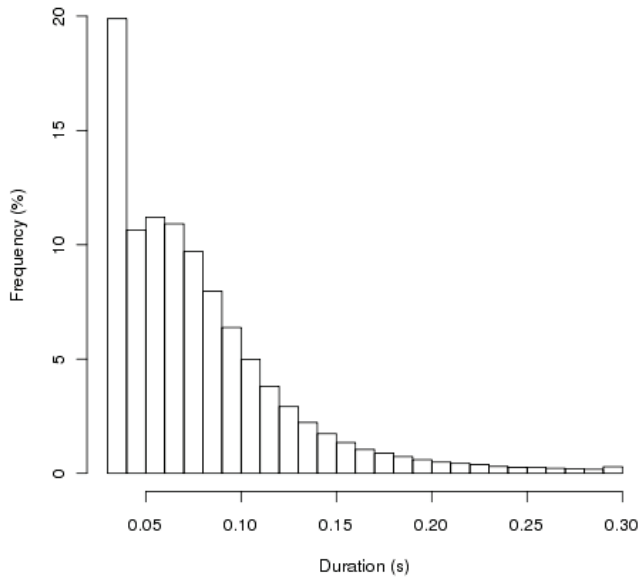


FIGURE 4.6 – Distribution de la durée des segments phonémiques en secondes au sein du corpus VOXDISS

L'alignement automatique permet d'obtenir les pauses silencieuse présentes dans le signal de parole du corpus, ainsi que la durée qui leur est associée. Comme présenté en section 4.2.2, nous avons reporté dans le module d'extraction acoustique cette information au niveau des mots, indiquant pour chacun d'entre eux la

durée du silence qui les précèdent et celle du silence qui les suit dans le flux de parole.

Le tableau 4.10 *infra* présente la répartition des mots en fonction de leur longueur syllabique au sein du corpus VOXDISS. La longueur syllabique des mots est une information issue de leur prononciation effective, non pas de leur longueur syllabique théorique. Par exemple, théoriquement, le mot « c'est-à-dire » comporte trois syllabes. Cependant, en prenant en compte ses différentes variantes de prononciations effectives dans le corpus VOXDISS, ce mot correspond en fait à deux *tokens*, « c' » et « est-à-dire », prononcés [sadir] : le premier token « c' » se voit donc assigner une longueur syllabique de 0, et le *token* « est-à-dire » une longueur syllabique de 2.

Lg syll.	**% Mots $\in$ dis**	% Mots $\notin$ dis	% Mots total
0	18,0	27,4	19,1 (25 308)
1	55,5	59,4	55,8 (74 218)
2	18,4	9,9	17,5 (23 102)
3	5,6	2,5	5,2 (6 885)
4	1,9	0,6	1,8 (2 359)
5	0,4	0,2	0,4 (518)
6	0,2	0,0	0,2 (171)
	100 % (15 624)	100 % (117 015)	100 % (132 639)

TABLEAU 4.10 – Répartition des mots en fonction de leur longueur syllabique et selon leur appartenance ou non à une disfluence au sein du corpus VOXDISS.

Une attention particulière a également été portée à la présence de fragments de morphèmes dans la parole, dans la mesure où ceux-ci peuvent être de forts indicateur de disfluence. Dans le corpus VOXDISS 0,6 % des mots sont des fragments, parmi lesquels près de 77 % apparaissent à l'intérieur d'un segment disfluent (dont 99 % dans le *reparandum*) : bien que ce constat plaide en faveur de l'utilisation de la présence d'amorces comme un indice fort de la présence de disfluence, celles-ci sont trop peu nombreuses dans ces données pour pouvoir les utiliser en lien avec d'autres informations ou pour les interpréter dans une caractérisation des disfluences d'édition. De plus, les schémas d'apparition des amorces dans le discours (*cf.* définition et description des amorces dans le chapitre 2, section 2.2.2) montrent une très forte ressemblance avec ceux des disfluences d'édition (répétitions et auto-corrections notamment), que ces derniers contiennent ou non une amorce. Nous faisons l'hypothèse que la détection des disfluences d'édition peut se faire indépendamment de la présence d'amorces. Par ailleurs, nous évaluons également nos travaux menés sur la détection automatique de disfluences sur des sorties de Reconnaissance Automatique de la Parole (RAP) : les moteurs de RAP,

sauf de rares exceptions, ne transcrivent pas les amorces, assignant à cet espace d'énonciation un mot complet présent dans le lexique utilisé.

Enfin, nous avons procédé à l'étude des mesures de fréquence fondamentale (F_0) dans le corpus VOXDISS. Ces analyses consistent en une étude des moyennes de F_0 au niveau du mot, en lien avec la présence de disfluences et de manière contrastive selon le segment à l'intérieur duquel se trouve le mot. Notre hypothèse de départ, en accord avec les observations menées dans la littérature (*cf.* par exemple Hokkanen (2001), ou plus récemment Christodoulides et Avanzi (2014) sur le français), est que la F_0 augmente au sein du *reparans*, par rapport aux mots inclus dans le *reparandum*. Le tableau 4.11*infra* résume les moyennes de F_0 au niveau du mot en fonction de la place du mot dans la parole. Les données montrent que contrairement à notre hypothèse de départ, la F_0 apparaît plus élevée pour le *reparandum* que pour le *reparans*.

Position du mot	Hors disfluence	reparandum	reparans
F_0 **moyenne**	187.5	185.6	183.8

TABLEAU 4.11 – F_0 moyenne des mots en fonction de leur place dans et hors disfluences d'édition dans le corpus VOXDISS.

De plus, la variation de F_0 entre les mots hors disfluences et les mots à l'intérieur d'une disfluence est beaucoup plus importante si l'on considère les mots débutant une disfluence uniquement (*cf.* tableau 4.12 *infra*). Enfin, les variations de F_0 sont beaucoup plus importantes au sein du *reparans*, qui présente à la fois une moyenne plus basse que celle du *reparandum* sur les mots d'« attaque » de la disfluence et une moyenne plus élevée concernant les autres mots du segment disfluent. Ces analyses font émerger les caractéristiques différentes de F_0 en mettant en avant une rupture de F_0, dans le flux de parole, entre l'ébauche (*reparandum*) et l'achèvement (*reparans*) d'une disfluence d'édition.

Position du mot	reparandum (premier mot)	reparandum (autre)	reparans (premier mot)	reparans (autre)
F_0 **moyenne**	181,4	187,8	177,2	189,5

TABLEAU 4.12 – F_0 moyenne des mots en fonction de leur place au sein des disfluences d'édition dans le corpus VOXDISS.

4.3.3 Caractérisation linguistique de la parole disfluente

Nous nous sommes penchée sur le comportement des disfluences à travers leur contexte lexical. En effet, dans l'objectif final d'utiliser les disfluences pour soute-

nir d'autres tâches d'extraction d'information sur des données conversationnelles issues de l'oral, il faut au préalable pouvoir identifier ces phénomènes disfluents. Nous avons vu en section 2.4 que les systèmes de détection des disfluences se basent soit sur des indices acoustiques, soit sur des indices lexicaux, soit sur une combinaison des deux. Dans cette perspective, quelles sont les caractéristiques lexicales des disfluences et de leur contexte immédiat ? Ces caractéristiques sont-elles suffisamment discriminantes (par rapport à des segments non disfluents) pour être de bons candidats de paramètres pour le développement d'un système d'identification adapté à des données orales conversationnelles (*cf.* chapitre 5) ?

Nous examinons ces contextes lexicaux grâce à une extraction des bigrammes entourant les événements disfluents, en nous inspirant des méthodes d'analyses proposées par Vasilescu *et al.* (2010b) appliquées à l'étude des fonctions de l'hésitation vocalique « euh ». En effet, les auteures font l'hypothèse que l'hésitation ainsi que les marqueurs discursifs classiques tels que « bon », « ben », « alors », *etc.* peuvent être indicateurs du travail de reformulation et de ce fait se retrouver dans des contextes porteur d'information particulièrement saillante. Cette spécificité est intéressante dans le cadre des systèmes question/réponse. À cet effet, les auteures de l'article sus-cité ont analysé les contextes d'occurrence de ces items dans le corpus RITEL (*cf.* section 3.2.2). En nous appuyant sur cette méthodologie nous avons procédé à une analyse des contextes d'occurrences des disfluences dans le corpus VOXDISS. L'hypothèse de travail est que les disfluences se retrouvent dans des contextes privilégiés où probablement le locuteur produit des ruptures dans une démarche de mise en mot de l'information pertinente pour son interlocuteur.

4.3.3.1 Méthode

Nous avons étudié le contexte lexical immédiat des zones disfluentes du corpus VOXDISS selon trois niveaux : les formes des mots prononcés telles qu'écrites dans les transcriptions manuelles, les lemmes correspondants à ces formes, et enfin leurs parties du discours. Pour cela, à partir des transcriptions manuelles, nous avons dans un premier temps réalisé une analyse morpho-syntaxique afin d'obtenir le lemme et la partie du discours de chacun des *tokens* du corpus, puis nous avons généré les bigrammes pour chacune des soixante conversations composant les données, et ce par tour de parole. Nous avons suite à ces pré-traitements pu étudier la fréquence d'apparition des bigrammes présents dans les contextes lexicaux amonts et avals des disfluences d'édition, dans le but de faire émerger des patrons de régularité permettant de caractériser les disfluences, selon leur catégorie.

Après comparaison de plusieurs étiqueteurs, nous avons utilisé le `Treetagger` pour ces analyses. Dans la mesure où cet outil n'a pas été adapté pour les données orales, et même si notre analyse (*cf.* section 4.2.3) a montré qu'il obtenait une bonne précision sur le corpus VOXDISS, l'étiquetage peut comporter des erreurs

de reconnaissance, en plus des erreurs classiques des outils d'étiquetage morpho-syntaxique. Par exemple, la forme « pouvoir » peut selon le contexte être un verbe à l'infinitif ou un nom commun singulier : cette ambiguïté inhérente à la langue française peut mener à une erreur d'étiquetage. La chaîne de traitement que nous avons développée dans le cadre de ces travaux permet en sortie de travailler sur deux formats différents, selon les besoins : un format `text/csv` classique et un format `text/xml` permettant un certain nombre de requêtes sur les bigrammes, pour lequel la figure 4.7 *infra* détaille la structure selon le schéma de documentation `Relax NG Compact`[3].

```
element conversation {
    element turn {
        attribute rank { xsd:positiveInteger },
        attribute speaker { "spk1" | "spk2" | "spk3" | "spk4" | "spk5" | "spk6" },
        attribute startTime { xsd:decimal },
        attribute endTime { xsd:decimal },
        attribute words { xsd:positiveInteger },
        attribute wordsInsideDisfluence { xsd:positiveInteger }
        element bigram {
            attribute rank { xsd:positiveInteger },
            element word {
                attribute lemma { text },
                attribute pos { text },
                text
            }+
        }+
    }+
}
```

FIGURE 4.7 – Documentation du format de structuration XML de traitement des bigrammes en Relax NG Compact.

4.3.3.2 Résultats

Nous avons étudié les bigrammes présents à la frontière des zones disfluentes, en séparant les contextes droits et gauches dans le but de faire émerger des régularités dans les patrons lexicaux. Pour chacun des deux contextes, nous avons regardé les bigrammes de formes, de lemmes et de parties du discours (ou POS pour *Part Of Speech*).

Le tableau 4.13 *infra* présente les bigrammes de POS les plus fréquents apparaissant à gauche des disfluences. Nous observons que le contexte amont des

3. *Regular Language for XML Next Generation* : schéma de documentation `XML`.

disfluences est hétérogène. Les disfluences combinées présentent cependant une particularité : le troisième bigramme apparaissant le plus fréquemment est « Nom commun – Interjection ». Ce bigramme n'apparaît dans aucun des autres contextes pour les autres classes de disfluences (d'où son absence dans le tableau 4.13 *infra*, qui présente les contextes communs aux différentes classes de disfluences d'édition) et pourrait être envisagé comme indice pour l'identification et/ou la caractérisation des disfluences combinées.

Les faux-départs présentent également quelques particularités : en effet, le bigramme « Verbe – Verbe » apparaît dans les cinq bigrammes les plus fréquents pour toutes les classes de disfluences excepté celui-ci. On peut également noter que le bigramme « Adjectif – Nom » semble caractériser les faux-départs, dans la mesure où il est exclu du contexte des autres classes de disfluences.

POS bigramme	**Classe de disfluence**	**Rang**	**Nb d'occurrences**
PRO :PER VER :pres	Répétition	1	59
	Auto-corr.	1	58
	Faux-départ	2	18
	Dis. combinée	1	42
DET :ART NOM	Répétition	2	55
	Auto-corr.	2	50
	Faux-dép.	1	24
	Dis. combinée	2	34
VERpres ADV	Répétition	3	31
	Auto-corr.	3	30
	Faux-départ	3	15
	Dis. combinée	4	20

TABLEAU 4.13 – Bigrammes de parties du discours les plus fréquentes communes aux quatre classes de disfluences dans le corpus VOXDISS.

Le contexte droit des disfluences dans ce corpus semble plus informatif pour la caractérisation des disfluences. Si l'on regarde les trois bigrammes de POS les plus fréquents, un seul concerne toutes les classes de disfluences (le bigramme « Pronom personnel – Verbe »). Tout comme mis en avant pour le contexte précédant les disfluences, les faux-départs semblent être la classe la plus hétérogène : elle est la seule a ne pas avoir le bigramme « Article – Nom » dans son contexte droit. En plus de cela, les deux bigrammes « Interjection – Adverbe » et « Interjection – Pronom personnel » sont présents mais n'apparaissent pas pour les autres classes de disfluences. Nous allons creuser ces informations afin d'expliquer ces spécificités. En effet, les caractéristiques du contexte aval des faux-départs peuvent être liées au fait que le *reparans* de cette disfluence d'édition est vide

(contrairement aux répétitions ou aux auto-corrections) ou qu'il y a une absence de lien sémantique entre le *reparandum* et le *reparans* ; du moins un fort changement de sens et/ou de formulation.

4.3.4 Profils de locuteurs : stratégies discursives et disfluences

Les disfluences ont également été considérées en relation avec le profil des locuteurs. Le tableau 4.14 *infra* résume les principaux paramètres des conversations en fonction des profils de locuteurs[4]. Nous nous sommes focalisée pour l'étude du corpus VOXDISS sur les profils de locuteurs Clients et Agents dans la mesure où les autres interactions (agents parlant entre eux ou clients parlant entre eux) représentent une minorité d'interactions dans ces données.

	TP Agent	**TP Client**
Nombre	5 493 (50,8 %)	4 897 (45,3 %)
Durée	4h27	4h30
Durée moyenne (sec.)	02'92	03'30
Nombre mots	64 057 (47,8 %)	64 617 (48,3 %)
Nombre moyen mots	11,6	13,1
Débit moyen (mots/sec.)	3,7	3,6

TABLEAU 4.14 – Caractéristiques principales des tours de parole (TP) du corpus VOXDISS par profil de locuteurs.

Les données montrent un équilibre relatif entre les tours de parole de la classe Agents et de la classe Clients, conformément aux précédentes observations : les interactions entre ces deux types de locuteurs sont prédominantes. Cet équilibre concerne la totalité des paramètres mesurés : le nombre de tours de parole, le ratio de parole, la durée des tours de parole. Les paragraphes suivants consistent en une description approfondie de la présence d'événements disfluents à travers les tours de parole et selon le profil de locuteur.

Nous avons mis en avant l'équilibre relatif entre les tours de parole des agents et des clients. D'un côté, la figure 4.8 *infra* montre que cet équilibre se retrouve également dans la répartition des classes de disfluences.

D'un autre côté, on constate que les profils de locuteurs présentent un déséquilibre quant au nombre de mots prononcés : tout d'abord, 25 % des mots prononcés par les locuteurs appartenant au profils Clients se trouvent à l'intérieur d'événements disfluents, contre 17,29 % pour les locuteurs appartenant au profil Agents. De plus, le tableau 4.15 *infra* montre que la taille des disfluences mesurée en nombre de mots est très inégale selon ces deux profils.

4. Certains tours assignés à la classe Agents ou Clients peuvent êtres vides (par exemple en cas de rire).

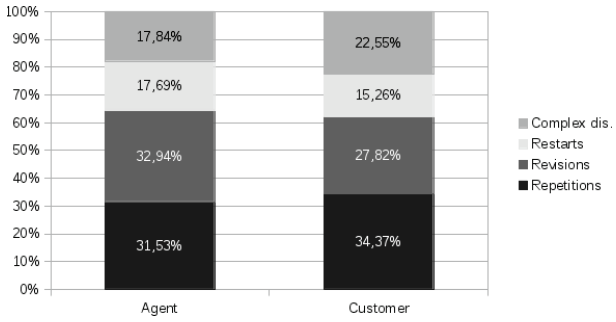


FIGURE 4.8 – Distribution des quatre classes de disfluences annotées par profil de locuteurs au sein du corpus VOXDISS.

Classe de disfluences	Nb mots Agent	Nb mots Client
Répétitions	1 378 (2,1 %)	2 276 (3,5 %)
Auto-corrections	1 984 (3,1 %)	2 427 (3,6 %)
Faux-départs	810 (1,3 %)	1 202 (1,9 %)
Disfluences combinées	1 661 (2,6 %)	3 179 (4,9 %)

TABLEAU 4.15 – Taille lexicale des classes de disfluences du corpus VOXDISS par profil de locuteurs.

	TP disfluents Agents	TP disfluents Clients
Nombre	954	1 214
Nombre mots	26 635	35 838
Nombre mots dans disfluence	5 833 (21,90 %)	9 084 (25,35 %)
Nombre mots hors disfluence	20 802 (78,10 %)	26 754 (74,65 %)

TABLEAU 4.16 – Caractéristiques principales des tours de parole (TP) disfluents du corpus VOXDISS par profil de locuteurs (Agents *versus* Clients).

Les disfluences des Agents et des Clients suivent la même tendance : les répétitions sont prédominantes, suivies des auto-corrections, des disfluences combinées et enfin des faux-départs. À la lumière de ces distributions, quelques observations s'imposent : de manière générale les stratégies discursives des Agents et des Clients semblent être différentes : les Clients semblent être plus disfluents dans la mesure où 25 % des mots qu'ils prononcent interviennent dans un événement disfluent, contre 21 % pour les Agents. Ils sont aussi auteurs de plus de tours de parole que les Agents.

4.4 Discussion

Nous avons présenté, dans ce chapitre de caractérisation des disfluences dans les conversations, d'une part la conception et la réalisation d'une chaîne de traitement et d'enrichissement des données orales conversationnelles – générique et extensible – et d'autre part des analyses de fouille de données et d'étude des phénomènes disfluents dans des données enrichies. Ces études permettent de faire émerger des spécificités du corpus étudié (le corpus VOXDISS, constitué de conversations issues des centres d'appel d'EDF) et de valider ou d'invalider des hypothèses théoriques (ou vérifiées dans d'autres données) quant à l'actualisation des disfluences d'édition dans la parole spontanée, notamment :

- la distribution et le comportement des phénomènes d'oral spontané au sein de conversations issues de centres d'appels correspond bien aux observations établies sur d'autres corpus d'oral spontané ;
- l'apparition de disfluences n'est pas un phénomène complètement aléatoire et certains contextes discursifs ou locaux permettent de les caractériser. Notamment, les disfluences sont réparties de manière homogène au sein des tours de parole, mais apparaissent de manière décroissante au fur et à mesure que la conversation progresse ; les tours de parole débutant par une disfluence d'édition sont beaucoup plus longs que les autres (y compris que ceux contenant une ou plusieurs disfluences en position non initiale) ;
- les disfluences d'édition et plus précisément les éléments qui les structurent présentent certaines caractéristiques acoustico-prosodiques ou lexicales distinctes de celles inhérentes aux segments non disfluents.

Nous avons lié la présence de disfluences à de nombreuses informations acoustiques, lexicales et discursives : de manière générale, les disfluences d'édition sont un phénomène difficile à appréhender, soumis à de nombreuses variations et dépendant d'un grand nombre d'informations extérieures. Nous avons toutefois pu distinguer quelques caractéristiques lexicales (émergence de certaines classes de mots autour des disfluences) ou acoustiques (variations de fréquence fondamentale entre l'ébauche et l'achèvement des disfluences). Nous avons également caractérisé les disfluences en prenant en compte le contexte dialogique (stratégies discursives en fonction des profils de locuteurs, positionnement des disfluences dans la progression du dialogue, *etc.*).

Les résultats de cette étude sont directement réutilisés dans nos travaux sur la détection automatique de disfluences dans les données orales conversationnelles. La chaîne de traitement et d'enrichissement des données orales conçue pour leur analyse permet d'injecter des traits caractéristiques des disfluences dans un système d'apprentissage supervisé que nous avons développé pour évaluer de nombreuses tâches de détection automatique ; ces tâches sont également conçues non seulement en fonction des enjeux scientifiques et applicatifs liés à la détection de disfluences dans la parole, mais aussi eu égard aux résultats des analyses menées

dans ce chapitre. Nous essaierons de répondre transversalement à la question suivante : est-ce que des traits qui semblent très caractéristiques des disfluences (et inversement) sont également des indices forts pour détecter automatiquement ces phénomènes ?

Chapitre 5

Détection automatique de disfluences d'édition en contexte conversationnel

Sommaire

Détecter des phénomènes d'oral spontané, et en particulier les disfluences d'édition en contexte conversationnel, relève d'enjeux théoriques et applicatifs importants pour la modélisation et le traitement de données issues de l'oral, comme nous avons pu le voir dans les chapitres 1 et 2. Il s'agit également de phénomènes particulièrement difficiles à détecter dès lors que l'on prend en compte la totalité de leurs réalisations dans l'énoncé (*cf.* chapitre 2), ce qui est perceptible dans le fait que les données orales conversationnelles, à un niveau global, tout comme les disfluences, à un niveau local, possèdent des caractéristiques relativement ténues à faire émerger (*cf.* chapitre 4). Nous explorons dans ce chapitre 5, eu égard à ces considérations théoriques et analyses de corpus, la faisabilité et les performances d'une méthode de détection automatique des disfluences dans des conversations en français issues de centres d'appel (corpus VoxDiss, *cf.* chapitre 3).

5.1 Introduction : enjeux de la détection automatique de disfluences dans des données de centres d'appels

Ce chapitre est dédié au développement d'un système capable de détecter les disfluences dans le contexte particulièrement difficile des données issues de centres d'appels (SADDE, pour « Système Automatique pour la Détection des Disfluences d'Édition »). L'approche présentée ici contribue à un défi important : les analyses sont en effet conduites dans un contexte industriel grâce à des données de centres d'appels produites par l'entreprise EDF. En effet, le contenu informationnel de ces données est très important pour les industriels, dans la mesure où il peut être utilisé pour améliorer la connaissance clients et la relation clients/entreprise. En particulier, il peut contribuer à identifier les interactions dysfonctionnelles et les bonnes pratiques comme les bonnes ou mauvaises stratégies de communication entre clients et agents, permettant ainsi l'élaboration d'une stratégie de communication optimisée pour les agents EDF. La finalité de ce travail est de développer un système de détection des disfluences permettant l'amélioration de la lisibilité de ce type de données mais aussi l'amélioration de l'efficacité des méthodes de fouille de texte actuellement mises en place chez EDF (Clavel *et al.*, 2013). Deux applications sont plus particulièrement ciblées : l'amélioration de la lisibilité des données transcrites, en particulier en lien avec l'interface de visualisation déjà développée dans le projet Vox Factory (Cailliau et Giraudel, 2008), et l'aide au fonctionnement de modules de Traitement Automatique des Langues (TAL) appliqués en aval sur les données, avec l'intégration ou la suppression des événements disfluents identifiés.

Nous décrivons tout d'abord, en section 5.2, les tâches de détection des disfluences que nous avons conçues, élaborées en lien avec ces enjeux scientifiques et applicatifs. Le corpus utilisé pour mener nos expérimentations, issu des centres d'appels EDF, est décrit en section 5.3. La section 5.4 est dédiée à la description du système nommé SADDE ; les résultats obtenus par notre système sur chacune des tâches sont présentés et commentés en section 5.5. Nous appliquons ensuite notre système aux sorties de reconnaissance automatique de la parole afin d'étudier la dégradation des résultats sur ces données (*cf.* section 5.6). Enfin, la section 5.7 clôture ce chapitre par une discussion d'ensemble à propos de ces travaux sur la détection automatique de disfluences d'édition sur des données orales conversationnelles.

5.2 Description des tâches de détection en fonction des enjeux scientifiques et applicatifs

Nous décrivons dans cette section l'ensemble des tâches de détection de disfluences que nous avons mises en place. Ces tâches sont envisagées pour répondre aux enjeux scientifiques et applicatifs qui relèvent de la détection de disfluences dans les données orales conversationnelles, que nous avons présentés dans le chapitre 1. L'identification des disfluences relève en effet de deux dimensions :

1. une meilleure modélisation de la parole par le biais de l'étude des moyens mis en place par des locuteurs pour construire leur discours et gérer l'interaction dans des dialogues homme-homme ; il s'agit donc d'un enjeu d'analyse des donnés ;
2. une amélioration des applications de TAL sur des données d'oral spontané et une meilleure lisibilité des données par des humains (dès lors que l'on n'est pas dans une perspective d'étude de la parole) ; il s'agit donc d'un enjeu de nettoyage des données.

Dans cette perspective, nous avons développé douze tâches de détection, qui répondent soit à l'enjeu d'analyse **(A)**, soit à l'enjeu de nettoyage **(N)**, soit aux deux enjeux conjointement **(AN)** ; pour chacun de ces trois objectifs, nous détaillons quels segments de la structure interne des disfluences d'édition (telle que décrite dans le chapitre 2, section 2.3.1) sont identifiés. Les tâches de détection sont conçues :

- soit de manière générique, c'est-à-dire sans distinction de classes de disfluences (répétitions, auto-corrections, faux-départs, disfluences combinées, *cf.* chapitre 2, section 2.3.2) : il s'agit des tâches $*$.A, $*$.B ou $*$.C ;
- soit en distinguant les répétitions, les auto-corrections, les faux-départs et les disfluences combinées : il s'agit des tâches $*$.X, $*$.Y ou $*$.Z.

La duplication des tâches entre celles distinguant les différentes classes de disfluences et les tâches génériques nous permet par ailleurs d'étudier la disparité des résultats en fonction des classes, qui présentent des caractéristiques différentes les unes des autres (*cf.* chapitre 4, section 4.3).

5.2.1 Détecter pour analyser les disfluences

Pour mieux étudier les stratégies discursives des locuteurs, la manière dont ils modifient leur énoncé et la raison pour laquelle ils le modifient, il est indispensable d'identifier et de lier le *reparandum* et le *reparans*, avec une inclusion optionnelle de l'*interregnum*, c'est-à-dire d'établir un lien entre ébauche et achèvement. Les tâches 1.A/1.X et 3.C/3.Z relèvent exclusivement de cet objectif d'analyse (**objectif A**). Le tableau 5.1 *infra* présente un exemple de segment disfluent avec étiquetage de référence pour ces quatre tâches.

Outre une détection considérant les différentes classes de disfluences, deux stratégies d'identification sont ici étudiées (tâches 1.A et 1.X *versus* tâches 3.C et 3.Z). Avec les tâches 1.A et 1.X, une seule séquence d'annotation est envisagée : l'ébauche n'est pas distinguée de l'achèvement, et la même étiquette est assignée à tous les mots inclus dans la disfluence (*reparandum*, *interregnum* et *reparans* confondus). *A contrario*, avec les tâches 3.C et 3.Z, deux séquences sont envisagées : ébauche et achèvement sont distincts et portent des étiquettes différentes. L'*interregnum* est également exclu de l'étiquetage, dans la mesure où il n'est pas indispensable à des fins d'analyse ; la priorité est ici mise sur le lien entretenu entre *reparandum* et *reparans*.

Une détection considérant ou non les différentes classes de disfluences est également examinée (tâches 1.A et 3.C *versus* tâches 1.X et 3.Z). La tâche 3.C, par exemple, comporte deux étiquettes : une pour la détection du *reparandum* (toutes classes confondues) et une pour la détection du *reparans* (toutes classes confondues également). La tâche 3.Z, reflet de la tâche 3.C mais avec distinction des classes de disfluences dans la détection, comprend sept étiquettes : quatre pour le *reparandum* (couvrant ainsi les quatre classes de disfluences) et trois pour le *reparans* (couvrant toutes les classes excepté les faux-départs, pour lesquels la disfluence ne présente pas d'achèvement ; le *reparans* est donc inexistant).

5.2.2 Détecter pour nettoyer les données des ébauches de disfluences

Afin d'améliorer les applications de traitement automatique sur des données issues de l'oral (**objectif N**), il est nécessaire de supprimer deux séquences associées, le *reparandum* et l'*interregnum*, soit l'ébauche de la disfluence. Les tâches 2.B/2.Y et 2.C/2.Z relèvent exclusivement de cet objectif de nettoyage (*cf.* tableau 5.2 *infra* pour un exemple de disfluence avec étiquetage de référence pour ces quatre

Tâche	Étiquetage de référence	Étiquettes	NbE
1.A	je voudrais payer < par euh enfin par > carte	<rpd+int+rpr>	1
1.X	je voudrais payer < par euh enfin par > carte	<rpd+int+rpr{cls}>	4
3.C	je voudrais payer < par > euh enfin < par > carte	<rpd> <rpr>	2
3.Z	je voudrais payer < par > euh enfin < par > carte	<rpd{cls}> <rpr{cls}>	7

TABLEAU 5.1 – Exemple d'énoncé annoté en disfluence d'édition pour les tâches de détection répondant à un objectif d'analyse exclusivement, où $\{cls\}$ représente les différentes classes de disfluences prises en compte (soit : répétitions, auto-corrections, faux-départs et disfluences combinées) et NbE correspond au nombre d'étiquettes.

tâches). De la même manière que pour les tâches de détection relevant de l'objectif **A**, deux niveaux de structuration sont étudiés ici, *via* des tâches distinctes. Pour les tâches 2.B et 2.Y, une seule séquence d'étiquetage est définie pour l'ensemble de l'ébauche, *reparandum* et *interregnum* confondus. *A contrario*, pour les tâches 2.C et 2.Z, une distinction est opérée entre *reparandum* et *interregnum*, dans une perspective de structuration de l'ébauche.

Tâche	Étiquetage de référence	Étiquettes	NbE
2.B	je voudrais payer < par euh enfin > par carte	<rpd+int>	1
2.Y	je voudrais payer < par euh enfin > par carte	<rpd+int{cls}>	4
2.C	je voudrais payer < par > < euh enfin > par carte	<rpd> <int>	2
2.Z	je voudrais payer < par > < euh enfin > par carte	<rpd{cls}> <int{cls}>	7

TABLEAU 5.2 – Exemple d'énoncé annoté en disfluence d'édition pour les tâches de détection répondant à un objectif de nettoyage exclusivement, où $\{cls\}$ représente les différentes classes de disfluences prises en compte (soit : répétitions, auto-corrections, faux-départs et disfluences combinées) et NbE correspond au nombre d'étiquettes.

5.2.3 Détecter pour analyser et nettoyer

Enfin, nous envisageons une série de tâches de détection utile à la fois pour un enjeu d'analyse, dans la mesure où l'on peut lier les parties d'ébauche – *reparandum* et *interregnum* – et d'achèvement – *reparans* – de la disfluence, et pour un enjeu de nettoyage, dans la mesure où l'on identifie la totalité de l'ébauche, que l'on peut ainsi supprimer (**objectif AN**). Les tâches 1.B/1.Y et 1.C/1.Z relèvent de ce double objectif (*cf.* tableau 5.3 *infra* pour un exemple de disfluence avec étiquetage de référence pour ces quatre tâches). Pour cet objectif également, nous testons deux

stratégies de structuration différentes pour détecter la disfluence. Concernant les tâches 1.B et 1.Y, deux séquences sont envisagées : l'ébauche (union du *reparandum* et de l'*interregnum*) et l'achèvement (*reparans*).

T	Étiquetage de référence	Étiquettes	NbE
1.B	[...] payer <par euh enfin><par> carte	<rpd+int><rpr>	2
1.Y	[...] payer <par euh enfin><par> carte	<rpd+int{cls}><rpr{cls}>	7
1.C	[...] payer <par><euh enfin><par> carte	<rpd><int><rpr>	3
1.Z	[...] payer <par><euh enfin><par> carte	<rpd{cls}><int{cls}><rpr{cls}>	10

TABLEAU 5.3 – Exemple d'énoncé annoté en disfluence d'édition pour les tâches de détection répondant à un double objectif d'analyse et de nettoyage, où $\{cls\}$ représente les différentes classes de disfluences prises en compte (soit : répétitions, auto-corrections, faux-départs et disfluences combinées) et NbE correspond au nombre d'étiquettes.

5.2.4 Discussion

Nous avons conçu plusieurs tâches de détection automatique des disfluences d'édition dans des données orales conversationnelles. L'élaboration de ces tâches répond à des objectifs de traitements liés à des enjeux scientifiques et applicatifs. Nous avons défini deux objectifs principaux : analyse et nettoyage des données orales.

Ces tâches de détections, pour lesquelles le tableau 5.4 *infra* propose une vue d'ensemble, répondent donc à l'un de ces deux objectifs ou aux deux de manière conjointe (objectifs **A**, **N** ou **AN**). Ce tableau résume également les autres points clef de conception de ces tâches : présentation en fonction du segment disfluent que nous souhaitons détecter (détection portant sur la totalité de la disfluence, son ébauche, son achèvement ou segment hybride) et en fonction des étiquettes d'annotation séquentielle impliquées dans ces tâches de détection. Enfin, nous nous attachons à évaluer deux séries de tâches : avec ou sans distinction des classes constituant la typologie des disfluences d'édition (répétitions, auto-corrections, faux-départs, disfluences combinées).

L'évaluation distincte et comparative de ces expérimentations permet une analyse fine et novatrice de la tâche de détection des disfluences d'édition, en lien avec différentes stratégies d'identification (différents degrés de structuration des séquences). Ces expériences nous permettent par exemple d'observer si l'on peut utiliser l'*interregnum* malgré sa faible représentation, son contenu hétéroclite, et la difficulté apparente dans la tâche d'annotation manuelle de ce segment. Nous pouvons également mesurer l'impact de l'ajout de certains segments sur la détection d'autres segments participant du même phénomène.

Tâche	Objectif	Étiquettes					Segment
		rpd	int	rpd+int	rpr	rpd+int+rpr	
1.A	A					✓	disfluence
1.B	AN			✓	✓		
1.C	AN	✓	✓		✓		
2.B	N			✓			ébauche
2.C	N	✓	✓				
3.C	A	✓			✓		hybride

(a) sans distinction de classes de disfluences.

Tâche	Objectif	Étiquettes $\{cls\}$					Segment
		rpd	int	rpd+int	rpr	rpd+int+rpr	
1.X	A					✓	disfluence
1.Y	AN			✓	✓		
1.Z	AN	✓	✓		✓		
2.Y	N			✓			ébauche
2.Z	N	✓	✓				
3.Z	A	✓			✓		hybride

(b) avec distinction de toutes les classes de disfluences, où $\{cls\}$ représente les différentes classes de disfluences prises en compte, soit : répétitions, autocorrections, faux-départs et disfluences combinées.

TABLEAU 5.4 – Vue d'ensemble des tâches de détection des disfluences d'édition avec (b) ou sans (a) distinction de classes, en fonction des objectifs (A = analyse, N = nettoyage, AN = analyse et nettoyage), des étiquettes d'annotation et des séquences impliquées.

5.3 Corpus d'expérimentation pour la détection automatique de disfluences

Les corpus dédiées à l'entraînement, au développement et à l'évaluation de SADDE sont issus du corpus VOXDISS. Ce corpus, décrit dans le chapitre 3 (section 3.3.2), est composé de soixante conversations issues des centres d'appels d'EDF. Il a été manuellement annoté en disfluences d'édition selon les principes également décrits en section 3.3.2 et nous y avons mené des analyses de corpus détaillées dans le chapitre 4 (section 4.3), afin de caractériser ces données et d'étudier la manière dont sont réalisées des disfluences sur un corpus d'oral spontané.

Nous décrivons en section 5.3.1 la méthode d'échantillonnage du corpus VOXDISS, en section 5.3.2 les caractéristiques contrastives des trois sous-ensembles ainsi produits. La section 5.3.3 est dédiée à la réalisation des disfluences au sein

de ces corpus d'entraînement, de développement et de test. Nous discutons en section 5.3.4 les particularités de nos corpus et les choix de méthode d'échantillonnage.

5.3.1 Méthode d'échantillonnage

Nous avons échantillonné le corpus VOXDISS, décrit en section 3.3.2, afin de créer des corpus d'entraînement (*train*), de développement (*dev*) et de test (*test*) pour le système de détection SADDE. Nous avons opéré une répartition classique des données pour ce type de tâche, à savoir allouer 80 % des données à l'entraînement du système, 10 % pour le développement et 10 % pour le test. Chaque sous-ensemble a par ailleurs été généré en tirant aléatoirement des appels au sein des quartiles du corpus VOXDISS, suivant une distribution homogène sur la durée des conversations. En effet, cette méthode permet de ne pas instaurer de biais dans l'échantillonnage du corpus, dans la mesure où aucun présupposé quant au contenu des données n'a été utilisé pour orienter leur découpage. Les sous-ensembles ainsi générés, ainsi que leurs principales caractéristiques, sont présentés dans le tableau 5.5 *infra*.

	train	**dev**	**test**
Nombre d'appels	48	5	7
Durée	9h19	1h01	1h13
Répartition des TP	78,8 %	9,3 %	11,9 %
Moyennes par appel :			
Durée	12m04	12m20	11m29
Nombre de locuteurs	2,1	2,2	3
Nombre de TP	183,4	206,6	189,9
Longueur des TP	17,4 mots	16,1 mots	13,3 mots

TABLEAU 5.5 – Subdivision du corpus VOXDISS en données d'entraînement (*train*), de développement (*dev*) et de test (*test*), où TP = tour de parole.

L'avantage de cette méthode d'échantillonnage, nous l'avons dit, réside dans le fait de ne pas introduire de biais en présupposant que les données sur lesquelles nous souhaitons évaluer SADDE possèdent bien telle ou telle caractéristique (en harmonisant un certain nombre de traits sur les trois échantillons) ; l'inconvénient, comme nous pouvons le voir dans le tableau 5.5 *supra*, réside dans l'absence de maîtrise des caractéristiques inhérentes aux échantillons produits. Ainsi, l'on remarquera que le corpus de test apparaît sensiblement différent des corpus d'entraînement et de développement : il présente notamment, en moyenne, un plus grand nombre de locuteurs par appel, le signe d'une conversation non stan-

dard, éloignée de l'appel archétypal constitué d'un échange bi-directionnel entre un client et un agent.

Concernant la répartition de la parole selon le genre du locuteur, une grande disparité est perceptible dans le corpus d'entraînement : 72,5 % des mots sont prononcés par des locuteurs de genre féminin, *versus* 27,5 % pour les locuteurs de genre masculin. De ce point de vue, les corpus de développement et de test sont plus équilibrés (respectivement 56,4 % et 52,1 % des mots prononcés par des femmes).

Enfin, concernant les profils de locuteurs (Agents *versus* Clients), ils sont assez bien répartis au sein des corpus d'entraînement et de développement, beaucoup moins au sein du corpus de test (60,2 % des mots sont prononcés par des locuteurs appartenant au profil Agents *versus* 39,8 % des mots prononcés par des locuteurs Clients). Toutefois, dans la mesure où nous avons démontré, à travers les analyses de corpus que nous avons menées (*cf.* chapitre 4, et plus précisément section 4.3), que les stratégies discursives de ces deux classes de locuteurs étaient très proches, nous faisons l'hypothèse que cette variation dans les échantillons aura peu d'impact sur les tâches de détection des disfluences.

5.3.2 Caractéristiques lexicales et acoustiques des sous-corpus

Le tableau 5.6 *infra* présente les principales caractéristiques lexicales et acoustiques des trois sous-corpus sélectionnés.

	train	**dev**	**test**
Nombre de mots graphiques	107 788	11 559	13 292
Distribution lexicale	81,3 %	8,7 %	10,0 %
Vocabulaire	2 548	950	935
Durée moyenne mots (s.)	0,220	0,225	0,224
Longueur syll. moy. mots	1,4	1,5	1,5
Nb phonèmes moy. mots	2,8	2,9	2,9
Durée moy. phonèmes (s.)	0,08	0,08	0,08
F_0 moyenne (Hz)	189,5	173,7	178,8
Moyennes par appel :			
Nombre de mots	2 246	2 312	1 899

TABLEAU 5.6 – Caractéristiques lexicales et acoustiques des sous-ensembles d'entraînement (*train*), de développement (*dev*) et de test (*test*). Le vocabulaire correspond au nombre de mots uniques présents dans le corpus.

D'un point de vue acoustique, les corpus d'entraînement, de développement et de test semblent bien équilibrés les uns par rapport aux autres. Le corpus d'entraînement présente une fréquence fondamentale moyenne plus élevée que les

deux autres jeux de données, ce qui est tout à fait normal compte tenu de la prédominance de locutrices dans ce corpus. En effet, les locuteurs de genre féminin prononcent 72,5 % des mots du corpus d'entraînement, *versus* 56,4 % et 52,1 % respectivement dans les corpus de développement et de test.

D'un point de vue lexical, les mots graphiques sont également bien répartis entre les différents échantillons, eu égard à une répartition identique à celle voulue à l'échelle des conversations (81,3 % des mots graphiques dans le corpus d'entraînement, 8,7 % dans le corpus de développement et 10,0 % dans le corpus de test). Les sous-ensembles ne présentent pas non plus d'hétérogénéité dans la répartition des différentes catégories morpho-syntaxiques, à l'exception d'une utilisation légèrement plus faible des pronoms pour le corpus de développement et d'une légère augmentation de l'utilisation de noms et de prépositions dans le corpus de test. La répartition des parties du discours pour chaque sous-corpus est présentée dans le tableau 5.7 *infra*. Le jeu d'étiquette pris en compte est celui des étiquettes simplifiées du `Treetagger`. Enfin, la mise en avant des vingt lemmes les plus fréquents de chaque sous-ensemble (également issus de la lemmatisation opérée par le `Treetagger`) appuie également l'homogénéité de l'échantillonnage ; les deux seuls lemmes à ne pas être présents pour les trois corpus sont « en » qui se trouve absent du corpus de développement, remplacé par « alors » (*cf.* tableau 5.8 *infra*).

POS	**train**	**dev**	**test**
PRO	21,4	19,7	20,5
VER	19,1	18,7	18,7
NOM	12,7	13,2	13,9
ADV	10,6	11,1	10,6
PRP	9,1	9,0	10,0
DET	6,7	7,0	7,1
INT	6,3	6,6	5,9
KON	6,4	5,8	5,7
ADJ	3,0	3,3	3,0
NUM	2,9	3,3	2,4
NAM	1,4	1,7	1,8
ABR	0,4	0,5	0,4
SYM	0,0	0,0	0,0

TABLEAU 5.7 – Répartition (en %) des parties du discours au sein des sous-ensembles d'entraînement (*train*), de développement (*dev*) et de test (*test*).

5.3.3 Distribution des disfluences dans les sous-ensembles

Il est par ailleurs nécessaire d'identifier les caractères concordants et discordants des trois sous-corpus utilisés pour entraîner, développer et évaluer *SADDE* du

train	dev	test
le	le	le
être	être	être
je	avoir	de
avoir	je	je
ce	@card@	<unknown>
de	de	avoir
vous	ce	ce
@card@	<unknown>	vous
que	vous	que
<unknown>	que	@card@
euh	euh	il
il	un	pas
pas	il	euh
un	pas	on
cela	oui	cela
et	cela	un
on	à	oui
à	on	à
oui	et	et
en	alors	en

TABLEAU 5.8 – Les vingt lemmes les plus fréquents des sous-ensembles d'entraînement (*train*), de développement (*dev*) et de test (*test*).

point de vue des phénomènes à détecter, les disfluences d'édition. Le tableau 5.9 *infra* indique les caractéristiques précises de chaque échantillon par rapport à la présence de disfluences d'édition. Les chiffres sont également présentés pour les segments structurant les disfluences et intervenant dans leur ébauche (*reparandum* et *interregnum*) et leur achèvement (*reparans*).

Tout d'abord, la répartition des phénomènes dans les différents corpus n'est pas optimale : en effet, moins de 8 % des disfluences sont présentes dans les données de développement et de test respectivement. Au sein de ces deux corpus, les disfluences se font en moyenne plus rares que dans le corpus d'entraînement. Ce constat est particulièrement frappant pour le corpus de test, qui contient en moyenne seulement 37,1 disfluences par appel, contre 62,5 en moyenne pour le corpus d'entraînement ; ce constat est toutefois à nuancer, dans la mesure où le corpus de test présente à la fois des appels en moyenne moins longs que ceux du corpus d'entraînement (*cf.* tableau 5.5 *supra*) et des tours de parole plus courts (13,3 mots *versus* 17,4 mots en moyenne).

Nous nous attachons ici au traitement automatique de phénomènes suffisamment fréquents dans l'oral spontané pour justifier l'attention qui leur est portée mais suffisamment peu fréquents pour impacter l'efficience de méthodes de détection automatique. Tant en présence qu'en couverture lexicale, nous faisons face à

	train				dev				test			
	rpd	int	rpr	dis	rpd	int	rpr	dis	rpd	int	rpr	dis
Nombre	3 001	309	2 319	3 001	271	29	206	271	260	31	212	260
Rep.	84,9	83,7	84,7	84,9	7,7	7,9	7,6	7,7	7,4	8,4	7,7	7,4
Dens_L	8,0	0,5	3,9	12,4	6,4	0,4	2,9	9,7	5,0	0,4	3,0	8,4
Lg_M	2,9	1,6	1,8	4,5	2,7	1,4	1,6	4,1	2,5	1,9	1,9	4,3
Moy/appel	62,5	6,4	48,3	62,5	54,2	5,8	41,2	54,2	37,1	4,4	30,3	37,1

TABLEAU 5.9 – Répartition des disfluences dans les sous-ensembles d'entraînement (*train*), de développement (*dev*) et de test (*test*), où Moy./appel correspond au nombre moyen de phénomènes par appel (enregistrement) dans le corpus, Lg_M correspond à la longueur moyenne des disfluences exprimée en nombre de mots ; la répartition (Rep.) et la densité lexicale (Dens_L) sont exprimées en pourcentages.

des données particulièrement parcimonieuses pour lesquelles il est de fait difficile de trouver des régularités dans leur schéma d'apparition. En effet, près de 90 % des mots de chacun de nos sous-corpus apparaissent hors d'une disfluence (plus précisément 87,6 % pour le *train*, 90,3 % pour le *dev* et 91,6 % pour le *test*).

5.3.4 Discussion

Nous avons présenté dans cette section les trois corpus utilisés pour l'entraînement, le développement et l'évaluation d'un système de détection automatique des disfluences, SADDE.

Nous avons échantillonné ces données à partir du corpus VOXDISS, manuellement annoté en disfluences, en respectant une homogénéité des conversations du point de vue de leur durée, sans ajouter d'autre biais dans la sélection. Les sous-ensembles ainsi produits se révèlent être en grande partie homogènes, avec une bonne distribution des différentes mesures prises en compte, aux niveaux discursif, lexical et acoustique. Les traits les moins bien répartis sont la prédominance des locuteurs de genre féminin dans le corpus d'apprentissage et la faible densité lexicale des disfluences au sein du corpus de test. Ce dernier point peut avoir un impact sur les performances du système sur ces données. Ce paramètre n'a pas été pris en compte dans la mesure où, par souci de généricité quant à l'application du modèle sur d'autres données, nous ne souhaitions pas présupposer de la quantité de disfluences présentes, mais cela pourrait être intéressant à tester, *via* l'utilisation d'un autre corpus par exemple.

Au-delà de ces analyses contrastives, il est important de signaler que nous disposons de très peu de données d'apprentissage pour l'entraînement de notre système, ce qui est généralement problématique pour l'élaboration de modèles de détection solides. De manière générale, nous manquons de ressources de type

oral conversationnel annotées pour le français, et en particulier pour l'étude des disfluences d'édition, pour laquelle il n'existe pas de corpus annoté de référence comme cela existe pour l'anglais par exemple (*cf.* notre étude de la littérature dans le chapitre 2 ainsi que notre discussion sur les corpus francophones au sein du chapitre 3).

5.4 Développement d'un Système Automatique pour la Détection des Disfluences d'Édition (SADDE)

Cette section expose notre méthode de détection automatique des disfluences à base de champs aléatoires conditionnels (ou « CRF », pour *Conditionnal Random Fields*). Les CRF constituent une approche d'apprentissage supervisé pour l'étiquetage séquentiel (Lafferty *et al.*, 2001). En plus d'être très efficace pour la détection des phénomènes disfluents, comme nous l'avons mis en avant au sein du chapitre 2, cette méthode présente également des performances à l'état de l'art pour de nombreuses tâches d'annotation, comme la transcription graphème-vers-phonème (Hahn *et al.*, 2011b), la reconnaissance d'entités nommées (Sang et De Meulder, 2003) et d'autres tâches en compréhension du langage parlé (Hahn *et al.*, 2011a). De plus, leur capacité à interpréter de larges ensembles de traits potentiellement redondants et à intégrer des dépendances structurelles entre les différentes classes recherchées contribue au choix de cette approche à base de CRF pour nos travaux.

Nous décrivons ici les différentes étapes de conception, choix théoriques et techniques ayant abouti au développement de `SADDE` : nous décrivons en section 5.4.1 nos choix de mise en œuvre et principes techniques ; les sections 5.4.2 et 5.4.3 sont respectivement dédiées à la description des indices utilisés et patrons d'apprentissage développés pour le système ; nous discutons ces choix et principes en section 5.4.4.

5.4.1 Mise en œuvre avec Wapiti

Nous utilisons la mise en œuvre des CRF fournie par `Wapiti` [1] (Lavergne *et al.*, 2010). `Wapiti` est une boîte à outil dédiée à la segmentation et à l'étiquetage de séquences développée au LIMSI, intégrant plusieurs modèles (Maxent, MEMM, CRF). Il met en œuvre plusieurs algorithmes standards et propose plusieurs méthodes de régularisation dans le but d'améliorer la performance de prédiction des modèles (sélection des traits pertinents, réduction du sur-apprentissage...). À des fins de transparence et de réutilisabilité du système développé dans ces travaux,

1. Site Web : http://wapiti.limsi.fr/, consulté le 16/01/15.

nous détaillons ici l'usage que nous avons fait de `4apiti`, soit la description des paramètres utilisés. `Wapiti` peut s'utiliser selon trois modes, pour réaliser :

i) un apprentissage (mode *train*) ;
ii) un étiquetage, en utilisant un modèle généré avec l'apprentissage (mode *label*) ;
iii) une copie d'un modèle précédemment généré sous une forme lisible (mode *dump*).

Le détail des options disponibles et paramètres que nous avons utilisés pour l'apprentissage et l'étiquetage sont décrits dans le tableau 5.10 *infra*.

Paramètres	**Utilisation**
Type de modèle	[CRF]
Algorithme	L-BFGS
Utilisation d'un corpus de développement	Activé
Sauvegarde/restauration de l'état d'optimisation	Désactivé
Nombre de tâches (*thread*)	16
Nombre de séquences à traiter par tâche	[64]
Gestion de la parcimonie	Activé
Nombre maximal d'itérations	Variable
Fenêtre de valeur objective pour le critère d'arrêt	[5]
Fenêtre de valeur du critère d'arrêt pour le développement	0
Intervalle pour le critère d'arrêt	[0,02 %]

(a) mode apprentissage (*train*).

Paramètres	**Utilisation**
Mode maxent pur	Désactivé
Taux d'erreur des étiquettes pour des données déjà étiquetées	Désactivé
Score des étiquettes	Désactivé
Posterior decoding à la place de Viterbi decoding	Désactivé
N meilleures étiquettes	1
Décodage forcé pour les données déjà en partie étiquetées	Désactivé

(b) mode étiquetage (*label*).

TABLEAU 5.10 – Paramétrage de `wapiti` pour l'apprentissage (a) et l'étiquetage (b) de SADDE. Les valeurs présentées entre crochets droits sont celles proposées par défaut.

Nous avons donc entraîné le système de détection des disfluences avec un modèle CRF, pour lequel deux algorithmes ont été testés : `rprop` et `l-bfgs`. Alors que `l-bfgs` est un algorithme très précis mais aussi gourmand en temps de calcul (plusieurs centaines d'itérations sont en général nécessaires avant d'atteindre une convergence acceptable), `rprop` est un algorithme plus rapide et mieux adapté aux contraintes de temps de traitement ; il permet également l'obtention de meilleurs

résultats sur des tâches complexes et structurées (*cf.* par exemple Dinarelli et Rosset (2011)).

Les deux algorithmes aboutissent à des performances proches pour la tâche de détection des disfluences d'édition, mais `l-bfgs` donne les meilleurs résultats sur notre corpus de développement, c'est donc celui que nous avons retenu pour SADDE. Le choix de l'algorithme est également dépendant de la configuration du système et de la tâche envisagée : nous avions en effet sélectionné l'algorithme `rprop` lors de précédentes expériences sur la détection des disfluences d'édition, ce dernier donnant pour celles-ci les meilleurs résultats avec des critères d'arrêt de l'algorithme (soit le nombre maximal d'itérations) déterminés de manière empirique, toujours d'après les résultats obtenus sur le corpus de développement (Dutrey *et al.*, 2014a).

Concernant les autres paramètres dédiés à l'entraînement du système, nous avons notamment optimisé les traitements sur le nombre de tâches (*thread*) allouées à l'exécution de `Wapiti`, paramètre dépendant des possibilités de calcul offertes par le serveur sur lequel est exécutée la tâche, ainsi que la fenêtre de valeur du critère d'arrêt pour le développement, empiriquement fixée à 0. Ce deuxième paramètre est lié à l'utilisation d'un corpus de développement au sein de l'outil. Nous avons également utilisé la fonctionnalité prévue par `Wapiti` pour activer les calculs avant-arrière pour la prise en compte du caractère parcimonieux des données, ce qui est le cas avec notre tâche de détection (*cf.* section 5.3.3).

Les paramètres dédiés à l'étiquetage sont essentiellement utilisés pour l'évaluation du système. Dans la mesure où nous utilisons des outils et mesures d'évaluation externes à `Wapiti`, ces paramètres n'ont pas été activés.

5.4.2 Indices utilisés pour l'entraînement et l'étiquetage

Nous détaillons dans cette section les indices impliqués dans le fonctionnement de SADDE. Nous avons utilisé trois types d'indices, extraits des données de manière automatique : des indices linguistiques, des indices acoustico-prosodiques et des indices discursifs. Dans la mesure où les CRF utilisent exclusivement des attributs discrets, tous les indices à caractère numérique continu ont été discrétisés à l'aide de l'outil `discretize4crf` (Raymond, 2011), qui met en œuvre la méthode MDLPC (« Minimum Description Length Principle ») appliquée à la discrétisation (cette méthode est décrite dans Fayyad et Irani (1993)).

Le tableau 5.11 *infra* résume les indices utilisés en détaillant leur type (linguistique, acoustique ou discursif), leur valeur (attribut discret ou continu) et leur méthode d'acquisition. Ces indices, sur lesquels reposent les patrons d'apprentissage utilisés par `Wapiti` afin de produire un modèle d'annotation séquentielle, sont choisis selon différents critères, combinés ou non. Il peut s'agir d'indices connus dans la littérature pour leur utilité dans des tâches de reconnaissance de disfluences ou des tâches de détection proches (*cf.* chapitre 2) ; il peut également

s'agir d'indices « théoriques », c'est-à-dire de traits caractéristiques des disfluences et de leur réalisation dans le discours (*cf.* chapitre 2) ; il peut enfin s'agir de traits à fort potentiel du fait de leur pertinence statistique évaluée en analyse de corpus (*cf.* chapitre 4). Ces indices ont été *in fine* sélectionnés manuellement selon qu'ils dégradent ou augmentent les performances obtenues par le système sur le corpus de développement (*cf.* section 5.3) pour les tâches de détection considérées (*cf.* section 5.2).

Indice	**Type**	**Valeur**	**Méthode/Outil d'acquisition**
Mots graphiques	Linguistique	Discret	Programme
Lemmes	Linguistique	Discret	`Treetagger`
Parties du discours	Linguistique	Discret	`Brill, MElt, Treetagger`
Genre du locuteur	Discursif	Discret	Programme
Progression du TP	Discursif	Continu	Programme
Progression du dialogue	Discursif	Continu	Programme
Chaîne phonémique	Acoustique	Discret	Alignement forcé
Longueur syllabique	Acoustique	Continu	Alignement forcé
Durée des phonèmes	Acoustique	Continu	Alignement forcé
Fréquence fondamentale	Acoustique	Continu	`Praat`
Taux de voisement	Acoustique	Continu/Discret	`Praat`

TABLEAU 5.11 – Description des indices utilisés pour construire les patrons de SADDE.

Les **indices linguistiques** sur lesquels se basent le système sont d'une part les formes des mots (au sens de leur réalisation graphique dans les transcriptions) et d'autre part les informations issues d'une étape d'étiquetage morpho-syntaxique (génération de lemmes et analyse en parties du discours). Afin d'obtenir ce deuxième type d'indice, nous avons combiné plusieurs outils d'étiquetage morpho-syntaxique : `Brill` (Brill, 1992) dans une version adaptée au français (Allauzen et Bonneau-Maynard, 2008), `MElt` (Denis et Sagot, 2009) et le `Treetagger` (Schmid, 1994). Le jeu d'étiquette de chacun de ces trois étiqueteurs est référencé en annexe A et nous avons comparé leurs résultats sur le corpus VOXDISS. Le résultat de cette comparaison est présenté au sein du chapitre 4, lors de la description de la chaîne d'enrichissement de données issues de l'oral (*cf.* section 4.2.3).

Les trois outils d'étiquetage considérés sont en effet assez disparates en matière de jeu d'étiquettes. Afin de pallier cette disparité et d'exploiter au mieux les sorties de ces outils, nous avons opéré une étape d'éclatement des étiquettes de `Brill` et de `MElt`, produisant huit traits supplémentaires aux étiquettes d'origine (soit neuf traits en tout pour chacun des deux étiqueteurs).

Les **indices acoustico-prosodiques** utilisés pour le système de détection ont été générés à partir de nos données, selon la méthode décrite dans le chapitre 4,

en section 4.2.2. Il s'agit en premier lieu de la transcription phonétique des mots du corpus (ou prononciation), directement issue de l'alignement forcé avec segmentation en phones, et des caractéristiques associées à la chaîne de phonèmes et à leur durée : durée du mot, durée moyenne des phonèmes, voyelles et consonnes composant le mot, nombre de phonèmes composant le mot et longueur syllabique du mot (soit le nombre de voyelles le composant, excepté le schwa final le cas échéant). Une seconde série d'indices concerne les mesures de fréquence fondamentale des phonèmes.

Nous utilisons également des **indices discursifs**, soit des informations relatives au déroulement du dialogue et aux locuteurs. Outre le genre du locuteur (« masculin » *versus* « féminin »), nous avons généré des informations concernant la progression du tour de parole et la progression de la conversation. Le mode de calcul de ces deux indices est décrit en section 4.2.4.

La classe du locuteur, telle que décrite et analysée en section 4.3.4, a été testée comme indice discursif mais pas retenue pour le système. En effet, les résultats obtenus sur le corpus de développement font état d'une dégradation des performances du système dès lors que cet indice est agrégé aux autres. Cela nous semble cohérent avec les spécificités du corpus VOXDISS, où comme nous l'avons discuté en section 4.3.4, les stratégies discursives des deux principales classes de locuteur de notre corpus (Agents et Clients) sont très proches : leur temps de parole est réparti de manière équilibrée, leur mise en œuvre des disfluences est également similaire, *etc.*

5.4.3 Patrons développés pour la construction des modèles de détection

Nous avons développé deux types de patrons pour quatre modèles à base de CRF. Ces patrons sont construits sur les indices linguistiques, acoustico-prosodiques et discursifs présentés en section 5.4.2. Outre la définition d'un modèle de base (*baseline*), nous avons en effet développé trois versions de SADDE : SADDE_L (basé sur des patrons lexicaux), SADDE_A (basé des patrons acoustiques) et SADDE_{LA} (basé sur une combinaison des patrons lexicaux et acoustiques).

Les tableaux 5.12 et 5.13 *infra* présentent pour chaque patron les informations suivantes :

- le trait sur lequel le patron s'appuie ;
- s'il s'applique sur des unigrammes ou des bigrammes ;
- sa fenêtre d'application (l'unité est le mot) ;
- s'il réfère à des unigrammes ou à des bigrammes de mots.

Nous avons opéré une sélection manuelle itérative des traits insérés dans SADDE, dans la mesure où les patrons ont systématiquement été testés de manière itérative sur le système : la configuration des patrons présentés ici (par exemple la dimension de la fenêtre d'application) découle directement des performances ob-

tenues sur le corpus de développement. Si dans toutes les configurations possibles des patrons les performances ont été dégradées sur le corpus de développement, alors les indices utilisés par ces patrons ont été écartés et ne sont pas utilisés dans la construction du modèle de détection.

Les patrons développés pour le **modèle linguistique** SADDE$_L$, considèrent des traits purement linguistiques, comme les parties du discours, ainsi que des indices para-linguistiques (genre du locuteur) ou discursifs (progression de la conversation et des tours de parole). Ces patrons sont décrits dans le tableau 5.12 *infra*.

Type	**N-grammes**	**Fenêtre**	**–**
Mots graphiques	uni-/bigrammes	-2/+2	u
Préfixes position 1	unigrammes	-2/+2	u
Préfixes position 2	unigrammes	0	u
Préfixes position 3	unigrammes	0	u
Suffixes position -1	unigrammes	-1/+1	*
Commence par une capitale ?	unigrammes	-1/+1	*
Contient un chiffre ?	unigrammes	-2/+2	u
POS `Brill` court	uni-/bigrammes	-2/+2	*
POS `Brill` long	uni-/bigrammes	-2/+2	*
POS `MElt` court 1	uni-/bigrammes	-2/+2	*
POS `MElt` court 2	uni-/bigrammes	-2/+2	*
POS `MElt` long	uni-/bigrammes	-2/+2	*
POS `Treetagger` court	uni-/bigrammes	-2/+2	*
POS `Treetagger` lemme	uni-/bigrammes	-2/+2	*
POS `Treetagger` long	uni-/bigrammes	-2/+2	*
Progression du TP	unigrammes	-2/+2	*
Progression du dialogue	unigrammes	-2/+2	u
Genre du locuteur	unigrammes	0	u

TABLEAU 5.12 – Patrons développés pour SADDE$_L$, un modèle CRF basé sur des indices lexicaux, où POS = parties du discours et TP = tour de parole.

Nos tests de sélection de traits pertinents pour la détection des disfluences peut aller à l'encontre de certains indices qui se sont révélés statistiquement pertinents dans l'étude de corpus menée dans le chapitre 4.

La classe du locuteur, notamment, n'a pas été retenue ; son utilisation dégrade les résultats obtenus par SADDE$_L$, alors même que nous avions fait émerger des stratégies discursives différentes entre les locuteurs Agents et Clients eu égard à la production d'énoncés disfluents. Nous avions par exemple montré que la densité lexicale des disfluences était plus importante pour la classe Clients. À l'inverse, la distribution des disfluences eu égard à la progression des tours de parole ne

semblait pas être un indice pertinent, dans la mesure où nous avons montré que les disfluences sont réparties de manière homogène dans la progression des tours de parole au sein desquels elles apparaissent (c'est-à-dire qu'elles n'apparaissent pas particulièrement en début ou en fin de tour de parole) : cet indice s'est pourtant avéré utile pour la détection automatique des disfluences.

Concernant les indices issus de l'étiquetage morpho-syntaxique, nous avons fait l'hypothèse que pour une utilisation au sein d'un outil d'apprentissage automatique l'étiqueteur qui donne les meilleurs résultats n'est pas nécessairement celui qui étiquette les mots « au plus juste ». Nous n'avons pas évalué ici leur qualité par rapport à un étiquetage de référence, comme présenté dans le chapitre 4 (et plus précisément en section 4.2.3) mais par rapport à leur efficacité en tant qu'indice dans le système de détection de disfluences à l'aide de CRF.

Les patrons développés pour le **modèle acoustique** $\texttt{SADDE}_A$ (*cf.* tableau 5.13 *infra*) sont des traits classiquement employés pour caractériser la parole d'un point de vue acoustique. Contrairement à notre hypothèse de départ, l'utilisation de la durée des pauses silencieuses avant et après les mots disfluents n'est pas un indice utile à la détection des disfluences.

Type	**N-grammes**	**Fenêtre**	**–**
Prononciation	uni-/bigrammes	-2/+2	*
Durée	unigrammes	-2/+2	u
Durée moyenne des phonèmes	uni-/bigrammes	-2/+2	u
Durée moyenne des voyelles	uni-/bigrammes	-2/+2	u
Durée moyenne des consonnes	uni-/bigrammes	-2/+2	u
Longueur syllabique	unigrammes	-2/+2	*
Nombre de phonèmes	unigrammes	-2/+2	*
Moyenne F0 (segments centraux)	uni-/bigrammes	-2/+2	*
Nb de voyelles taux de voisement < 70 %	uni-/bigrammes	-2/+2	*
Au moins une voyelle avec un taux de voisement < 70 % ?	uni-/bigrammes	-2/+2	*
Taux de voisement moyen < 70 % ?	uni-/bigrammes	-2/+2	*
$\Delta F0_{max-min}$ voyelles	unigrammes	-2/+2	u
$\Delta F0_{fin-deb}$ voyelles	unigrammes	-2/+2	u

TABLEAU 5.13 – Patrons développés pour le modèle basé sur des indices acoustiques, $\texttt{SADDE}_A$.

Enfin, nous avons généré un **modèle lexico-acoustique,** $\texttt{SADDE}_{LA}$, est construit en utilisant les indices acoustiques et linguistiques développés pour les deux modèles $\texttt{SADDE}_L$ et $\texttt{SADDE}_A$, dans leur configuration optimisée pour chacun de ces deux modèles séparément.

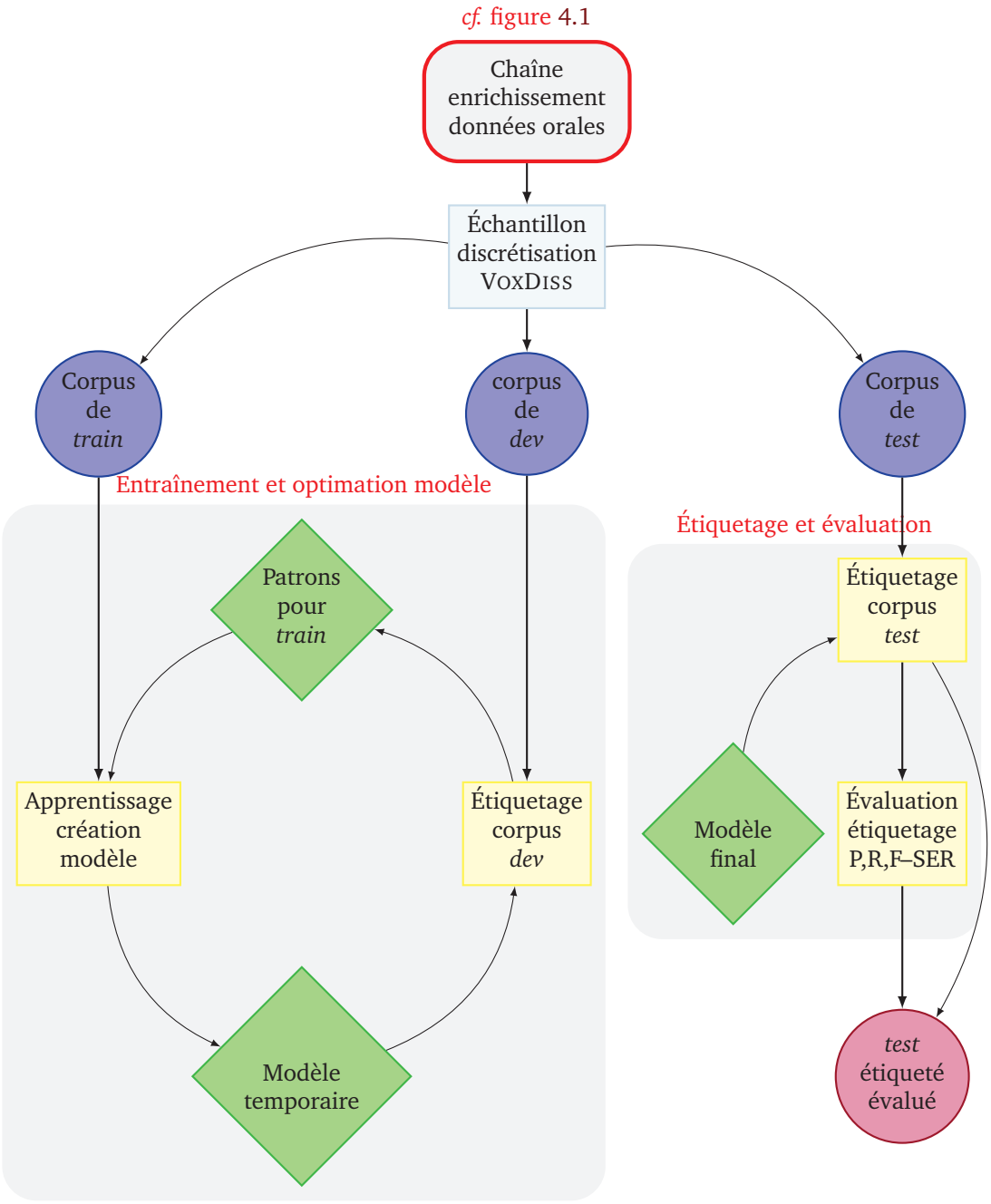


FIGURE 5.1 – Chaîne de développement, de fonctionnement et d'évaluation du système de détection des disfluences (SADDE). Le module d'entraînement et d'optimisation du modèle est utilisé à deux reprises de manière distincte, *(i)* pour mettre au point le modèle CRF fondé sur des indices linguistiques et *(ii)* pour mettre au point le modèle CRF fondé sur des indices acoustiques.

5.4.4 Discussion

Nous avons présenté dans cette section le développement d'un système de détection des disfluences, SADDE. Nous avons décliné ce système en trois modèles, afin d'évaluer sur nos tâches de détection les performances d'un système fondé sur des indices linguistiques et d'un système fondé sur des indices acoustiques ; le troisième modèle est généré en fusionnant les patrons développés pour les deux premiers modèles. Pour construire ces modèles, nous avons testé une série d'indices à partir d'un état de l'art sur l'analyse et la détection automatique de disfluences, de même qu'en nous basant sur nos propres observations sur la réalisation des disfluences dans un corpus d'oral conversationnel en français. Nous avons opéré une sélection manuelle des indices en optimisant la configuration des patrons associés sur un corpus de développement. Nous avons également testé deux algorithmes pour les CRF, et optimisé les résultats obtenus par l'outil utilisé pour implémenter les CRF, Wapiti. Le schéma présenté en figure 5.1 *supra* résume la chaîne de conception, de fonctionnement et d'évaluation de SADDE.

5.5 Évaluation de SADDE sur les tâches de détection des disfluences

Nous évaluons SADDE sur l'ensemble des tâches de détection que nous avons conçues pour l'identification des disfluences d'édition en contexte conversationnel. La section 5.5.1 est dédiée à l'exposition des principes méthodologiques régissant l'évaluation de notre système ; les sections suivantes sont dédiées à la présentation des résultats et aux analyses liées, suivant les principes énoncés ici.

En premier lieu, nous discutons les résultats en fonction du segment que nous cherchons à détecter (totalité de la disfluence ? seulement son ébauche ? un segment hybride ?) : pour chacun de ces ensembles, nous nous attachons à identifier le meilleur modèle CRF compte tenu de l'objectif lié à chacune des tâches (analyse des disfluences ? nettoyage des données conversationnelles ? les deux ?) ; nous discutons les résultats en fonction des mesures à privilégier selon ces objectifs.

Nous nous attachons en second lieu à comparer les différentes stratégies découlant des tâches envisagées pour chaque segment à identifier : celles-ci correspondent en effet à différents niveaux de structuration (une à trois séquences détectées pour un segment donné).

Enfin, nous menons deux analyses transverses : la première, qui couvre les sections 5.5.2 et 5.5.3, concerne la pertinence de mener une détection distinguant les différentes classes de disfluences (répétitions, auto-corrections, faux-départs, disfluences combinées) ; la seconde, interne à la section 5.5.2, concerne l'impact de la détection de l'achèvement sur la qualité de la détection de l'ébauche.

Nous présentons par ailleurs pour chaque tâche les résultats obtenus en précision, rappel, f-mesure et Slot Error Rate. Les mesures de précision, rappel et f-mesure sont systématiquement présentées pour l'ensemble des étiquettes impliquées dans la tâche de détection mais aussi pour chaque étiquette évaluée individuellement. Lorsque cela est pertinent, nous présentons également les résultats détaillés du Slot Error Rate, c'est-à-dire les détails des performances du système sur les erreurs d'insertion, de suppression et de substitution (toujours pour l'ensemble des étiquettes impliquées dans chaque tâche).

5.5.1 Principes méthodologiques d'évaluation

Les principes méthodologiques régissant l'évaluation du système sur l'ensemble des tâches de détection des disfluences traitées portent sur deux aspects : *(i)* l'évaluation des performances de chaque modèle CRF créé par rapport à un modèle de base (*cf.* section 5.5.1.1) et *(ii)* le choix des mesures d'évaluation choisies pour déterminer les performances de SADDE de la manière la plus précise possible, tout en prenant en compte les objectifs liés à chaque tâche (*cf.* section 5.5.1.2).

5.5.1.1 Modèle minimal de SADDE (*baseline*)

Un modèle minimal a été développé sur l'exemple de celui décrit dans Qian et Liu (2013). Cette *baseline* utilise des indices lexicaux uniquement. Plus précisément, les patrons développés sont simplement fondés sur la forme fléchie des mots et sont appliqués sur une fenêtre de -2/+2 mots, comme formalisé dans le tableau 5.14 *infra*. Utiliser uniquement les informations lexicales au niveau local permet de mieux mesurer l'apport des traits et des patrons lexico-acoustiques utilisés pour enrichir le système.

Type	**N-grammes**	**Fenêtre**	**–**
Forme des mots	uni-/bigrammes	-2/+2	u

TABLEAU 5.14 – Patrons développés pour le modèle minimal de SADDE (baseline).

5.5.1.2 Métriques d'évaluation

SADDE est évalué à l'aide de différentes métriques d'évaluation, qui donnent des informations différentes et complémentaires quant à l'efficacité du système. Pour la description des mesures de performance, nous utilisons les notations suivantes :

- *vp*+*fn* représente la référence ;

- *vp* (vrai positif) représente les étiquettes trouvées par le système et présentes dans la référence (considérées comme correcte) ;
- *fp* (faux positif) les étiquettes trouvées par le système mais absentes de la référence (erreurs d'insertion) ;
- *vn* (vrai négatif) les étiquettes non trouvées par le système et absentes de la référence (correctes) ;
- *fn* (faux négatif) les étiquettes non trouvées par le système mais présentes dans la référence (erreurs de suppression).

Précision, Rappel et F-mesure Ce triplet de mesures binaires, décrit dans la formule *infra*, constitue les mesures de performance les plus conventionnelles en extraction d'information. Elles considèrent comme faux – de manière indifférenciée – les trois types d'erreurs suivants : les substitutions, les suppressions et les insertions. La f-mesure est une moyenne harmonique pondérée des mesures distinctes que sont rappel et précision.

$$\text{P} = \frac{vp}{vp+fp} \qquad \text{R} = \frac{vp}{vp+fn} \qquad \text{F-Mesure} = \frac{2\times(P\times R)}{P+R}$$

Slot Error Rate Cette métrique (Makhoul *et al.*, 1999), contrairement au rappel et à la précision, consiste en la mesure d'un taux d'erreurs : proche du taux d'erreurs de mots ou WER (pour *Word Error Rate*), utilisé en reconnaissance automatique de la parole, le SER comptabilise en effet les erreurs selon leur catégorie (insertions, suppressions, substitutions). Dans le cadre de notre tâche de détection des disfluences d'édition, la substitution notamment peut porter sur le type (identification de l'étiquette), la frontière ou les deux. L'utilisation du SER, tel que présenté dans la formule *infra*, nous permet de pondérer le poids de ce type d'erreurs par rapport aux erreurs d'insertion et de suppression.

$$\text{Slot Error Rate} = \frac{fn+fp+tf+0.5\times(t+f)}{vp+fn}$$

L'utilisation de l'ensemble de ces métriques permet d'évaluer le système de détection de disfluences. Le SER consistant en une mesure plus fine, venant compléter les classiques précision, rappel et f-mesure, nous le privilégions pour une analyse fine des résultats obtenus par SADDE.

5.5.2 Performances sur les tâches de détection sans distinction des classes de disfluences

Cette section, conformément aux principes énoncés en introduction de chapitre, est dédiée à l'évaluation de SADDE sur les tâches de détection des disfluences sans identification des quatre classes de disfluences d'édition. Nous analysons les résultats obtenus par le système en fonction du segment disfluent concerné par

la détection automatique. Pour chacun de ces ensembles, nous nous attachons à analyser ces résultats eu égard aux enjeux de détection liés à chaque tâche. Nous comparons également les différentes stratégies de structuration et l'apport de la détection des *reparans* pour la détection des ébauches.

Pour chacune des tâches de détection, regroupées en trois ensembles cohérents au sein des sections 5.5.2.1, 5.5.2.2 et 5.5.2.3, les tableaux de résultats présentent les informations suivantes :

- les performances obtenues par le système en précision, rappel, f-mesure et SER pour l'*ensemble* des étiquettes impliquées dans la tâche. Les résultats globaux sont indiqués en noir dans les tableaux de résultats ;
- le détail des performances obtenues par le système en précision, rappel et f-mesure pour *chacune* des étiquettes impliquées dans la tâche, dans le cas où cette dernière comprend au moins deux étiquettes. Dans ce cas, les étiquettes correspondantes sont indiquées en gris, de même que les résultats associés.

Les meilleures performances et les performances les plus basses obtenues pour chaque mesure (globale et par étiquette) sont respectivement présentées en vert et en rouge dans les tableaux de résultats.

5.5.2.1 Détection de la totalité des segments disfluents

Le tableau 5.15 *infra* présente l'évaluation de SADDE pour les trois tâches ayant pour but de détecter la totalité des segments disfluents, sans distinguer les différentes classes d'événements disfluents. Ces tâches divergent de par leur stratégie de structuration des disfluences : en effet, la tâche 1.A détecte la totalité de la disfluence sans différencier les segments qui la compose ; la tâche 1.B propose une structuration intermédiaire, en distinguant deux séquences, son ébauche (constituée du *reparandum* et de l'*interregnum*) et son achèvement (constitué du *reparans*) ; la tâche 1.C propose une structuration complète : les trois séquences composant la disfluences sont différenciées (*reparandum*, *interregnum*, *reparans*).

Tout d'abord, l'évaluation des différents modèles de détection montre que la *baseline* obtient la meilleure précision pour l'ensemble des tâches (précision globale = 53,4 % pour la tâche 1.A, 64,7 % pour la tâche 1.B et 64.8 % pour la tâche 1.C, soit respectivement des gains de 7,5 points, 5,1 points et 4,3 points par rapport au deuxième meilleur modèle en précision) et pour la totalité des séquences de détection (à l'exception de l'identification de l'*interregnum* dans la tâche 1.C, pour laquelle SADDE_L obtient 15 points de plus, passant d'une précision de 50 % à une précision de 75 %).

Concernant la tâche sans aucune structuration des disfluences (tâche 1.A), le meilleur modèle (c'est-à-dire le modèle obtenant le meilleur équilibre entre les différentes mesures d'évaluation) est le modèle combinant des indices linguistiques et acoustiques, SADDE_{LA}.

	BSL	SADDE$_L$	SADDE$_A$	SADDE$_{LA}$	Étiquettes
P	53.4	45.9	49.5	45.9	
R	21.1	27.6	21.1	29.9	rpd+int+rpr
F	30.2	34.4	29.6	36.2	
SER	77.8	69.9	78.0	69.7	

(a) Tâche 1.A – objectif (A) – séquence unique.

	BSL	SADDE$_L$	SADDE$_A$	SADDE$_{LA}$	Étiquettes
P	64.7	59.6	57.9	54.1	
	61.6	59.4	55.5	52.5	rpd+int
	68.1	59.7	60.6	56.0	rpr
R	26.1	39.6	25.6	37.5	
	23.5	38.8	23.5	35.8	rpd+int
	29.2	40.6	28.3	39.6	rpr
F	37.2	47.6	35.5	44.3	
	34.0	47.0	33.0	42.6	rpd+int
	40.9	48.3	38.6	46.4	rpr
SER	77.0	69.0	80.9	73.4	

(b) Tâche 1.B – objectif (AN) – deux séquences.

	BSL	SADDE$_L$	SADDE$_A$	SADDE$_{LA}$	Étiquettes
P	64.8	60.5	57.4	56.2	
	61.2	58.3	55.0	55.3	rpd
	50.0	75.0	60.0	60.0	int
	70.0	62.0	60.0	57.0	rpr
R	25.6	39.6	24.7	34.2	
	24.2	39.2	23.5	34.2	rpd
	9.7	29.0	9.7	19.4	int
	29.7	41.5	28.3	36.3	rpr
F	36.8	47.8	34.5	42.5	
	34.7	46.9	32.9	42.3	rpd
	16.2	41.9	16.7	29.3	int
	41.7	49.7	38.5	44.4	rpr
SER	78.7	69.3	82.3	74.6	

(c) Tâche 1.C – objectif (AN) – trois séquences.

TABLEAU 5.15 – Performances obtenues par la *baseline* et les trois versions de SADDE dans la tâche de détection des disfluences sans distinction de classes, pour les stratégies d'identification sans structuration (a), avec structuration intermédiaire (b) et avec structuration complète (c).

Les deux autres tâches, qui cherchent à structurer la disfluence au moment de la détection (tâches 1.B et 1.C), présentent quant à elles des tendances similaires. Pour ces deux tâches, le modèle qui obtient les meilleures performances est celui fondé sur des indices linguistiques uniquement ($\texttt{SADDE}_L$). Le modèle à base d'indices acoustiques ($\texttt{SADDE}_A$) présente un rappel particulièrement faible (25,6 % pour la tâche 1.B et 24,7 % pour la tâche 1.C sur l'ensemble des étiquettes), soit 10 à 15 points de moins que le meilleur rappel, obtenu dans les deux cas par $\texttt{SADDE}_L$. Il est par ailleurs particulièrement étonnant de voir que $\texttt{SADDE}_{LA}$ obtient une précision à la fois plus faible que $\texttt{SADDE}_L$ et que $\texttt{SADDE}_A$. Contrairement à notre hypothèse, dans ce cas la combinaison des patrons optimisés pour les deux modèles « purs » ne permet pas une amélioration des performances lorsqu'ils sont combinés. Enfin, si l'on analyse le détail des performances par étiquette (et donc par segment identifié), nous pouvons remarquer que les *reparans* sont plus faciles à détecter que les séquences de *reparandum* et d'*interregnum* – que ces deux dernières séquences soient confondues ou non. Enfin, la détection de l'*interregnum* seul obtient une très bonne précision (75 % avec $\texttt{SADDE}_L$) ; en contrepartie, elle obtient globalement un rappel très faible (29 % au mieux), ce qui est compréhensible étant donné sa très faible fréquence d'apparition dans les données.

Nous avons évalué le meilleur système pour détecter la totalité d'une disfluence : les meilleurs résultats sont obtenus par le modèle fondé sur des indices linguistiques uniquement ($\texttt{SADDE}_L$). Si l'on s'attache à identifier la meilleure stratégie pour la détection de ce segment, entre les différentes stratégies de structuration évaluées, nous pouvons affirmer que plus on structure le segment, meilleurs sont les résultats. En effet, si l'on considère le meilleur modèle pour chacune des trois tâches, la f-mesure augmente significativement entre la tâche de détection sans structuration (tâche 1.A, avec une f-mesure de 36,2 %) et la tâche de détection avec structuration complète (tâche 1.C, avec une f-mesure de 47,8 %, soit un gain de 11,6 points).

Ces résultats sont cependant à mettre en perspective avec les objectifs finaux de la détection. Si l'on souhaite détecter les disfluences dans un objectif d'analyse (objectif **A**), alors la meilleure méthode est d'utiliser notre *baseline* en distinguant le *reparandum* du *reparans*. En effet, en particulier si l'on s'attache à l'étude de la manière dont les locuteurs modifient leur énoncé en contexte conversationnel – que ce soit pour préciser leur pensée, corriger une faute sémantique, *etc.* – on souhaitera construire des couples de mots ou segments textuels liés entre eux. À ces fins, la tâche 1.B est la plus adaptée et donne les meilleurs résultats (61,0 % de précision pour la détection du *reparandum* et 68,1 % pour celle du *reparans*). En revanche, si l'objectif de la détection est de nettoyer les données en supprimant l'ébauche de la disfluence (objectif **N**), c'est-à-dire à la fois le *reparandum* et l'*interregnum*, alors nous souhaitons obtenir le meilleur équilibre possible entre les différentes mesures d'évaluation. Dans ce cas la tâche 1.C est à privilégier, avec le modèle $\texttt{SADDE}_L$, cette dernière donnant dans l'ensemble les meilleurs résultats

(précision globale à 60,5 %, rappel global à 39,6 %, f-mesure globale à 47,8 % et SER à 69,3 %).

5.5.2.2 Détection des ébauches

Les tâches 2.B et 2.C ont été conçues pour détecter l'ébauche des disfluences d'édition et participent de ce fait d'un objectif de nettoyage exclusivement (objectif **N**). Elles confrontent deux stratégies d'annotation (en miroir aux tâches 1.B et 1.C) : alors que dans la tâche 2.B les éléments composant l'ébauche sont confondus (*reparandum* et *interregnum* considérés comme une séquence unique) ils sont différenciés dans la tâche 2.C (détection de deux séquences). Les résultats obtenus par les différentes versions de SADDE sont présentés dans le tableau 5.16 *infra*.

	BSL	SADDE$_L$	SADDE$_A$	SADDE$_{LA}$	Étiquettes
P	61.1	52.0	50.4	51.3	
R	26.5	39.2	25.4	38.1	rpd+int
F	37.0	44.7	33.8	43.7	
SER	73.7	67.5	76.0	68.1	

(a) Tâche 2.B – objectif (N) – séquence unique.

	BSL	SADDE$_L$	SADDE$_A$	SADDE$_{LA}$	Étiquettes
P	55.9	56.0	51.8	52.2	
	55.9	55.0	53.1	51.3	rpd
	60.9	72.7	28.6	70.0	int
R	23.7	38.5	24.4	37.5	
	25.4	40.0	26.5	39.2	rpd
	9.7	25.8	6.5	22.6	int
F	33.3	45.6	33.2	43.6	
	34.9	46.3	35.4	44.4	rpd
	16.7	38.1	10.5	34.1	int
SER	78.9	70.1	80.2	71.1	

(b) Tâche 2.C – objectif (N) – deux séquences.

TABLEAU 5.16 – Performances obtenues par la *baseline* et les trois versions de SADDE dans la tâche de détection des ébauches de disfluences sans distinction de classes, pour les stratégies d'identification sans structuration (a) et avec structuration (b).

Tout d'abord, si l'on cherche à identifier le meilleur modèle pour chacune des deux tâches, les tendances sont globalement les mêmes : le modèle fondé sur des indices linguistiques exclusivement (SADDE$_L$) obtient les meilleures performances

de manière globale comme par étiquette (c'est-à-dire par séquence d'identification).

Contrairement aux tâches de détection portant sur la totalité de la disfluence (tâches 1.A, 1.B et 1.C) l'unique objectif poursuivi ici est le nettoyage des données orales, dans la mesure où le *reparans* n'est pas détecté. Nous sélectionnons donc la meilleure stratégie compte tenu cet objectif. Ainsi, malgré une précision et une f-mesure globales moins bonnes (-4 points de précision et -0,9 points de f-mesure), la stratégie de confusion du *reparandum* et de l'*interregnum* (tâche 2.B) est la meilleure stratégie pour détecter l'ébauche des disfluences d'édition. En effet, si l'on analyse en détail les performances obtenues par $\texttt{SADDE}_L$ sur les deux tâches (*cf.* tableau 5.17), la tâche 2.B obtient un meilleur SER dans la mesure où elle fait beaucoup moins d'erreurs de suppression : trente séquences supplémentaires détectées par rapport à la tâche 2.C, pour seulement trois séquences détectées à tort (erreurs d'insertion). Nous voulons en effet éviter à la fois la suppression de segments non disfluents (liée aux erreurs d'insertion) et la conservation d'ébauches (liée aux erreurs de suppression). Pour la tâche 2.B, $\texttt{SADDE}_L$ produit en revanche davantage d'erreurs de substitution (correspondant uniquement à des erreurs de frontière dans la mesure où une seule étiquette est considérée), type d'erreur auquel nous avons accordé moins de poids.

Tâches	**Corrects**	**Insertions**	**Suppressions**	**Substitutions**
2.B	39,2	16,5	41,2	19,6
2.C	38,5	15,8	47,1	14,4

TABLEAU 5.17 – Détails du SER (en %) obtenu par $\texttt{SADDE}_L$ sur les deux tâches de détection de l'ébauche des disfluences.

En effet, pour les erreurs de substitution, même si la totalité de l'ébauche n'est pas identifiée, l'identification d'un segment d'ébauche de disfluence améliore de toute manière la lisibilité des données et leur fluidité syntaxique, comme l'illustre la série d'exemples 5.1 *infra*.

Nous présentons pour chacun de ces trois exemples la référence et l'hypothèse, c'est-à-dire la séquence effectivement détectée par $\texttt{SADDE}_L$ pour la tâche 2.B. Dans l'exemple (a) la totalité du premier tour de parole est annotée dans la référence ; SADDE conserve le segment « c'est ma compagne », ce qui nous semble être une meilleure alternative : l'information principale est conservée, sans rupture dans l'axe syntagmatique. Dans les exemples (b) et (c), SADDE réduit considérablement le dérèglement syntaxique des énoncés, ne laissant que des répétitions simples à identifier dans une deuxième passe par exemple (« vous vous » et « oui oui »). Nous pouvons même nous interroger sur le statut disfluent de cette répétition « oui oui » qui nous semble d'avantage lier le discours que rendre la parole disfluente.

EXEMPLE SONORE 5.1 – Erreurs de substitution produites par SADDE$_L$ sur la tâche 2.B.

(a) REF : — < c'est ma compagne qui a c'est euh au nom de > — dites-moi
HYP : — c'est ma compagne < qui a c'est euh au nom de > — dites-moi

(b) REF : au niveau du raccordement de chantier < vous co- > vous savez
HYP : au niveau du raccordement de chantier vous < co- > vous savez

(c) REF : < oui oui oui oui non mais je me > ça m'a servi de base
HYP : oui oui < oui oui non mais je me > ça m'a servi de base

5.5.2.3 Détection hybride

La tâche de détection 3.C est dédiée à l'analyse des disfluences (objectif **A**), *via* la détection d'un segment hybride composé de deux séquences liées : le *reparandum* et le *reparans*. L'*interregnum* est donc exclu de la détection ; cette exclusion est motivée par la faible quantité d'*interregnum* lexicalement pleins (hors pauses silencieuses donc) et la prédominance d'*interregnum* essentiellement constitués d'hésitations vocaliques (« euh ») et de marqueurs discursifs (« enfin », « alors », *etc.*), relativement aisés à détecter par ailleurs. Les performances obtenues par les différentes versions du système sont décrites dans le tableau 5.18 *infra*.

	BSL	SADDE$_L$	SADDE$_A$	SADDE$_{LA}$	Étiquettes
P	69.7	57.5	58.9	56.5	
	62.9	56.2	56.0	54.4	rpd
	78.2	59.3	62.4	59.2	rpr
R	25.8	39.6	25.2	38.8	
	23.5	40.0	23.5	38.1	rpd
	28.8	39.2	27.4	39.6	rpr
F	37.7	46.9	35.3	46.0	
	34.2	46.7	33.1	44.8	rpd
	42.1	47.2	38.0	47.5	rpr
SER	75.5	70.9	79.7	71.7	

TABLEAU 5.18 – Performances obtenues par la *baseline* et les trois versions de SADDE dans la tâche de détection hybride des disfluences sans distinction de classes (Tâche 3.C – objectif (A) – deux séquences).

L'objectif de la tâche 3.C est donc d'obtenir des couples de segments liés *reparandum* et *reparans*. Les résultats obtenus par les différents modèles sont assez

contrastés : d'une part la meilleure précision est obtenue avec la *baseline*, d'autre part les meilleurs rappel et f-mesure sont obtenus avec le modèle à base d'indice linguistique uniquement (SADDE$_L$) ; les performances les moins bonnes étant obtenues par le modèle construit sur des indices acoustiques exclusivement.

Eu égard à l'objectif d'analyse, nous souhaitons soit privilégier la précision afin d'obtenir les couples de segments les plus justes possibles (dans le cas où nous souhaitons constituer automatiquement une ressource de disfluences en contexte conversationnel par exemple) soit privilégier un bon équilibre avec un rappel honorable et faisant fi d'erreurs de substitution (dans le cas où nous souhaitons constituer la ressource de manière semi-automatique, avec correction manuelle des exemples en aval de leur extraction avec SADDE) ; dans ce deuxième cas de figure, l'utilisation de SADDE$_L$ est à privilégier, ce modèle obtenant la meilleure f-mesure et le SER le plus bas pour cette tâche.

5.5.2.4 Analyses transversales : impact de la détection de l'achèvement sur l'identification de l'ébauche

Une analyse transversale des tâches de détection, dont les résultats sont présentés dans les sections 5.5.2.1 et 5.5.2.2, permet d'étudier l'impact de la détection de l'achèvement sur l'identification de l'ébauche en répondant aux questions suivantes :

- l'identification du *reparans* soutient-elle l'identification de l'ébauche d'une disfluence ?
- l'identification du *reparans* soutient-elle l'identification du *reparandum* et de l'*interregnum* ?

Ainsi, nous comparons les expériences 1.B et 2.B (avec le meilleur modèle, SADDE$_L$) en vue de répondre à la question suivante : l'identification du *reparans* soutient-elle l'identification de l'ébauche d'une disfluence ? Malgré une légère baisse du rappel (-0,4 points), les performances du système concernant la détection de l'ébauche (identification d'une seule séquence, *reparandum* et *interregnum* confondus) sont sensiblement améliorées lorsque le système y associe une détection conjointe du *reparans* : la précision passe de 52,0 % à 59,4 %, pour un gain de 2,3 points en f-mesure.

De la même manière que pour l'analyse précédente, une confrontation des résultats obtenus pour les expériences 1.C et 2.C permet de répondre à la question suivante : l'identification du *reparans* soutient-elle l'identification du *reparandum* et de l'*interregnum* ? Contrairement à la stratégie évaluée dans le paragraphe précédent, ici l'adjonction d'une détection du *reparans* dégrade les résultats obtenus pour le *reparandum* et l'*interregnum* (envisagés comme deux séquences distinctes). Excepté une légère augmentation du rappel pour la détection du *reparandum*, toutes les autres mesures sont dégradées, spécialement pour la détection de l'*interregnum*.

Il n'y a donc pas de réponse absolue : si l'on cherche à détecter l'ébauche des disfluences sans structurer les éléments qui la compose, alors la meilleure stratégie est d'y adjoindre une détection du *reparans*, qui vient améliorer les performances de détection de cette séquence. Si l'on cherche à isoler l'*interregnum* du *reparans* à des fin de nettoyage des données disfluentes, objectif qui ne nécessite pas d'identifier l'achèvement des disfluences, alors mieux vaut ne pas adjoindre à la tâche une détection du *reparans*, qui dans ce cas n'aide pas à détecter les deux premières séquences.

5.5.3 Performances sur les tâches de détection avec identification des classes de disfluences

En complément des trois séries de tâches de détection (présentées en section 5.5.2), nous avons également mené une série d'expériences visant à étudier la faisabilité d'une détection des disfluences considérant les quatre classes d'événements inclus dans les disfluences d'édition (répétitions, auto-corrections, faux-départs, disfluences combinées). Nous avons évalué SADDE en suivant la même méthode que pour nos précédentes expériences, et les tâches ci-dessous, pour lesquelles les résultats sont présentés suivant le même regroupement par type de segment concerné, sont le miroir des détections sans distinction des classes de disfluences.

Distinguer les classes de disfluences implique une augmentation significative du nombre d'étiquettes incluses dans le système de détection (par exemple, entre la tâche 1.A et sa tâche miroir 1.X, on passe de une à quatre étiquettes ; entre la tâche 1.C et sa tâche miroir 1.Z, on passe de trois à dix étiquettes). Ainsi, bien qu'elle permette d'analyser le comportement d'un modèle de détection cherchant à identifier de manière distincte des classes de disfluences au comportement parfois différent les unes des autres, ces résultats sont à relativiser : en effet, augmenter le nombre d'étiquettes est synonyme d'une augmentation de la parcimonie de chacune d'entre elles, ce qui peut entraîner des effets de bords sur les performances obtenues par le système. C'est notamment le cas concernant les faux-départs et les disfluences combinées, les deux classes les moins fréquentes.

5.5.3.1 Détection de la totalité des segments disfluents par classes de disfluences

Les tableaux 5.19, 5.20 et 5.21 *infra* présentent l'ensemble des résultats pour les tâches de détection de la totalité du segment de disfluences d'édition, avec distinction des différentes classes. Ces tâches sont les tâches miroir de celles évaluées en section 5.5.2.1 : elles présentent en effet les mêmes stratégies de structuration du segment à identifier, distinction des classes en sus. Ainsi, le tableau 5.19 présente les performances obtenues pour une stratégie d'identification des disfluences

sans structuration ; le tableau 5.20 présente les performances obtenues avec structuration intermédiaire ; le tableau 5.21 présente les performances obtenues avec structuration complète.

	BSL	SADDE$_L$	SADDE$_A$	SADDE$_{LA}$	Étiquettes
P	56.0	41.7	45.5	41.9	
	72.7	69.5	63.6	62.5	dis.rep
	44.4	33.3	37.5	33.3	dis.rev
	0.0	16.7	0.0	7.7	dis.rest
	18.8	11.8	18.2	23.1	dis.div
R	18.0	23.0	17.2	23.8	
	42.6	43.6	37.2	42.6	dis.rep
	4.7	15.3	7.1	17.6	dis.rev
	0.0	5.6	0.0	2.8	dis.rest
	6.5	8.7	8.7	13.0	dis.div
F	27.2	29.6	25.0	30.0	
	53.7	53.6	47.0	50.6	dis.rep
	8.5	21.0	11.9	23.1	dis.rev
	0.0	8.3	0.0	4.1	dis.rest
	9.7	10.0	11.8	16.7	dis.div
SER	86.8	82.4	91.4	84.7	

TABLEAU 5.19 – Performances obtenues par la *baseline* et les trois versions de SADDE dans la tâche de détection des classes de disfluences, pour la stratégie d'identification sans structuration (Tâche 1.X – objectif (A) – séquence unique).

Tout d'abord, il est intéressant de voir que pour la sélection du meilleur modèle CRF les tendances sont identiques à celles des tâches de la section 5.5.2.1 : la *baseline* obtient globalement la meilleure précision (excepté pour la tâche 1.Z, pour laquelle l'absence de résultats de la *baseline* est discutée dans le paragraphe suivant) ; concernant la tâche 1.X, SADDE$_L$ obtient les meilleurs résultats toutes mesures équilibrées, à l'image de la tâche 1.A ; SADDE$_L$ obtient les meilleurs résultats pour les tâches 1.Y et 1.Z, à l'image des tâches 1.B et 1.C. De ce point de vue, la distinction des classes d'événements ne modifie donc pas la tendance des performances obtenues par les différents modèles évalués.

Par ailleurs, il est très étonnant d'observer que la *baseline* ne parvient à détecter aucune des disfluences d'édition annotées dans la référence, ni ne parvient à détecter d'autres séquences qui pourraient être absentes de la référence. Bien qu'il s'agisse de la tâche la plus complexe parmi celles évaluées dans nos travaux et donc celle où chaque séquence à détecter comporte le moins d'occurrences dans nos données (autrement dit c'est pour cette tâche que la parcimonie des données, fonction des étiquettes impliquées dans la tâche, est la plus forte), nous devrions

	BSL	SADDE$_L$	SADDE$_A$	SADDE$_{LA}$	Étiquettes
	57.8	52.6	54.4	48.3	
	68.2	63.5	64.6	63.4	rpd+int.rep
	50.0	52.4	46.2	49.0	rpd+int.rev
	0.0	7.7	0.0	0.0	rpd+int.rest
P	27.3	36.0	21.1	21.7	rpd+int.div
	68.2	62.2	64.6	63.4	rpr.rep
	41.7	42.9	46.2	32.7	rpr.rev
	33.3	46.7	41.7	28.6	rpr.div
	22.7	31.8	22.2	29.4	
	48.4	50.5	45.2	48.4	rpd+int.rep
	7.1	25.9	7.1	28.2	rpd+int.rev
	0.0	2.8	0.0	0.0	rpd+int.rest
R	6.5	19.6	8.7	10.9	rpd+int.div
	48.4	49.5	45.2	48.4	rpr.rep
	5.9	21.2	7.1	18.8	rpr.rev
	8.8	20.6	14.7	11.8	rpr.div
	32.6	39.6	31.6	36.6	
	56.6	56.3	53.2	54.9	rpd+int.rep
	12.4	34.6	12.2	35.8	rpd+int.rev
	0.0	4.1	0.0	0.0	rpd+int.rest
F	10.5	25.4	12.3	14.5	rpd+int.div
	56.6	55.1	53.2	54.9	rpr.rep
	10.3	28.3	12.2	23.9	rpr.rev
	14.0	28.6	21.7	16.7	rpr.div
SER	86.1	78.3	88.8	82.4	

TABLEAU 5.20 – Performances obtenues par la *baseline* et les trois versions de SADDE dans la tâche de détection des classes de disfluences, pour la stratégie d'identification avec structuration intermédiaire (Tâche 1.Y – objectif (AN) – deux séquences).

obtenir au moins un résultat proche de celui obtenu par le modèle CRF le moins bon par ailleurs, SADDE$_A$. Nous n'avons pour l'instant aucune explication à donner sur ce résultat pour le moins surprenant.

Nous avons déterminé que, de manière générale, une structuration précise des séquences à identifier permet d'augmenter les performances du système de détection, par rapport à une détection sans structuration interne de la séquence. Nous souhaitons également déterminer, par la mise en regard des expériences menées avec et sans distinction des classes de disfluences, si la distinction de ces dernières permet d'identifier certaines classes particulièrement problématiques lorsque l'on cherche à détecter la totalité d'un segment disfluent.

Nous comparons à ces fins les résultats obtenus dans la configuration suivante : meilleure stratégie de structuration (c'est-à-dire identification de trois séquences

	BSL	SADDE$_L$	SADDE$_A$	SADDE$_{LA}$	Étiquettes
P	0.0	52.9	51.0	50.3	
	0.0	69.6	63.2	63.0	rpd.rep
	0.0	43.5	37.5	50.0	rpd.rev
	0.0	6.2	12.5	8.3	rpd.rest
	0.0	34.8	36.4	26.9	rpd.div
	0.0	85.7	50.0	80.0	int.rep
	0.0	50.0	0.0	33.3	int.rev
	0.0	0.0	0.0	0.0	int.div
	0.0	66.7	63.2	63.0	rpr.rep
	0.0	41.3	37.5	38.6	rpr.rev
	0.0	40.0	37.5	33.3	rpr.div
R	0.0	31.2	19.9	29.6	
	0.0	51.6	38.7	49.5	rpd.rep
	0.0	23.5	7.1	25.9	rpd.rev
	0.0	2.8	2.8	2.8	rpd.rest
	0.0	17.4	17.4	15.2	rpd.div
	0.0	33.3	5.6	22.2	int.rep
	0.0	25.0	0.0	8.3	int.rev
	0.0	0.0	0.0	0.0	int.div
	0.0	49.5	38.7	49.5	rpr.rep
	0.0	22.4	7.1	20.0	rpr.rev
	0.0	17.6	17.6	14.7	rpr.div
F	0.0	39.2	28.6	37.3	
	0.0	59.3	48.0	55.4	rpd.rep
	0.0	30.5	11.9	34.1	rpd.rev
	0.0	3.8	4.5	4.2	rpd.rest
	0.0	23.2	23.5	19.4	rpd.div
	0.0	48.0	10.0	34.8	int.rep
	0.0	33.3	0.0	13.3	int.rev
	0.0	0.0	0.0	0.0	int.div
	0.0	56.8	48.0	55.4	rpr.rep
	0.0	29.0	11.9	26.4	rpr.rev
	0.0	24.5	24.0	20.4	rpr.div
SER	100.0	77.3	90.9	79.8	

TABLEAU 5.21 – Performances obtenues par la *baseline* et les trois versions de SADDE dans la tâche de détection des classes de disfluences, pour la stratégie d'identification avec structuration complète (Tâche 1.Z – objectif (AN) – trois séquences).

successives : *reparandum*, *interregnum* et *reparans*), meilleur modèle CRF pour cette stratégie (c'est-à-dire construit sur des indices linguistiques : SADDE$_L$), et avec ou sans distinction de classes (c'est-à-dire tâche 1.C *versus* tâche 1.Z).

Une identification des classes dégrade les résultats pour toutes les mesures d'évaluation considérées avec une perte de 7,6 points en précision, une perte de 8,4 points en rappel et une perte de 8,6 points en f-mesure ; le Slot Error Rate est également moins bon, passant de 69,3 % avec la tâche 1.C à 77,3 % pour la tâche 1.Z, avec notamment une nette augmentation des erreurs de suppression et de substitution ; à noter toutefois une baisse des erreurs d'insertion, correspondant à treize séquences correctement non-détectées (*cf.* tableau 5.22 *infra*).

Tâches	**Corrects**	**Insertions**	**Suppressions**	**Substitutions**
1.C	39,6	14,5	49,1	11,3
1.Z	31,0	13,1	55,5	13,5

TABLEAU 5.22 – Détails du SER (en %) obtenu par SADDE_L sur la tâche de détection des segments disfluents sans identification des classes (1.C) et avec identification des classes (1.Z).

Concernant le détail des résultats par classe de disfluences, les répétitions sont particulièrement bien détectées par le système, suivies des auto-corrections et des disfluences combinées. Cette distribution suit à la fois la fréquence des classes au sein du corpus VOSDISS et le degré de complexité inhérent à chacune des classes (les répétitions sont la classe la plus simple, les disfluences combinées la classe la plus complexe). Les faux-départs sont particulièrement difficiles à détecter.

5.5.3.2 Détection des ébauches par classes de disfluences

Les résultats obtenus par les différents modèles du système SADDE pour la tâche de détection des ébauches avec distinction des classes de disfluences (*cf.* tableau 5.23 *infra*) suit la tendance des performances obtenues sur les mêmes tâches sans identification des classes (*cf.* tâches 2.B et 2.C en section 5.5.2.2). Les performances globales les plus élevées sont en effet obtenues en construisant un modèle fondé sur des indices linguistiques (SADDE_L) et en détectant l'ébauche en une seule séquence (tâche 2.Y : *reparandum* et *interregnum* confondus).

Ceci dit, concernant l'apport d'une détection par classes à la tâche d'identification des ébauches, le constat est le même que pour la détection de l'intégralité des segments disfluents : l'éclatement des segments par classe dégrade l'ensemble des résultats obtenus par comparaison avec la même tâche menée sans distinction de classes. En effet, en contrastant les résultats obtenus par SADDE_L pour les tâches 2.B et 2.Y, l'on observe à la fois une baisse de la f-mesure (-4,6 points en f-mesure) et une augmentation du Slot Error Rate (+9,8 points).

Un résultat particulièrement frappant pour cette tâche 2.Y est l'incapacité du système CRF à construire un modèle fondé sur des indices acoustiques, alors même qu'il parvient à identifier des séquences pour la même tâche sans détection des

	BSL	SADDE$_L$	SADDE$_A$	SADDE$_{LA}$	Étiquettes
P	60.8	51.8	0.0	51.0	
	74.2	68.1	0.0	67.2	rpd+int.rep
	52.6	47.4	0.0	46.0	rpd+int.rev
	0.0	7.1	0.0	11.1	rpd+int.rest
	40.0	41.7	0.0	32.0	rpd+int.div
R	23.8	32.7	0.0	29.6	
	49.5	50.5	0.0	48.4	rpd+int.rep
	11.8	31.8	0.0	27.1	rpd+int.rev
	0.0	2.8	0.0	2.8	rpd+int.rest
	13.0	21.7	0.0	17.4	rpd+int.div
F	34.3	40.1	0.0	37.5	
	59.4	58.0	0.0	56.2	rpd+int.rep
	19.2	38.0	0.0	34.1	rpd+int.rev
	0.0	4.0	0.0	4.4	rpd+int.rest
	19.7	28.6	0.0	22.5	rpd+int.div
SER	84.4	77.3	100.4	81.3	

(a) Tâche 2.Y – objectif (N) – séquence unique.

	BSL	SADDE$_L$	SADDE$_A$	SADDE$_{LA}$	Étiquettes
P	61.3	49.7	46.1	48.2	
	73.8	67.1	68.9	61.2	rpd.rep
	57.9	42.4	35.3	43.6	rpd.rev
	11.1	7.1	0.0	30.0	rpd.rest
	46.2	37.5	14.8	30.8	rpd.div
	66.7	62.5	33.3	40.0	int.rep
	0.0	42.9	0.0	57.1	int.rev
	0.0	0.0	0.0	0.0	int.div
R	22.3	31.6	18.2	28.2	
	48.4	52.7	45.2	44.1	rpd.rep
	12.9	29.4	7.1	28.2	rpd.rev
	2.8	2.8	0.0	8.3	rpd.rest
	13.0	19.6	8.7	17.4	rpd.div
	11.1	27.8	5.6	11.1	int.rep
	0.0	25.0	0.0	33.3	int.rev
	0.0	0.0	0.0	0.0	int.div
F	32.7	38.7	26.1	35.6	
	58.4	59.0	54.5	51.2	rpd.rep
	21.2	34.7	11.8	34.3	rpd.rev
	4.4	4.0	0.0	13.0	rpd.rest
	20.3	25.7	11.0	22.2	rpd.div
	19.0	38.5	9.5	17.4	int.rep
	0.0	31.6	0.0	42.1	int.rev
	0.0	0.0	0.0	0.0	int.div
SER	85.2	82.3	92.6	83.3	

(b) Tâche 2.Z – objectif (N) – deux séquences.

TABLEAU 5.23 – Performances obtenues par la *baseline* et les trois versions de SADDE dans la tâche de détection des classes d'ébauches de disfluences, pour les stratégies d'identification sans structuration (a) et avec structuration (b).

classes d'une part, et pour des tâches plus complexes d'autre part (c'est-à-dire avec un nombre d'étiquettes supérieur, comme la tâche 2.Z). Il présente d'ailleurs un Slot Error Rate supérieur à 100 %, car sa seule action sur les données est une erreur d'insertion, soit la détection (erronée) de l'ébauche d'une disfluence combinée sur un énoncé à la fois long et complexe, contenant par ailleurs l'ébauche d'une auto-correction. Cette ébauche n'est d'ailleurs détectée dans aucune des autres onze tâches de détection envisagées. Le tableau 5.24 *infra* compare l'hypothèse et la référence de cet énoncé pour la tâche 2.Y, résultats obtenus par SADDE_A. Cet exemple est particulièrement intéressant, dans la mesure où l'objectif poursuivi avec cette tâche est rempli : l'énoncé produit par le système (« hè ça tourne tous les soirs les machines ») est effectivement nettoyé et non disfluent mais on perd conjointement de l'information sur l'enjeu du dialogue (la consommation des appareils électroménagers).

Hypothèse insertion rpd+int{ *disfluence combinée* }	**Référence** rpd+int{ *auto-correction* }
— d'accord — ouais — donc c'est pas là où ça où on consomme le plus là — hè < c'est pas là où vous consommez le plus mais euh le fait de pouvoir faire partir derrière les a- les machines si en bas et en haut > ça tourne tous les soirs les machines — hm	— d'accord — ouais — donc c'est pas là où ça où on consomme le plus là — hè c'est pas là où vous consommez le plus mais euh le fait de pouvoir faire partir derrière < les a- > les machines si en bas et en haut ça tourne tous les soirs les machines — hm

TABLEAU 5.24 – Double erreur de suppression et d'insertion sur un énoncé complexe produite par SADDE_A sur la tâche 2.Y.

5.5.3.3 Détection hybride par classes de disfluences

Dans cette dernière expérience, dont la tâche est le miroir de celle présentée en section 5.5.2.3 (objectif **A**), l'objectif est de détecter un segment hybride constitué du *reparandum* et de l'achèvement des disfluences (*reparans*) : la différence avec la tâche 3.C réside dans l'identification des classes de disfluences impliquées dans l'élaboration d'une disfluence d'édition. Nous souhaitions notamment pouvoir distinguer les faux-départs des autres catégories de disfluences, qui relèvent de stratégies discursives bien particulières et sont caractérisés par la seule énonciation de l'ébauche de la disfluence (pas de *reparans* associé). Cette catégorie est déjà particulièrement difficile à annoter par un être humain, du fait de sa fréquente confusion avec les auto-corrections. Il est en effet difficile de déterminer avec précision quels sont les facteurs objectifs à partir desquels on peut considérer qu'un

locuteur abandonne son énoncé (et produit donc un faux départ) ou le corrige (et s'auto-corrige, matérialisant l'achèvement de son ébauche d'énonciation).

Le tableau 5.25 *infra* présente les résultats obtenus par les différentes versions du système pour cette tâche de détection (tâche 3.Z). Tout comme mis en avant pour la tâche 3.C, nous sommes ici dans un enjeu d'analyse des disfluences. Avec cette tâche, l'objectif est notamment de pouvoir construire une base de disfluences d'édition pour laquelle il est possible non seulement de lier ébauche et achèvement, mais aussi de catégoriser le type de disfluence. Hormis la catégorisation, considérant ces objectifs communs, les tendances de résultats sont identiques à celles de la tâche 1.C : si l'on a pour contrainte une construction automatique de la ressource, alors la précision du résultat est à privilégier et la *baseline* est alors considérée comme le meilleur système. Si l'on s'abolit de cette contrainte et que l'on accepte une construction semi-automatique avec validation manuelle, alors nous privilégions le système avec le meilleur équilibre pour toutes les mesures d'évaluation : il s'agit du modèle fondé sur des indices linguistiques (SADDE$_L$) est alors à utiliser.

Enfin, comme pour les ensembles de tâches précédents, une détection conjointe de toutes les classes de disfluences avec leur identification dégrade les résultats pour toutes les mesures. Ce n'est donc pas une bonne stratégie pour détecter les disfluences.

5.5.4 Discussion

Nous avons évalué l système de détection automatique des disfluences d'édition SADDE sur les tâches de détections décrites en section 5.2. Ces évaluations permettent de mieux comprendre la détection de disfluences en contexte conversationnel, en répondant à des questions en lien avec *(i)* les indices impliqués dans la détection, *(ii)* les stratégies de structuration des segments disfluents et *(iii)* les classes de disfluences. De manière transversale, nos discussions sur ces aspects sont envisagées eu égard aux deux enjeux principaux d'une détection des disfluences dans des données orales : le nettoyage des données et/ou l'analyse des disfluences.

Nous avons déterminé que la meilleure stratégie est de mener une détection des segments disfluents en structurant ces derniers, avec une identification conjointe des trois éléments constitutifs d'une disfluence d'édition : le *reparandum*, l'*interregnum* et le *reparans*. En effet, plus la tâche est structurée, meilleures sont les performances, avec notamment une amélioration de la précision. Nous avons établi ces tendances en prenant le meilleur modèle CRF développé (SADDE$_L$, fondé sur des indices linguistiques et discursifs).

En effet, nous avons également identifié le type d'indice le plus à même de détecter des segments disfluents sur ce type de tâche complexe et structurée : nous avons remarqué qu'une utilisation d'indices acoustiques seuls ne constitue pas une

	BSL	SADDE$_L$	SADDE$_A$	SADDE$_{LA}$	Étiquettes
P	64.0	53.9	54.0	48.8	
	72.6	69.1	65.0	63.4	rpd.rep
	70.0	45.7	50.0	44.4	rpd.rev
	25.0	8.3	0.0	8.3	rpd.rest
	27.3	36.4	25.0	25.9	rpd.div
	71.4	70.8	65.5	65.2	rpr.rep
	50.0	37.5	41.2	34.1	rpr.rev
	37.5	42.9	38.5	33.3	rpr.div
R	23.3	30.5	21.6	29.2	
	48.4	50.5	41.9	48.4	rpd.rep
	8.2	24.7	9.4	23.5	rpd.rev
	5.6	2.8	0.0	2.8	rpd.rest
	6.5	17.4	10.9	15.2	rpd.div
	48.4	49.5	40.9	48.4	rpr.rep
	5.9	17.6	8.2	16.5	rpr.rev
	8.8	17.6	14.7	17.6	rpr.div
F	34.2	39.0	30.9	36.6	
	58.1	58.4	51.0	54.9	rpd.rep
	14.7	32.1	15.8	30.8	rpd.rev
	9.1	4.2	0.0	4.2	rpd.rest
	10.5	23.5	15.2	19.2	rpd.div
	57.7	58.2	50.3	55.6	rpr.rep
	10.5	24.0	13.7	22.2	rpr.rev
	14.3	25.0	21.3	23.1	rpr.div
SER	82.4	76.4	88.7	79.7	

TABLEAU 5.25 – Performances obtenues par la *baseline* et les trois versions de SADDE dans la tâche de détection hybride des disfluences avec distinction de classes (Tâche 3.Z – objectif (A) – deux séquences).

bonne configuration pour la détection. En revanche, l'utilisation d'indices linguistiques et discursifs permet d'obtenir de bien meilleurs résultats (les meilleures performances sont obtenues avec SADDE$_L$).

Cependant, contrairement à notre hypothèse de départ et aux résultats obtenus sur des tâches similaires faisant usage d'un champ réduit d'indices (*cf.* Dutrey *et al.* (2014b) et Dutrey *et al.* (2014a)), la combinaison d'indices pour la génération d'un nouveau modèle fondé sur des traits linguistiques, discursifs et acoustiques dégrade les résultats. Dans cette configuration, les performances sont même moins bonnes que chacun des deux modèles construits indépendamment l'un de l'autre. Nous pensons toutefois qu'une autre méthode que celle utilisée ici pour exploiter l'ensemble des traits pourrait permettre à ces indices de mieux se combiner (par exemple *via* la fusion de modèle ou une nouvelle étape de sélection de traits et d'optimisation de patrons effectuée pour la construction de ce modèle) ; en effet,

la meilleure configuration (pour la génération de patrons et la sélection d'indices) pour chacun des modèles pris indépendamment n'est pas celle qui permet aux indices de s'exprimer au mieux lorsqu'ils sont combinés pour générer un nouveau modèle d'apprentissage.

La tâche mise en avant pour ces travaux, avec une structuration complète des disfluences, est d'autant plus intéressante qu'elle permet avec l'application d'un seul modèle de détection de satisfaire les deux enjeux principaux de la détection de disfluences sur des données conversationnelles (analyse et nettoyage). Le système est ainsi proposé sans *a priori* sur la finalité envisagée de la détection.

Enfin, nos expériences menées sur la distinction des différentes classes de disfluences nous permettent d'en tirer deux conclusions principales. En premier lieu, malgré des caractéristiques différentes les unes des autres, les classes identifiées dans le cours de la détection dégradent les résultats globaux. Nous pensons que cela est aussi dû à une augmentation de la parcimonie des classes envisagées, parcimonie inhérente à une multiplication des étiquettes de détection. Avec très peu de données pour chacune des classes, il paraît difficile de construire un modèle solide. En second lieu, ces expériences ont toutefois montré les disparités de résultats selon les classes. Les répétitions, qui constituent par ailleurs 35,8 % des disfluences dans le corpus de test, sont particulièrement bien détectées. À l'inverse, les faux-départs constituent la classe la plus difficile à identifier : il s'agit d'une part de la classe la moins représentée dans nos données (13,9 % des disfluences) et d'autre part celle portant le plus à confusion dans des tâches d'annotation manuelle des disfluences (difficulté d'identification des frontières, confusion avec les auto-corrections… ; *cf.* Clavel *et al.* (2013) sur l'annotation du corpus VOXDISS).

5.6 Application de SADDE aux sorties de reconnaissance automatique de la parole

En complément des expériences menées sur les transcriptions manuelles, il est indispensable d'étudier les performances de SADDE dès lors que ce dernier est appliqué à des hypothèses issues des systèmes de reconnaissance automatique de la parole (RAP). En effet, bien que nous souhaitions développer un système indépendant des performances obtenues par ces systèmes, qui sont nécessairement amenés à évoluer, nous souhaitons tout autant évaluer notre système en « conditions réelles ».

Nous faisons l'hypothèse que les disfluences subissent de grandes variations entre leur apparition dans le signal et leur représentation dans les transcriptions. En effet, de nombreux paramètres influent sur la transcription de l'énoncé, en premier lieu la performance du système utilisé lors de l'étape de reconnaissance automatique de la parole. De ce fait, notre étude concernant les transcriptions automatiques est double : en premier lieu, nous nous attachons à identifier quels

types de variations concernent les disfluences (sont-elles supprimées des transcriptions ? transcrites avec justesse ? modifiées ? et dans quelles proportions ?) ; en second lieu, outre une évaluation générale du système de détection des disfluences développé dans ces travaux (car nous supposons une dégradation manifeste des performances), nous souhaitons identifier les éventuels changements dans le comportement et l'utilité globale des indices que nous avons mis en place pour cette détection (un modèle construit sur des indices acoustiques verra-t-il ses performances augmenter par rapport à son application à des transcriptions manuelles ?).

Nous décrivons en section 5.6.1 les principes expérimentaux de cette étude des disfluences et de leur détection dans des transcriptions automatiques. La section 5.6.2 est dédiée à une analyse de l'impact de l'étape de RAP sur les disfluences, par le biais d'une typologie des formes prises par cet impact. Nous évaluons en section 5.6.3 le système de détection automatique des disfluences d'édition, `SADDE`, sur les tâches les plus complètes parmi celles conçues et évaluées sur les transcriptions manuelles.

5.6.1 Principes expérimentaux

Nous avons mis au point une chaîne de traitement des transcriptions automatiques permettant l'analyse et la détection des disfluences d'édition. Cette chaîne, décrite en figure 5.2 *infra*, s'articule autour des deux chaînes de traitement respectivement décrites en section 4.2 et 5.4.4.

Les deux objectifs de cette étape de traitements sont les suivants : *(i)* enrichir les transcriptions automatiques avec les indices acoustiques, lexicaux et discursifs que nous avons mis au point sur les transcriptions manuelles et le signal de parole, et *(ii)* obtenir, sur les transcriptions automatiques, une annotation de référence en disfluences d'édition correspondant aux disfluences manuellement annotées sur les transcriptions manuelles.

Nous avons pour cela automatiquement projeté les annotations de référence sur les transcriptions automatiques, en suivant la méthode présentée dans Galibert *et al.* (2011) ; pour cette étape nous avons opéré une correction manuelle des annotations afin de prendre en compte les erreurs d'alignement. Nous avons également mis en œuvre une étape d'appariement de ces annotations avec les sorties de la chaîne d'enrichissement des données orales conversationnelles, afin d'obtenir un corpus de transcriptions automatiques enrichi et annoté en disfluences. Nous avons ensuite transmis ce corpus au module d'étiquetage en disfluences mis en œuvre avec `Wapiti` et d'évaluation des performances de ce dernier.

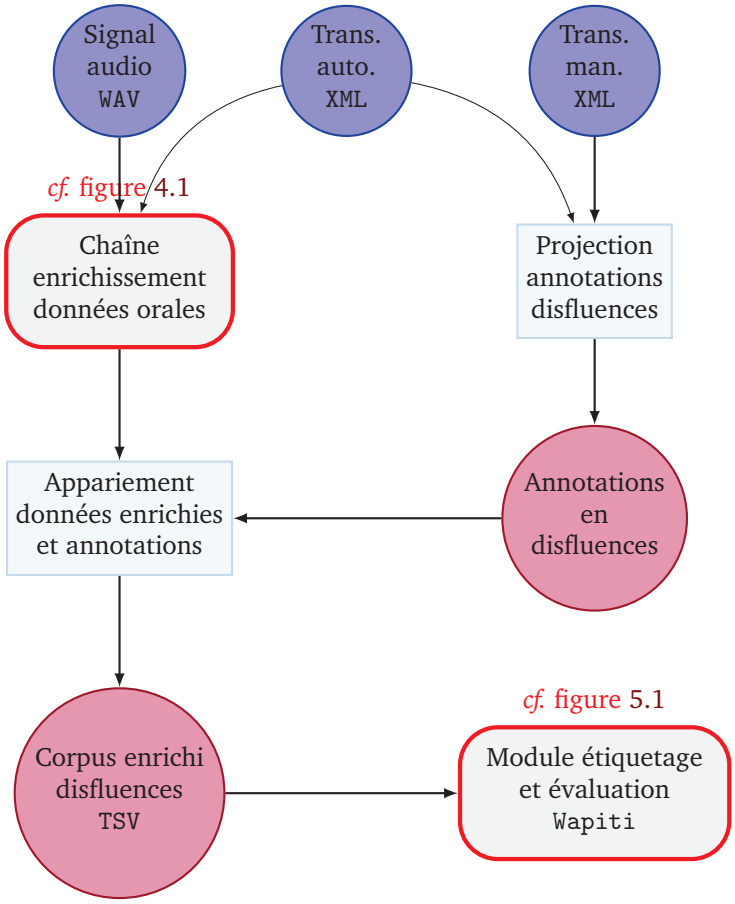


FIGURE 5.2 – Chaîne de traitement de transcriptions automatiques.

5.6.2 Impact de l'étape de Reconnaissance Automatique de la Parole sur les disfluences

De par leur fonctionnement, les systèmes de reconnaissance automatique de la parole ont un fort impact sur la transcription des disfluences d'édition. Un des facteurs principaux de cet impact est la capacité du système à identifier et à retranscrire correctement un mot. Le taux d'erreur mot du corpus VOXDISS avoisine les 33,8 % : il a été calculé pour un jeu de données constitué de 6h de parole piochée au sein des 10h du corpus VOXDISS (*cf.* Clavel *et al.* (2013) pour une description détaillée du système de reconnaissance automatique de la parole utilisé pour transcrire ces données).

Nous comparons tout d'abord les caractéristiques principales du corpus de test entre les transcriptions manuelles et les transcriptions automatiques (*cf.* ta-

bleau 5.26 *infra*). Outre une perte effective de 10 % de mots sur les transcriptions automatiques, on observe une augmentation du nombre moyen de phonèmes par mot : on peut ainsi supposer que le système a tendance à identifier et transcrire des mots plus longs que ceux réellement prononcés. La durée moyenne des mots est cependant plus brève dans les transcriptions automatiques.

	test : trans. man.	**test : trans. auto.**
Nombre de mots graphiques	13 292	11 697
Vocabulaire	935	933
Nb moy. phonèmes / mots	2,9	3,2
Durée moy. phonèmes (s.)	0,08	0,07
Durée moyenne mots (s.)	0,224	0,218
Longueur syll. moy. mots	1,5	1,5
F_0 moyenne (Hz)	178,8	175,8
Moyennes par appel :		
Nombre de mots	1 899	1 671

TABLEAU 5.26 – Caractéristiques lexicales et acoustiques du corpus de test, transcriptions manuelles *versus* transcriptions automatiques.

En lien avec ce constat, nous avons également pu observer l'impact de cette « perte » de mots sur les disfluences. Le nombre de disfluences présentes dans les sorties de RAP baisse par rapport à la référence (transcriptions manuelles), comme illustré dans le tableau 5.27 *infra*. En effet, on note une perte de 17 % sur le nombre total de disfluences, bien que cette perte ne touche pas de manière identique toutes les classes de disfluences considérées (*cf.* figure 5.3 *infra*).

	test : trans. man.				**test : trans. auto**			
	rpd	int	rpr	dis	rpd	int	rpr	dis
Nombre	260	31	212	260	215	5	164	215
Densité lex. (%)	5,0	0,4	3,0	8,4	4,7	0,2	2,3	7,2
Longueur moy. (mots)	2,5	1,9	1,9	4,3	2,5	5,6	1,6	3,9
Moyennes par appel :								
Nombre	37,1	4,4	30,3	37,1	30,7	0,7	23,4	30,7

TABLEAU 5.27 – Répartition des disfluences dans le corpus de test : comparaison entre transcriptions manuelles et transcriptions automatiques.

Cette perte est particulièrement frappante pour les *interregnum*, avec 83,9 % d'*interregnum* supprimés. De plus, les sorties de RAP contiennent 53,3 % de mots en moins au sein des *interregnum* (32,2 % en moins sur les *reparans* et 17,1 % sur les *reparandum*). Le nombre de disfluences présentes dans les transcriptions

automatiques a été calculé à partir des annotations qui y ont été projetées et manuellement corrigées.

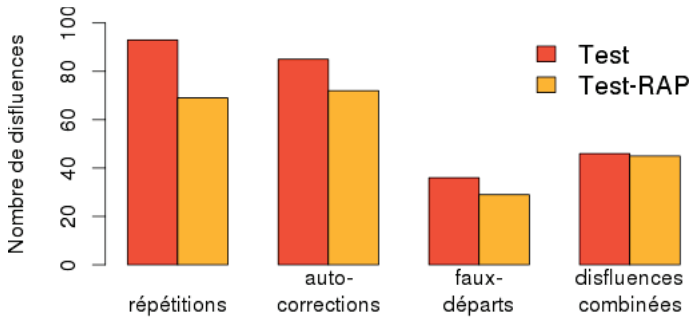


FIGURE 5.3 – Comparaison du nombre de disfluences par classe sur le corpus de test entre les transcription manuelles (Test) et les transcriptions automatiques (Test-RAP).

Nous expliquons la forte proportion de suppression des *interregnum* par la suppression des hésitations vocaliques (telles « hm », « euh » ou « hum », qui ont respectivement 70, 213 et 17 occurrences dans les transcriptions manuelles) et la forte diminution des interjections (-55,5 % dans les sorties de reconnaissance automatique de la parole par rapport aux transcriptions manuelles). En effet, comme nous l'avons analysé au sein du chapitre 4, ces éléments sont hautement caractéristique des *interregnum*, qui sont sauf exception composés de pauses silencieuses, d'hésitations vocaliques et de marqueurs discursifs (*cf.* structuration et typologie des disfluences d'édition, chapitre 2).

Un autre facteur d'impact dans les variations possibles de l'actualisation des disfluences réside dans la disparition des amorces avec le processus de reconnaissance automatique de la parole. En effet, là où un annotateur humain reconnaît un fragment de morphème, qu'il est capable d'annoter, le système remplace ce dernier par un mot appartenant à son lexique. Or, nous avons déterminé lors de nos analyses des disfluences en corpus (*cf.* chapitre 4) que les amorces pouvaient avoir un fort potentiel de caractérisation des disfluences, dans la mesure où 77 % des fragments présents dans le corpus VOXDISS apparaissent à l'intérieur d'un segment disfluent, dont 99 % dans le *reparandum*.

De manière générale, nous avons dressé une typologie de l'impact de l'étape de reconnaissance automatique de la parole sur les disfluences selon trois classes :

1. suppression des disfluences produites par les locuteurs ;
2. modification de la classe des disfluences ;
3. génération de disfluences non énoncées par le locuteur (c'est-à-dire absentes du signal de parole).

La suppression des disfluences peut survenir selon deux procédés illustrés dans le tableau 5.28 *infra* : soit une omission pure et simple dans la transcription automatique ; soit une transformation de la transcription, comme illustré avec la transformation d'une amorce. Ici, le système remplace « pas s- » par « parce que », produisant de ce fait un énoncé disfluent (mais différent de l'énoncé effectivement prononcé par le locuteur).

Hypothèse : suppression de la répétition	**Référence** : répétition
— une nouvelle	— une euh une nouvelle
Hypothèse : suppression de l'auto-correction	**Référence** : auto-correction
— les acceptez *parce que* je n'y peux plus rien	— les acceptez *pas* s- je n'y peux plus rien

TABLEAU 5.28 – Exemples de suppression de disfluences par omission ou par transformation dans la transcription automatique générée par le système de reconnaissance de la parole.

Les disfluences peuvent, sans être modifiées, subir des modifications de classe. Dans ce cas, contrairement à une suppression des disfluences, l'énoncé produit par le système, en étant toujours différent de la référence, n'est pas correct sur un plan syntaxique. Le tableau 5.29 *infra* présente des exemples emblématiques de ce type de transformation.

Hypothèse : auto-correction	**Référence** : répétition
— vous envoyer un manque que ce soit	— vous envoyer avant que que ce soit
Hypothèse : répétition	**Référence** : auto-correction
— dans ce cas dans ce cas	— dans ce ca- dans ce cadre-là
Hypothèse : répétition	**Référence** : auto-correction
— quand on a résilié résilié le contrat	— quand on a réalisé euh résilié le contrat
Hypothèse : répétition	**Référence** : auto-correction
— j'avais j'avais rendez-vous	— j'ai hum j'avais un rendez-vous

TABLEAU 5.29 – Exemples de modifications de classes de disfluences dans la transcription automatique générée par le système de reconnaissance de la parole.

Enfin, les hypothèses proposées par le système de reconnaissance automatique de la parole peuvent générer des disfluences, soit en accentuant une disfluence déjà présente dans la référence (et donc effectivement prononcée par le locuteur, *cf.* exemple présenté dans le tableau 5.30 *infra*), soit en générant à tort une disfluence alors même que l'énoncé du locuteur n'en contient pas.

Hypothèse : disfluence combinée	**Référence** : répétition
— pas encore les les les caractéristiques totales	— pas encore les les caractéristiques totales

TABLEAU 5.30 – Exemple de génération de classes de disfluences dans la transcription automatique générée par le système de reconnaissance de la parole.

5.6.3 Évaluation des performances de SADDE sur les transcriptions automatiques générées par le système de Reconnaissance Automatique de la Parole

Nous avons évalué les performances de SADDE sur les transcriptions automatiques générées par le système de Reconnaissance Automatique de la Parole. Nous avons pour cela choisi d'appliquer une tâche de détection très structurée, dans la mesure où ce type de tâche donne de bons résultats sur les transcriptions manuelles, tout en étant générique du point de vue des applications de la détection de disfluences. Cependant, nos analyses concernant l'impact de la RAP sur les disfluences ont mis en avant une suppression des *interregnum*. Nous avons de ce fait choisi d'appliquer une tâche de détection hybride (*cf.* tâche 3.C). Les résultats pour cette tâche (3.C$_{RAP}$), sont présentés dans le tableau 5.31 *infra*.

	BSL		SADDE$_L$		SADDE$_A$		SADDE$_{LA}$		Étiquettes
	reco	man	reco	man	reco	man	reco	man	
P	36,2	*69,7*	22,7	*57,5*	5,6	*58,9*	7,8	*56,5*	
	34,1	*62,9*	21,8	*56,2*	5,3	*56,0*	7,6	*54,4*	rpd
	38,4	*78,2*	23,8	*59,3*	6,1	*62,4*	7,9	*59,2*	rpr
R	16,2	*25,8*	20,3	*39,6*	23,7	*25,2*	27,8	*38,8*	
	13,5	*23,5*	18,5	*40,0*	22,5	*23,5*	26,1	*38,1*	rpd
	19,8	*28,8*	22,8	*39,2*	25,1	*27,4*	29,9	*39,6*	rpr
F	22,4	*37,7*	21,4	*46,9*	9,1	*35,3*	12,1	*46,0*	
	19,4	*34,2*	20,0	*46,7*	8,6	*33,1*	11,8	*44,8*	rpd
	26,1	*42,1*	23,2	*47,2*	9,8	*38,0*	12,5	*47,5*	rpr
SER	103,3	*75,5*	131,9	*70,9*	432,5	*79,7*	363,0	*71,7*	

TABLEAU 5.31 – Performances obtenues sur les sorties de reconnaissance automatique de la parole par la *baseline* et les trois versions de SADDE pour la tâche de détection hybride sans distinction de classes (Tâche 3.C$_{RAP}$) ; comparaison avec les résultats obtenus sur les transcriptions manuelles (Tâche 3.C) ; man = transcriptions manuelles et reco = transcriptions automatiques.

Tout d'abord, les performances de SADDE sont extrêmement dégradées sur les sorties de RAP. De plus, nous avions fait l'hypothèse que le modèle fondé sur des indices acoustico-prosodiques obtiendrait de meilleures performances sur les transcriptions automatiques que sur les transcriptions manuelles, comparativement aux autres modèles. Cette hypothèse est donc invalidée, et c'est ici aussi les modèles fondés sur des indices linguistiques qui donnent de meilleurs résultats que ceux fondés sur des indices acoustiques. À noter que d'une part la *baseline* permet d'obtenir la meilleure précision, alors que SADDE_{LA} permet l'obtention du meilleur rappel. Les résultats sont donc très contrastés.

5.7 Discussion

Nous avons mis au point, au sein de ce chapitre, un système de détection automatique de disfluences d'édition en contexte conversationnel fondé sur les champs aléatoires conditionnels : SADDE. L'identification de phénomènes disfluents dans les données issues de la parole constitue un défi important pour l'amélioration de la connaissance de ce type de données, tant du point de vue de leur traitement automatique (analyses syntaxiques, étiquetage morpho-syntaxique, analyse de sentiments, *etc.*) que de leur analyse (compréhension des stratégies discursives et dialogiques, caractérisation de l'oral spontané, *etc.*). Au-delà de ces aspects, nous travaillons dans le contexte particulièrement difficile des données de centres d'appels, sur un corpus comportant des cadres dialogiques variés et des spécificités techniques (par exemple, enregistrement des appels sur un seul canal).

En prenant en compte les méthodes de détection automatique à l'état de l'art pour des tâches similaires de détection des phénomènes disfluents, notre apport au domaine est de mener une détection des disfluences complète, structurée et répondant aux enjeux scientifiques et applicatifs liés :

1. Détection complète car SADDE est capable de prendre en compte toutes les classes de disfluences d'édition possibles, pour lesquelles nous avons au préalable dressée une typologie (*cf.* chapitre 2). Ces classes incluent des répétitions, des auto-corrections, des faux-départs et des disfluences combinées (répétitions et/ou enchevêtrement des classes sus-citées) ;
2. Détection structurée, car SADDE est capable d'identifier à la fois l'ébauche et l'achèvement des disfluences d'édition, en détectant les trois éléments qui en sont constitutifs : *reparandum*, *interregnum* et *reparans* ;
3. Détection en lien avec les enjeux scientifiques et applicatifs, car de par sa finesse de structuration SADDE permet à la fois d'identifier et de supprimer les segments réellement disfluents au sein des énoncés, et donc de nettoyer les données orales afin d'augmenter l'efficacité des traitements automatiques, et d'identifier les achèvements liés à ces segments, de manière à pouvoir

notamment étudier les stratégies discursives mises en place par les locuteurs.

Pour construire nos modèles de détection, nous avons mis à contribution des indices caractérisant les disfluences sur quatre plans : linguistique, acoustique et prosodique, discursif et para-linguistique (utilisation des méta-données inhérentes aux corpus conversationnels). Nous avons ainsi été en mesure d'évaluer l'apport de ces indices pour la détection, d'une part en lien avec les analyses de corpus décrites dans le chapitre 4, d'autre part en lien avec le type de tâche de détection. En effet, nous avions au préalable construit une chaîne d'enrichissement des données orales conversationnelles avec de nombreux traits lexicaux et acoustico-prosodiques. Ces traits ont été analysés vis-à-vis de leur pertinence dans la caractérisation des disfluences d'édition dans des données d'oral spontané. Nous avons par la suite testé leur apport vis-à-vis de la détection automatique, et avons notamment établi que certains événements *a priori* liés à l'apparition de disfluences, comme les pauses silencieuses, n'étaient en fait pas utiles pour détecter automatiquement ces dernières (pour les modèles CRF tels que nous les avons construits).

Nous avons par ailleurs créé trois modèles pour la détection de disfluences à base de CRF : un modèle fondé sur des indices linguistiques et discursifs, un modèle fondé sur des indices acoustico-prosodiques, et un modèle mixte utilisant tous les types d'indices sus-cités. L'évaluation comparative de ces trois modèles sur chacune des tâches de détection créées a ainsi mis en avant la faiblesse d'une méthode basée sur des indices acoustiques exclusivement, et le fait que les disfluences d'édition sont davantage caractérisées par des indices linguistiques et discursifs. Notre hypothèse était cependant qu'une utilisation conjointe d'indices acoustico-prosodiques et linguistiques permettrait une meilleure expression de ces traits afin de générer un modèle de détection plus efficace. Nos résultats invalident cette hypothèse ; toutefois, nous pensons qu'une autre méthode d'utilisation conjointe de tous ces indices permettrait d'inverser cette tendance. En effet, nous avons optimisé l'utilisation des indices indépendamment les uns des autres, c'est-à-dire nous avons optimisé l'utilisation d'indices acoustico-prosodiques pour la construction de $\texttt{SADDE}_A$, et parallèlement et de manière indépendante l'utilisation d'indices linguistiques pour la construction de $\texttt{SADDE}_L$, étapes à l'issue desquelles nous avons agrégé les patrons utilisant ces indices pour construire $\texttt{SADDE}_{LA}$. Nous pensons à d'autres stratégies de combinaisons :

- soit une optimisation conjointe de l'utilisation de tous les indices afin de construire un nouveau modèle mixte ;
- soit une fusion des modèles « purs » de manière à générer un modèle mixte à partir de ces deux modèles (sans redéfinition de patrons) ;
- soit la mise en œuvre d'un système de vote permettant de choisir la meilleure stratégie d'étiquetage entre les sorties des deux modèles $\texttt{SADDE}_A$ et $\texttt{SADDE}_L$.

Envisager d'autres stratégies de combinaisons impliquerait de disposer de plus de données d'entraînement ; nous sommes en effet limitée dans nos expériences

par la petite taille des corpus. Toutefois, nous pensons que les résultats obtenus sont suffisants pour être utilisés dans le cadre d'une annotation semi-automatique, précisément pour augmenter la taille des données d'apprentissage ; la mise en place d'un apprentissage actif, par le biais d'une pré-annotation des données avec SADDE, accompagnée d'une étape de correction manuelle, permettrait en effet d'obtenir de plus grands corpus et donc la génération de meilleurs modèles.

Enfin, bien qu'ayant construits et appliqués ces modèles sur des transcriptions manuelles afin de rendre le système indépendant des performances de systèmes de RAP, nous avons évalué la dégradation des résultats sur les transcriptions automatiques du corpus de test ayant servi à l'évaluation des tâches de détection. Contrairement à ce que nous pensions, les indices acoustiques donnent toujours les résultats les moins bons. Nous avons dressé une typologie de l'impact de la RAP sur la présence de disfluences dans les transcriptions automatiques, analyse qui à notre connaissance n'a jamais été menée sur des données orales conversationnelles.

Chapitre 6

Conclusion & Perspectives

Sommaire

CETTE thèse en Traitement Automatique des Langues naturelles (TAL) s'inscrit dans le cadre du développement de méthodes d'analyse robuste pour le traitement de donnés issues de la parole spontanée. En effet, extraire de l'information langagière de données de parole spontanée nécessite le développement de méthodes et l'utilisation de techniques adaptées, prenant en compte à la fois *(i)* les spécificités liées au signal de parole et *(ii)* celles liées aux retranscriptions manuelle et automatique de ce dernier. De nombreux défis liés à la gestion de la variabilité dans la parole et aux différents modes d'expression sont notamment à prendre en compte.

Ces travaux de thèse se situent donc à la frontière entre le TAL et le Traitement de la Parole. Une dimension supplémentaire est caractérisée par le fait que nos travaux portent plus particulièrement sur des conversations téléphoniques, impliquant de prendre en compte à la fois la dimension interaction homme-homme et les spécificités liées aux enregistrement du signal.

Plus particulièrement, cette thèse est dédiée à **l'analyse et à la détection automatique de disfluences dans la parole spontanée conversationnelle**. Ces phénomènes caractéristiques des énoncés spontanés posent de nombreux défis dans le traitement des corpus oraux, tout en ayant un rôle très important à jouer dans la production de l'énoncé et le déroulement du dialogue.

Nous dressons, en section 6.1, un bilan des travaux que nous avons réalisés dans cette perspective, avant de décrire en section 6.2 les nombreuses perspectives de recherches ouvertes à l'issue de ces travaux, tant dans la poursuite de recherches s'inscrivant directement dans la continuité de nos résultats que dans la réutilisation de ces travaux pour répondre à d'autres défis scientifiques et applicatifs.

6.1 Conclusions suite au travail de thèse

Dans cette thèse, nous avons décrit les défis posés par la mise en place de méthodes d'analyse robuste destinées au traitement de données issues de l'oral, et plus spécifiquement lorsque ces méthodes ont pour objectif de permettre la réalisation de tâches de fouille de données et d'extraction d'information sur des données d'oral spontané. Nous avons également défini les enjeux liés à l'analyse et surtout à l'identification des phénomènes caractéristiques de l'oral spontané dans des données conversationnelles, qui relèvent de deux dimensions :

- une meilleure modélisation de la parole par le biais de l'étude des moyens mis en place par des locuteurs pour construire leur discours et gérer l'interaction dans des dialogues homme-homme ; il s'agit d'un enjeu d'analyse des donnés ;
- une amélioration des applications de TAL sur des données d'oral spontané et une meilleure lisibilité de ces données par des humains (dès lors que l'on

n'est pas dans une perspective d'étude de la parole) ; il s'agit d'un enjeu de nettoyage de données.

Afin de circonscrire notre objet d'étude, les disfluences, et d'orienter nos travaux liés à leur analyse et à leur détection, nous avons dressé un état de l'art des recherches menées sur les phénomènes d'oral spontané : nous avons ainsi délimité le champ de phénomènes auxquels se rattachent les événements disfluents et avons mis en exergue les disfluences d'édition, qui regroupent un ensemble d'événements ayant une forte incidence sur le déroulement syntagmatique de l'énoncé, à l'inverse d'autres phénomènes comme les *fillers* [1]. En effet, ces derniers (qui regroupent les amorces, les hésitations vocaliques et les marqueurs discursifs) représentent des unités bien circonscrites dans le discours et les difficultés liées à leur identification ne reposent plus sur des défis de détection de leurs bornes dans l'énoncé mais résident plutôt dans le fait de valider leur appartenance à une classe relevant de la disfluence. Autrement dit, les défis scientifiques liés à leur identification relèvent davantage de la sémantique afin de déterminer si, une fois détectés dans la parole, ces éléments sont effectivement disfluents ou exercent une autre fonction dans le discours.

Les défis liés à l'identification des disfluences d'édition sont tout autre et se situent au niveau de la capacité des systèmes à les identifier dans le discours. Nous avons ainsi dressé un état de l'art des méthodes et systèmes développés pour détecter ces phénomènes dans des corpus d'oral spontané. De nombreuses méthodes ont été mises au point, sur l'anglais notamment, pour identifier toutes les classes de disfluences d'édition existantes (répétitions, auto-corrections, faux-départs, disfluences combinées). Seulement, nous avons également mis en exergue le caractère partiel de ces méthodes, qui s'attachent à la détection d'une partie seulement des disfluences d'édition dans le discours.

En complément à cet état de l'art sur la caractérisation des disfluences dans les données d'oral spontané, nous avons mis au point une chaîne de traitement et d'enrichissement de corpus de parole conversationnelle, que nous avons pour ces travaux orientée vers l'analyse des disfluences en corpus. En effet, cette méthode d'analyse permet :

1. d'enrichir automatiquement un corpus brut de nombreuses informations liées à l'analyse du signal de parole (extractions acoustiques et génération de traits de prosodie), à l'analyse des transcriptions manuelles ou automatiques de ce signal (analyse morpho-syntaxique) et à l'extraction d'informations discursives en lien avec le traitement de données conversationnelles (progression du dialogue).

2. d'analyser un aspect des données (dans notre cas les disfluences, grâce à l'extraction d'annotations présentes dans le corpus EDF sur lequel nous

1. *Cf.* section 2.2.2 page 20 pour une description de ces phénomènes.

avons travaillé) en confrontation avec ces nouveaux indices issus de l'étape d'enrichissement des données.

Grâce à cette méthode, nous avons pu mettre en exergue des traits caractéristiques de l'apparition des disfluences dans le discours et au sein du dialogue. Les résultats de cette analyse de corpus pointent toutefois le fait que les disfluences sont soumises à de nombreuses variations et que leur apparition dépend d'un grand nombre d'informations extérieures.

Les résultats de ces recherches et travaux sur la caractérisation des disfluences en contexte conversationnel et les méthodes de détection automatique nous ont permis de développer un système de détection automatique, SADDE (Système Automatique de Détection des Disfluences d'Édition), qui mène une tâche de détection complète, structurée, et répondant aux enjeux scientifiques et applicatifs liés. En effet, ce système basé sur une méthode d'apprentissage automatique supervisé utilisant les champs aléatoires conditionnels (CRF) est capable de prendre en compte toutes les classes de disfluences d'édition et de structurer les éléments inclus dans la construction énonciative de ces phénomènes. Cette finesse de structuration permet à la fois un nettoyage des données orales pour faciliter des traitements automatiques et une étude des stratégies discursives mises en place par les locuteurs. Nous avons par ailleurs pris en compte toutes les dimensions caractérisant les disfluences dans les données orales spontanées, en utilisant des indices linguistiques, acoustico-prosodiques, discursifs et para-linguistiques.

En lien avec les enjeux que nous avons rappelés au début de cette section, nous avons mis en place différentes expériences, chacune modélisant une tâche correspondant à un besoin applicatif spécifique. Pour l'ensemble de ces tâches, trois modèles d'étiquetage séquentiel ont été évalués et comparés, afin d'analyser l'efficacité des indices permettant la détection des disfluences d'édition : un modèle à base d'indices linguistiques, un modèle à base d'indices acoustico-prosodiques, un modèle utilisant ces deux types d'indices. Les résultats, s'ils ne sont pas très élevés, permettent d'ores et déjà de tirer quelques conclusions intéressantes.

Nous avons notamment déterminé que la meilleure stratégie est de mener une détection des segments disfluents en structurant ces derniers, avec une identification conjointe des trois éléments constitutifs d'une disfluence : le *reparandum*, l'*interregnum* et le *reparans* [2]. En effet, plus la tâche est structurée, meilleures sont les performances, avec notamment une amélioration de la précision.

Nous avons également montré qu'une détection identifiant de manière distincte les classes de disfluences considérées (soit toutes les classes possibles : répétitions, auto-corrections, faux-départs et disfluences combinées) dégradait les résultats obtenus. L'application d'un classifieur sur les sorties du système pour les tâches de détection sans distinction de classes serait certainement la bonne méthode pour compléter cette tâche.

2. *Cf.* section 2.3.1 page 25 pour une étude de la structure interne des disfluences d'édition.

Enfin, le meilleur modèle pour l'évaluation de la tâche la plus complète (dans la mesure où elle peut répondre à tous les enjeux liés à l'identification de disfluences d'édition, dont elle identifie la structure complète) a été appliqué sur le corpus de test transcrit automatiquement. Concernant l'évaluation du système, nous pouvons observer que les résultats sont extrêmement dégradés, en particulier du point de vue de la précision du système. Cette analyse nous a cependant permis de mettre en évidence une typologie des modifications subies par les zones disfluentes lors du processus de reconnaissance automatique de la parole.

6.2 Perspectives de recherche

Les résultats de cette thèse ouvrent de très nombreuses perspectives de recherche, tant sur la continuité directe des travaux menés sur l'analyse et la détection de disfluences que sur leur réutilisation pour répondre à d'autres défis scientifiques et applicatifs.

6.2.1 Amélioration du système de détection des disfluences

Tout d'abord, nous envisageons un certain nombre de pistes de recherches visant à l'amélioration du système de détection automatique, SADDE.

Gestion de la parcimonie La littérature sur l'analyse des disfluences et nos travaux menés sur leur caractérisation en contexte conversationnel montrent qu'à l'exception de certains indices discursifs ou liés au cadre conversationnel (genre du locuteur, distribution des disfluences au sein des conversations, *etc.*) les phénomènes disfluents sont surtout caractérisés par des variations lexicales et acoustico-prosodiques dans leur contexte local. Ce constat, couplé avec le fait que les disfluences apparaissent de manière très parcimonieuse dans les données, nous donne à penser que les performances de notre système pourraient être grandement améliorées si nous prenions en compte cette parcimonie dans les données *avant* d'entraîner le système de détection. Nous faisons ainsi l'hypothèse qu'en donnant au système des données d'entraînement présentant un très fort taux de disfluences (ce qui pourrait revenir à lui donner des séquences d'exemples de disfluences d'édition avec un contexte local minimal) les nombreux indices sollicités pour la détection pourraient mieux s'exprimer et mieux être mis à contribution.

Alternatives de construction de modèle mixte Contrairement à notre hypothèse, qui était que la meilleure manière de détecter des disfluences était de faire usage conjointement d'indices acoustico-prosodiques et linguistiques (dans la mesure où les disfluences sont caractérisées sur ces deux plans), nos résultats montrent que la construction d'un modèle fondé sur des indices mixtes donne de

moins bonnes performances que les autres modèles : en particulier, cela donne sur certaines tâches de plus mauvais résultats que les deux modèles « purs » appliqués individuellement. Nous souhaitons expérimenter d'autres manières d'utiliser les indices mixtes, dans la mesure où *(i)* les travaux scientifiques portant sur la détection de disfluences avec des méthodes d'apprentissage supervisé montrent que la combinaison d'indices donnent les meilleurs résultats et *(ii)* dans nos propres travaux, avec un jeu d'indices beaucoup moins développé que celui mis au point dans la thèse, les modèles combinant indices lexicaux et acoustiques donnent également les meilleurs résultats. Nous envisageons deux méthodes pour améliorer les performances obtenues par SADDE avec des indices mixtes :

- opérer une étape d'optimisation des patrons et de sélection des indices au moment de la construction du modèle d'apprentissage ;
- construire un modèle mixte à partir des modèles construits indépendamment sur les indices acoustico-prosodiques et linguistiques (fusion de modèles).

Analyse des stratégies d'annotation La détection de disfluences combinées pose un défi particulièrement difficile non seulement pour leur reconnaissance, mais aussi pour l'évaluation des systèmes de détection : nous souhaitons étudier plus attentivement la stratégie mise en place par le système pour repérer et surtout structurer cette classe de disfluences. En effet, lors de l'apparition successive ou imbriquée de plusieurs disfluences d'édition, appartenant potentiellement à des classes différentes, quelle est la stratégie mise en œuvre par le système ? peut-on dégager une tendance permettant d'affirmer qu'il va « aller au plus précis », c'est-à-dire détecter plusieurs disfluences successives, ou au contraire identifier un segment plus large englobant toute la zone de disfluence (auquel cas la structuration sera moins fine) ?

Augmentation des données d'apprentissage De manière générale, les systèmes d'apprentissage automatique sont très sensibles à la quantité de données utilisées pour l'apprentissage : une augmentation significative de la quantité de données d'apprentissage permettrait certainement l'obtention d'un système plus performant et plus robuste. De plus, les résultats obtenus par SADDE sur la tâche de détection des disfluences devrait pouvoir augmenter, *via* une approche fondée sur de l'apprentissage actif par exemple, à la fois la taille des corpus annotés et les performances du système.

Amélioration de la référence Nous avons évalué notre système sur des données qui ont été manuellement annotées dans le cadre du projet Vox Factory : nous avons perçu lors de plusieurs étapes d'analyse que cette référence était loin d'être parfaite et contenait de nombreuses erreurs d'annotation (nous ne disposons malheureusement pas d'informations précises comme des mesures d'accords

inter-annotateurs) : il serait nécessaire de pouvoir entraîner et évaluer le système de détection sur une meilleure référence.

Adaptation aux sorties de Reconnaissance Automatique de la Parole Nous avons par ailleurs mené une tâche d'évaluation de la dégradation des performances obtenues par notre système sur des sorties de reconnaissance automatique de la parole (RAP) : nous envisageons d'une part d'approfondir l'analyse de l'impact de cette étape de RAP sur les données orales disfluentes et d'autre part de mener des recherches afin d'adapter le système aux transcriptions automatiques de la parole, notamment en utilisant des indices spécifiques à ces transcriptions (par exemple, les indices de confiance associés à la reconnaissance des mots).

6.2.2 Évaluation des méthodes de détection des disfluences en français

Au-delà des améliorations que nous pourrons apporter au système de détection des disfluences que nous avons développé, il nous semble indispensable d'ancrer cette tâche dans une perspective de comparaison avec les travaux menés au sein de la communauté francophone (construction de ressources, propositions de méthodes, *etc.*). Nous souhaitons mener à bien une évaluation de la généricité de notre système en l'appliquant *(i)* de manière générale à d'autres données d'oral spontané conversationnel et *(ii)* plus spécifiquement à d'autres données issues de centres d'appels. Nous envisageons, dans cette perspective, d'utiliser un deuxième corpus EDF constitué de conversations entre des agents et des clients professionnels et un corpus d'appels de la RATP, réalisé dans le cadre du projet DECODA).

6.2.3 Valorisation industrielle

En matière de valorisation industrielle, nous rappelons que nos travaux s'inscrivent chez EDF dans la continuité du projet Vox Factory, qui a notamment mis au point une interface dédiée à la fouille de données conversationnelles issues des centres d'appels. Nos travaux permettent d'intégrer à cet outil les conversations au sein desquelles les disfluences d'édition ont été détectées. Afin de valoriser au mieux cet apport, nous souhaitons mettre en œuvre une couche de traitement des disfluences préalablement détectées, afin de permettre à l'utilisateur de visualiser soit les dialogues tels qu'ils ont été énoncés, soit les dialogues exempts de disfluences d'édition (afin de faciliter la lecture et l'exploration des données), soit les dialogues présentés non plus de manière linéaire mais avec une visualisation des énoncés jouant sur une représentation conjointe des axes syntagmatique et paradigmatique, comme proposé par C. Blanche-Benveniste.

Concernant la réutilisation des travaux de thèse dans une perspective de valorisation industrielle, nous envisageons d'évaluer l'impact de la suppression des

ébauches de disfluences d'édition (permise par le fait que SADDE structure les disfluences d'édition lorsqu'il les détecte) sur l'efficacité d'autres tâches d'extraction d'information et de fouille de texte actuellement mises en place chez EDF pour l'analyse des conversations issues de centres d'appels.

6.2.4 Ouverture des travaux à d'autres domaines scientifiques

Notre travail de recherche ouvre de nombreuses possibilités pour l'utilisation des disfluences d'édition dans des domaines scientifiques connexes. En particulier, nous nous intéressons aux axes de recherches suivants, qui reflètent la place que peuvent prendre ces travaux au sein de problématiques parfois proches, parfois plus éloignées.

Pour l'étude des stratégies discursives et dialogiques, ces travaux pourraient permettre d'aboutir à la construction d'un corpus d'analyse des disfluences en contexte et à la mise en relation des disfluences avec d'autres phénomènes, comme les émotions, les opinions ou encore les actes de dialogue. Plus spécifiquement, sur des corpus de centres d'appels, établir un lien entre phénomènes d'oral spontané et présence de lexique métier serait aussi très intéressant.

L'extraction structurée de disfluences d'édition (produisant des couples ébauche et achèvement) pourrait également permettre l'enrichissement de ressources portant sur les phénomènes de correction (concernant les auto-corrections avec changement de sens) et sur les paraphrases (concernant les auto-corrections sans changement de sens).

Enfin, dans le cadre des systèmes de dialogues homme-machine, nous pensons que la mise en exergue d'indices pertinents pour caractériser et typer l'apparition de disfluences au sein du flux de parole pourrait permettre d'améliorer la synthèse vocale et la gestion de l'interaction par des avatars : en effet, ces indices pourraient être utilisés non plus pour la détection de disfluences dans la parole mais *a contrario* pour la génération de phénomènes d'oral spontané de manière, par exemple, à augmenter le caractère naturel des synthèses vocales ou à contrôler l'énonciation de la machine (lorsque l'avatar doit par exemple inviter l'humain à prendre la parole).

Bibliographie

ADDA-DECKER, M. (2006). De la reconnaissance automatique de la parole à l'analyse linguistique de corpus oraux. *In Acte des 26e Journées d'Étude sur la Parole (JEP'06)*, pages 389–400.

ADDA-DECKER, M., GENDROT, C. et NGUYEN, N. (2008). Contributions du traitement automatique de la parole à l'étude des voyelles orales du français. *Traitement Automatique des Langues*, 49(3):13–46.

ADDA-DECKER, M., HABERT, B., BARRAS, C., ADDA, G., BOULA DE MAREÜIL, P. et PAROUBEK, P. (2003). A disfluency study for cleaning spontaneous speech automatic transcripts and improving speech language models. *In Proceedings of the 3rd Workshop on Disfluency in Spontaneous Speech (DiSS'03)*, pages 67–70.

ADDA-DECKER, M., HABERT, B., BARRAS, C., Boula de MAREÜIL, 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 25e Journées d'Étude sur la Parole (JEP'04)*.

ADDA-DECKER, M., NEMOTO, R. et DURAND, J. (2009). Stratégies de démarcation du mot en français : une étude expérimentale sur grand corpus. *In Actes des 6èmes Journées Linguistiques de Nantes*, pages 91–96.

ALLAUZEN, A. et BONNEAU-MAYNARD, H. (2008). Training and evaluation of pos taggers on the french multitag corpus. *In Proceedings of LREC*.

AMBLARD, M. et FORT, K. (2014). Étude quantitative des disfluences dans le discours de schizophrènes : automatiser pour limiter les biais. *In Actes de Traitement Automatique des Langues Naturelles (TALN)*.

AUDHKHASI, K., KANDHWAY, K., DESHMUKH, O. D. et VERMA, A. (2009). Formant-based technique for automatic filled-pauses detection in spontaneous spoken english. *In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'09)*.

BARRAS, C., GEOFFROIS, E., WU, Z. et LIBERMAN, M. (1998). Transcriber : a free tool for segmenting, labeling and transcribing speech. *In Proceedings of the 1st*

International Conference on Language Resources and Evaluation (LREC'98), pages 1373–1376.

BAZILLON, T., DEPLANO, M., BÉCHET, F., NASR, A. et FAVRE, B. (2012). Syntactic annotation of spontaneous speech : application to call-center conversation data. *In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*.

BELIAO, J. et LACHERET, A. (2013). Disfluency and discursive markers : when prosody and syntax plan discourse. *In Proceedings of the 6th Workshop on Disfluency in Spontaneous Speech (DiSS'13)*, pages 5–8.

BELIAO, J., LACHERET, A. et KAHANE, S. (2014). Discourse and prosody in spoken french : Why, what and how should one count ? a comparative statistical perspective. *Nouveaux cahiers de linguistique française*, 31:33–44.

BLANCHE-BENVENISTE, C. (1997). *Approches de la langue parlée en français*. Ophrys, 2e (2010) édition.

BLANCHE-BENVENISTE, C., BILGER, M., ROUGET, C. et VAN DEN EYNDE, K. (1990). *Le français parlé : Études grammaticales*. Éditions du CNRS.

BLANCHE-BENVENISTE, C., DELOFEU, J., STEFANINI, J. et VAN DEN EYNDE, K. (1987). *Pronom et syntaxe. L'approche pronominale et son application au français*. Peeters, 2 édition.

BOERSMA, P. et WEENINK, D. (2001). Praat, a system for doing phonetics by computer. *Glot International*, 5(9):341–345.

BONNEAU-MAYNARD, H., AYACHE, C., BÉCHET, F., DENIS, A., KUHN, A., LEFÈVRE, F., MOSTEFA, D., QUIGNARD, M., ROSSET, S., SERVAN, C. et VILLANEAU, J. (2006). Results of the french evalda-media evaluation campaign for literal understanding. *In Proceedings of the 5th international Conference on Language Resources and Evaluation (LREC'06)*, pages 2054–2059, Genoa, Italy.

BOUFADEN, N. (2004). *Extraction d'information à partir de transcriptions de conversations téléphoniques spécialisées*. Thèse de doctorat, Université de Montréal.

BOUFADEN, N., DELISLE, S. et MOULIN, B. (1998). Analyse syntaxique robuste de dialogues retranscrits : peut-on vraiment traiter l'oral à partir de l'écrit ? *In Actes de la 5e conférence annuelle sur le Traitement automatique des langues naturelle (TALN'98)*.

Boula de MAREÜIL, P., HABERT, B., BÉNARD, F., ADDA-DECKER, M., BARRAS, C., ADDA, G. et PAROUBEK, P. (2005). A quantitative study of disfluencies in french broadcast interviews. *In Proceedings of the 4th Workshop on Disfluency in Spontaneous Speech (DiSS'05)*.

BOURAOUI, J.-L. et VIGOUROUX, N. (2009). Traitement automatique de disfluences dans un corpus linguistiquement contraint. *In Actes de la 16e conférence annuelle sur le Traitement automatique des langues naturelle (TALN'09)*, Senlis, France.

BOURAOUI, J.-L. M. (2008). *Analyse, modélisation et détection automatique des disfluences dans le dialogue oral spontané contraint : le cas du contrôle aérien*. Thèse de doctorat, Université Toulouse III – Paul Sabatier.

BOVE, R. (2008). *Analyse syntaxique automatique de l'oral : étude des disfluences*. Thèse de doctorat, Université d'Aix-Marseille I.

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 13e conférence annuelle sur le Traitement automatique des langues naturelle (TALN'06)*.

BRILL, E. (1992). A simple rule-based part of speech tagger. *In Proceedings of the third conference on Applied natural language processing (ANLC '92)*, pages 152–155, Stroudsburg, PA, USA. Association for Computational Linguistics.

BÉCHET, F., MAZA, B., BIGOUROUX, N., BAZILLON, T., EL-BÈZE, M., DE MORI, R. et ARBILLOT, E. (2012). Decoda : a call-center human-human spoken conversation corpus. *In Proceedings of the 8th international Conference on Language Resources and Evaluation (LREC'12)*.

CAILLIAU, F. et GIRAUDEL, A. (2008). Enhanced search and navigation on conversational speech. *In Proceedings of SIGIR Workshop of Searching Spontaneous Conversational Speech (SSCS'08)*, pages 66–70, Singapore.

CAMPIONE, E. (2001). *Étiquetage prosodique semi-automatique de corpus oraux : algorithmes et méthodologie*. Thèse de doctorat, Université de Provence.

CANDEA, M. (2000a). *Contribution à l'étude des pauses silencieuses et des phénomènes dits d'« hésitation » en français oral spontané*. Thèse de doctorat, Université Sorbonne Nouvelle – Paris III.

CANDEA, M. (2000b). Les euh et les allongements dits d'« hésitation » : deux phénomènes soumis à certaines contraintes en français oral non lu. *In Actes des 23e Journées d'Étude sur la Parole (JEP'00)*, Aussois, France.

CANDEA, M., VASILESCU, I. et ADDA-DECKER, M. (2005). Inter- and intra-language acoustic analysis of autonomous fillers. *In Proceedings of the 4th Workshop on Disfluency in Spontaneous Speech (DiSS'05)*.

CAPPEAU, P. et SEIJIDO, M. (2005). Les corpus oraux en français. Rapport technique, DGLFLF.

CHRISTODOULIDES, G. et AVANZI, M. (2014). Phonetic and prosodic characteristics of disfluencies in french spontaneous speech. *In 14th Conference on Laboratory Phonology (LabPhon'14)*.

CLARK, H. H. (2002). Speaking in time. *Speech Communication*, 36:5–13.

CLARK, H. H. et FOX TREE, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition*, 84:73–111.

CLAVEL, C., ADDA, G., CAILLIAU, F., GARNIER-RIZET, M., CAVET, A., CHAPUIS, G., COURCINOUS, S., DANESI, C., DAQUO, A.-L., DELDOSSI, M., GUILLEMIN-LANNE, S., SEIZOU, M. et SUIGNARD, P. (2013). Spontaneous speech and opinion detection : mining call-centre transcripts. *Lang. Resources & Evaluation*, 1:40.

CONSTANT, M. et DISTER, a. (2010). *Spoken Communication*, chapitre Automatic detection of disfluencies in speech transcriptions, pages 259–272. Cambridge Scholars Publishing.

DANESI, C. et CLAVEL, C. (2010). Impact of spontaneous speech features on business concept detection : a study of call-centre data. *In Proceedings of the International Workshop on Searching Spontaneous Conversational Speech (SSCS '10)*, pages 11–14. Association for Computing Machinery (ACM).

DELIC (2004). Présentation du corpus de référence du français parlé. *Recherches sur le français parlé*, 18:11–42.

DENIS, P. et SAGOT, B. (2009). Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art pos tagging with less human effort. *In Proceedings of PACLIC'09*, Hong-Kong, China.

DINARELLI, M. et ROSSET, S. (2011). Models cascade for tree-structured named entity detection. *In Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP'11)*, pages 1269–1278, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

DINARELLI, M., STEPANOV, S., VARGES, G. et RICCARDI, G. (2010). The luna spoken dialog system : Beyond utterance classification. *In Proceedings of ICASSP*, Dallas, USA.

DUEZ, D. (2001). Acoustic-phonetic characteristics of filled pauses in spontaneous french speech : Preliminary results. *In Proceedings of the 2nd Workshop on Disfluency in Spontaneous Speech (DiSS'01)*.

DUTREY, C., CLAVEL, C., ROSSET, S., VASILESCU, I. et ADDA-DECKER, M. (2014a). A crf-based approach to automatic disfluency detection in a french call-centre corpus. *In Proceedings of InterSpeech'14*.

DUTREY, C., ROSSET, S., ADDA-DECKER, M., CLAVEL, C. et VASILESCU, I. (2014b). Disfluences dans la parole spontanée conversationnelle : détection automatique utilisant des indices lexicaux et acoustiques. *In Actes des JEP'14*.

FAYYAD, U. M. et IRANI, K. B. (1993). Multi-interval discretization of continuous-valued attributes for classification learning. *In Proceedings of the International Joint Conference on Uncertainty in AI (IJCAI'93)*, pages 1022–1027.

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.

GADET, F. (1989). *Le français ordinaire*. Armand Colin.

GALIBERT, O., ILLOUZ, G. et ROSSET, S. (2005). Ritel : An open-domain, human-computer dialog system. *In Proceedings of the 9th International Conference on Speech Communication and Technology (Interspeech'05)*, pages 2789–2792.

GALIBERT, O., LEIXA, J., ADDA, G., CHOUKRI, K. et GRAVIER, G. (2014). The ETAPE speech processing evaluation. *In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3995–3999, Reykjavik, Iceland.

GALIBERT, O., ROSSET, S., GROUIN, C., ZWEIGENBAUM, P. et QUINTARD, L. (2011). Structured and extended named entity evaluation in automatic speech transcriptions. *In Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP'11)*, pages 518–526, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

GALLIANO, S., GEOFFROIS, E., GRAVIER, G., BONASTRE, J.-F., MOSTEFA, D. et CHOUKRI, K. (2006). Corpus description of the ester evaluation campaign for the riche transcription of french broadcast. *In Proceedings of the 5th international Conference on Language Resources and Evaluation (LREC'06)*.

GALLIANO, S., GRAVIER, G. et CHAUBARD, L. (2009). The ester 2 evaluation campaign for the rich transcription of french radio broadcasts. *In Proceedings of the 10th International Conference on Speech Communication and Technology (Interspeech'09)*, pages 2583–2586.

GARCIA-FERNANDEZ, A., VASILESCU, I. et ROSSET, S. (2010). Euh as cue for speaker confidence and word searching in human spoken answers in french. *In Proceedings of the 5th Workshop on Disfluency in Spontaneous Speech – The 2nd International Symposium on Linguistic Patterns in Spontaneous Speech (DiSS-LPSS'10)*, Tokyo, Japan.

GARNIER-RIZET, M., ADDA, G., CAILLIAU, F., GAUVAIN, J.-L., GUILLEMIN-LANNE, S. et LAMEL, L. (2008). Callsurf : Automatic transcription, indexing and structuration of call center conversational speech for knowledge extraction and query by content. *In Proceedings of the 6th International Language Resources and Evaluation (LREC'08)*, pages 2623–2628.

GAUVAIN, J.-L., ADDA, G., LAMEL, L., LEFÈVRE, F. et SCHWENK, H. (2005). Transcription de la parole conversationnelle. *Traitement Automatique des Langues*, 45(3):35–47.

GEORGILA, K. (2009). Using Integer Linear Programming for Detecting Speech Disfluencies. *In Proceedings of HLT/NAACL*, pages 109–112.

GEORGILA, K., WANG, N. et GRATCH, J. (2010). Cross-domain speech disfluency detection. *In Proccedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL'10)*.

GODFREY, J., HOLLIMAN, E. et MCDANIEL, J. (1997). Switchboard : Telephone speech corpus for research and development. *IEEE ICASSP*, 1:517–520.

GROUIN, C. (2013). *Anonymisation de documents cliniques : performances et limites des méthodes symboliques et par apprentissage statistique*. Thèse de doctorat, UPMC – Paris VI.

HAHN, S., DINARELLI, M., RAYMOND, C., LEFEVRE, F., LEHNEN, P., de MORI, R., MOSCHITTI, A., NEY, H. et RICCARDI, G. (2011a). Comparing Stochastic Approaches to Spoken Language Understanding in Multiple Languages. *IEEE Transactions on Audio, Speech & Language Processing*, (6):1569–1583.

HAHN, S., LEHNEN, P. et NEY, H. (2011b). Powerful Extensions to CRFs for Grapheme to Phoneme Conversion. *In Proceedings of ICASSP*, pages 4912–4915.

HARDY, H., BAKER, K., BONNEAU-MAYNARD, H., DEVILLERS, L., ROSSET, S. et STRZALKOWSKI, T. (2003). Semantic and dialogic annotation for automated multilingual customer service. *In ISCA Eurospeech*, Genève, Suisse.

HEEMAN, P. A. (1997). *Speech Repairs, International Boundaries and Discourse Markers : Modeling Speaker's Utterances in Spoken Dialog*. Thèse de doctorat, University of Rochester (New-York, USA).

HENRY, S., CAMPIONE, E. et VÉRONIS, J. (2004). Répétitions et pauses (silencieuses et remplies) en français spontané. *In Actes des 25e Journées d'Étude sur la Parole (JEP'04)*, Fès, Morocco.

HINDLE, D. (1983). Deterministic parsing of syntactic non-fluencies. *In Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics (ACL'83)*, pages 123—-128.

HOKKANEN, T. (2001). Prosodic marking of self-repairs. *In Proceedings of the 2nd Workshop on Disfluency in Spontaneous Speech (DiSS'01)*.

JOHNSON, M. et CHARNIAK, E. (2004). A TAG-based Noisy Channel Model of Speech Repairs. *In Proceedings of ACL*.

KAUSHIK, M., TRINKLE, M. et HASHEMI-SAKHTSARI, A. (2010). Automatic detection and removal of disfluencies from spontaneous speech. *In Proceedings of the 13th Australasian International Conference on Speech Science and Technology (SST'10)*, pages 98–101, Melbourne, Australia.

KIM, J., SCHWARM, S. E. et OSTENDORF, M. (2004). Detecting structural metadata with decision trees and transformation-based learning. *In Proceedings of the Human Language Technology Conference/North American Chapter of the Association for Computational Linguistics Meeting (HLT/NAACL'04)*, pages 137–144.

LABOV, W. (1966). Hypercorrection by the lower middle class as a factor in linguistic change. *In* BRIGHT, W., éditeur : *Sociolinguistics : Proceedings of the UCLA Sociolinguistics Conference, 1964*, pages 84–113. Mouton.

LAFFERTY, J. D., MCCALLUM, A. et PEREIRA., F. C. N. (2001). Conditional Random Fields : Probabilistic Models for Segmenting and Labeling Sequence Data. *In Proceedings of International Conference on Machine Learning*, pages 282–289.

LAVERGNE, T., CAPPÉ, O. et YVON, F. (2010). Practical Very Large Scale CRFs. *In Proceedings ACL*, pages 504–513.

LEVELT, W. J. M. (1983). Monitoring and self-repair in speech. *Cognition*, 14: 41–104.

LICKLEY, R. J. (1994). *Detecting Disfluency in Spontaneous Speech*. Thèse de doctorat, University of Edinburgh.

LIU, Y., SHRIBERG, E. E., STOLCKE, A., HILLARD, D., OSTENDORF, M. et HARPER, M. (2006). Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE Transactions on Audio, Speech, and Language Processing*, 14:1526 – 1540.

MACLAY, H. et OSGOOD, C. (1959). Hesitation phenomena in spontaneous english speech. *Words*, 15:19–44.

MAKHOUL, J., KUBALA, F., SCHWARTZ, R. et WEISCHEDEL, R. (1999). Performance Measures For Information Extraction. *In Proceedings of DARPA Broadcast News Workshop*, pages 249–252.

MOREL, M.-A. et DANON-BOILEAU, L. (1998). *Grammaire de l'intonation. L'exemple du français oral*. Ophrys.

NEMOTO, R. (2011). *Large-scale Acoustic and Prosodic Investigations of French*. Thèse de doctorat, Université Paris-Sud XI.

NGUYEN, N. et ADDA-DECKER, M. (2013). *Traité IC2 Cognition et Traitement de l'Information – Méthodes et outils pour l'analyse phonétique des grands corpus oraux*. Hermes – Lavoisier.

O'CONNELL, D. C. et KOWAL, S. (2005). Uh and um revisited : Are they interjections for signaling delay ? *Journal of Psycholinguistic Research*, 34:555–576.

PALLAUD, B. (1999). Lapsus et phénomènes voisins dans la langue parlée : problème d'identification. *Recherches sur le français parlé*, 15:9–40.

PALLAUD, B. et HENRY, S. (2004). Amorces de mots et répétitions : des hésitations plus que des erreurs en français parlé. *In Le poids des mots. Actes des 7èmes Journées Internationales d'Analyse statistique des Données Textuelles*, pages 848–858.

PAROUBEK, P. (2000). Language resources as by-product of evaluation : the multitag example. *In Proceedings of LREC*, pages 151–154.

PESHKOV, K., PRÉVOT, L., RAUZY, S. et PALLAUD, B. (2013). Categorizing syntactic chunks for marking disfluent speech in french language. *In* EKLUND, R., éditeur : *Proceedings of the 6th Workshop on Disfluency in Spontaneous Speech (DiSS'13)*, pages 59–62, Stockholm, Sweden. KTH Royal Institute of Technology.

PIU, M. et BOVE, R. (2007). Annotation des disfluences dans les corpus oraux. *In Actes des 9e Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL'07)*, Toulouse, France.

QIAN, X. et LIU, Y. (2013). Disfluency Detection Using Multi-step Stacked Learning. *In Proceedings of HLT/NAACL*, pages 820–825.

RASOOLI, M. S. et TETREAULT, J. (2013). Joint Parsing and Disfluency Detection in Linear Time. *In Proceedings of EMNLP*, pages 124–129.

RATNAPARKHI, A. (1996). A maximum entropy model for part-of-speech tagging.

RAYMOND, C. (2011). discretize4crf. https://gforge.inria.fr/projects/discretize4crf/, consulté le 16/01/2015.

ROSSET, S., ILLOUZ, G. et MAX, A. (2005). Interaction et recherche d'information : le projet ritel. *Traitement Automatique des Langues*, 46(3):155–179.

ROSSET, S. et PETEL, S. (2006). The ritel corpus – an annotated human machine open-domain question answering spoken dialog corpus. *In Proceedings of the 5th international Conference on Language Resources and Evaluation (LREC'06)*.

ROSSET, S., TRIBOUT, D. et LAMEL, L. (2008). Multi-level information and automatic dialog act detection in human-human spoken dialogs. *Speech Communication*.

SANG, E. F. T. K. et DE MEULDER, F. (2003). Introduction to the CoNLL-2003 Shared Task : Language-Independent Named Entity Recognition. *In Proceedings of CoNLL*, pages 142–147.

SCHMID, H. (1994). Probabilistic part-of-speech tagging using decision trees. *In Proceedings of the International Conference on New Methods in Language Processing (NeMLaP'94)*.

SHRIBERG, E. E. (1994). *Preliminaries to a Theory of Speech Disfluencies*. Thèse de doctorat, Berkeley University of California.

SHRIBERG, E. E. (1999). Phonetic consequences of speech disfluency. *In Proceedings of the 14th International Congress of Phonetic Sciences (ICPhS'99)*, pages 619–622, San Francisco, USA.

SHRIBERG, E. E. et STOLCKE, A. (2002). Prosody modeling for automatic speech recognition and understanding. *In Proceedings of the Workshop on Mathematical Foundations of Natural Language Modeling*.

SNOVER, M., DORR, B. et SCHWARTZ, R. (2004). A Lexically-Driven Algorithm for Disfluency Detection. *In Proceedings of HLT-NAACL*.

STRASSEL, S. (2004). *Simple Metadata Annotation Specification V6.2*. Linguistic Data Consortium, 6.2 édition.

SWERTS, M., WICHMANN, A. et BEUN, R.-J. (1998). Filled pauses as markers of discourse structure. *Journal of Pragmatics*, 30:485–496.

TASKAR, B., GUESTRIN, C. et KOLLER, D. (2004). Max-margin markov networks. *In* THRUN, S., SAUL, L. et SCHÖLKOPF, B., éditeurs : *Advances in Neural Information Processing Systems 16*, pages 25–32. MIT Press.

TORREIRA, F., ADDA-DECKER, M. et ERNESTUS, M. (2010). The nijmegen corpus of casual french. *Speech Communication*, 52:201–221.

VAN SCHOOTEN, B., ROSSET, S., GALIBERT, O., MAX, A., op den AKKER, R. et ILLOUZ, G. (2007). Handling speech input in the ritel qa dialogue system. *In Proceedings of the 8th International Conference on Speech Communication and Technology (Interspeech'07)*, Antwerp, Belgique.

VASILESCU, I., ADDA-DECKER, M. et NEMOTO, R. (2008). Caractéristiques acoustiques et prosodiques des hésitations vocaliques dans trois langues. *Traitement Automatique des Langues*, 49(3):199–228.

VASILESCU, I., ROSSET, S. et ADDA-DECKER, M. (2010a). On the functions of the vocalic hesitation euh in interactive man-machine question answering dialogs in french. *In Proceedings of the 5th Workshop on Disfluency in Spontaneous Speech – The 2nd International Symposium on Linguistic Patterns in Spontaneous Speech (DiSS-LPSS'10)*, Tokyo, Japan.

VASILESCU, I., ROSSET, S. et ADDA-DECKER, M. (2010b). On the role of discourse markers in interactive spoken question answering systems. *In Proceedings of the 7th international Conference on Language Resources and Evaluation (LREC'10)*, pages 2450–2456, Valletta, Malta. European Language Resources Association (ELRA).

VAUDABLE, C. (2012). *Analyse et reconnaissance des émotions lors de conversations de centres d'appels*. Thèse de doctorat, Université Paris-Sud XI.

YOSHIDA, E. et LICKLEY, R. J. (2010). Disfluency patterns in dialogue processing. *In Proceedings of the 5th Workshop on Disfluency in Spontaneous Speech – The 2nd International Symposium on Linguistic Patterns in Spontaneous Speech (DiSS-LPSS'10)*, pages 115–118.

ZWARTS, S. et JOHNSON, M. (2011). The Impact of Language Models and Loss Functions on Repair Disfluency Detection. *In Proceedings of ACL*, pages 703–711.

Appendices

Annexe A

Jeux d'étiquettes des étiqueteurs morpho-syntaxiques

Étiquette	**Description**
ADJ	adjectif
ADJWH	adjectif interrogatif
ADV	adverbe
ADVWH	adverbe interrogatif
CC	conjonction de coordination
CLO	pronom clitique objet
CLR	pronom clitique réflexif
CLS	pronom clitique sujet
CS	conjonction de subordination
DET	déterminant
DETWH	déterminant interrogatif
ET	mot étranger
I	interjection
NC	nom commun
NPP	nom propre
P	préposition
P+D	amalgame préposition et déterminant
P+PRO	amalgame préposition et pronom
PONCT	marque de ponctuation
PREF	préfixe
PRO	pronom complet
PROREL	pronom relatif
PROWH	pronom interrogatif
V	verbe à l'indicatif ou au conditionnel
VIMP	verbe à l'impératif
VINF	verbe à l'infinitif
VPP	participe passé
VPR	participe présent
VS	verbe au subjonctif

TABLEAU A.1 – Jeu d'étiquette de l'étiqueteur morpho-syntaxique `MElt`

Étiquette	Description
ABR	abréviation
ADJ	adjectif
ADV	adverbe
DET :ART	article
DET :POS	pronom possessif
INT	interjection
KON	conjonction
NAM	nom propre
NOM	nom
NUM	chiffre
PRO	pronom
PRO :DEM	pronom démonstratif
PRO :IND	pronom indéfini
PRO :PER	pronom personnel
PRO :POS	pronom possessif
PRO :REL	pronom relatif
PRP	préposition
PRP :det	contraction préposition et article
PUN	ponctuation
PUN :cit	ponctuation de citation
SENT	marque de fin de phrase
SYM	symbole
VER :cond	verbe au conditionnel
VER :futu	verbe au futur
VER :impe	verbe à l'impératif
VER :impf	verbe à l'imparfait
VER :infi	verbe à l'infinitif
VER :pper	verbe au participe passé
VER :ppre	verbe au participe présent
VER :pres	verbe au présent
VER :simp	verbe au passé simple
VER :subi	verbe au subjonctif imparfait
VER :subp	verbe au subjonctif présent

TABLEAU A.2 – Jeu d'étiquette de l'étiqueteur morpho-syntaxique `Treetagger`