Mots composés et disfluences

Matthieu Constant[1], Anne Dister[2]

[1] Université Paris-Est – Matthieu.Constant@univ-mlv.fr
[2]Facultés universitaires Saint-Louis – dister@fusl.ac.be

Abstract

Disfluencies, phenomenon specific to spoken texts, have the characteristic of breaking the syntactic linearity of speech. Multiword lexical units tend to form semantic units. In this paper, we show, through an automatic procedure, that multiword lexical units are less likely to contain a disfluency than a free sequence of words. The procedure consists first in locating disfluencies with a resource-free iterative algorithm, and then in recognizing multiword units thanks to a probabilistic model.

Résumé

Les disfluences, phénomène propre à l'oral, ont la particularité de briser la linéarité syntaxique de l'énoncé. Les mots composés ont tendance à former des unités syntaxiques et sémantiques. Dans cet article, nous montrons que l'énonciation de telles expressions dans un discours oral est moins propice à l'apparition de disfluences qu'une séquence libre de mots. Pour cela, nous avons mis au point une procédure automatique de reconnaissance probabiliste des mots composés incluant une détection itérative préalable des disfluences.

Mots-clés : disfluences, mots composés, étiquetage morphosyntaxique, segmentation.

1. Introduction

Les disfluences, phénomène propre à l'oral, ont la particularité de briser la linéarité syntaxique de l'énoncé dans lequel elles apparaissent. Elles constituent une interruption (souvent momentanée, parfois définitive) dans le déroulement de l'énoncé. Les expressions polylexicales telles que les mots composés forment des unités syntaxiques et sémantiques. Du fait de cette double propriété, l'énonciation de telles expressions dans un discours oral nous parait moins propice à l'apparition de disfluences qu'une séquence libre de mots. C'est ce que nous allons vérifier dans cet article.

Afin de vérifier notre hypothèse, nous avons utilisé la procédure suivante. Nous sommes partis d'un corpus de transcriptions orales formé de près de 500 000 mots graphiques (Dister, 2007). Nous avons repéré les disfluences à l'aide d'un algorithme itératif, implanté dans l'outil *Distagger* (Constant et Dister, 2010). Les disfluences ont alors été supprimées du texte afin de rendre à ce dernier sa linéarité syntaxique. Nous avons ensuite balisé les mots composés au moyen du segmenteur-étiqueteur *lgtagger* basé sur le modèle probabiliste des champs aléatoires markoviens et sur un lexique morphosyntaxique à large couverture (Constant et Sigogne, 2011).

Les disfluences repérées initialement ont alors été réinsérées dans le texte balisé afin de procéder à divers calculs statistiques nous permettant de vérifier notre hypothèse de départ.

Cet article se divise en 7 parties distinctes. Dans un premier temps, nous décrivons notre corpus et précisons la notion de disfluence (section 2). Nous présentons ensuite l'algorithme de repérage des disfluences, ainsi qu'une évaluation de l'outil utilisé sur notre corpus (section 3). Nous discutons la notion de mot composé (section 4), puis nous détaillons et évaluons la procédure automatique de reconnaissance des mots composés (section 5). Enfin, nous expliquons les divers calculs statistiques mis en place afin de vérifier notre hypothèse (section 6). La section 7 est consacrée aux conclusions et perspectives.

2. Transcriptions orales et disfluences

Les données sur lesquelles nous travaillons sont issues de la banque de données VALIBEL (acronyme pour *Variétés Linguistiques du français en Belgique*). Les transcriptions suivent des conventions explicites (Dister *et al.*, 2006), qui convergent largement avec celles adoptées dans d'autres projets (cf. *Corpus de référence du français parlé* de l'équipe DELIC, les données du projet Rhapsodie, pour ne citer qu'eux). Celles-ci suivent trois grands principes : adoption de l'orthographe standard, non-recours à la ponctuation de l'écrit (et donc pas de découpage du texte en phrases) et « valorisation » de l'oralité des données. Par là, nous voulons dire que sont transcrits un certain nombre de traits propres à la production de la langue parlée, d'« achoppements » dans la linéarité de l'énoncé, de marques du discours en cours d'élaboration ; ces phénomènes sont inhérents aux productions orales (même les plus surveillées) et on les regroupe souvent sous l'appellation générale de *disfluences*[1]. Ce sont ces phénomènes qui brisent la linéarité et entravent bien souvent le bon fonctionnement des analyseurs (morpho)syntaxiques. Dans l'exemple suivant, la locutrice répare son énoncé, en répétant une construction amorcée (*qui s'en/*) et en la complétant (*qui s'enfonce*)

> **ilePA2** or une trémie euh grammaticalement c'est une chose **qui s'en/** qui s'enfonce plutôt dans la terre [ilePA2r] [2]

Le terme *disfluence* regroupe donc différents types de réalisations qui ont toutes cette particularité qu'à un moment donné de l'énoncé, on constate un piétinement sur un même point de l'axe syntagmatique (pour une modélisation des séquences disfluences, voir Shriberg, 1994). Dans cet article, nous nous concentrons en particulier sur 4 types de disfluences fréquentes et relativement faciles à repérer automatiquement. Outre le *euh*, nous avons traité la répétition, l'autocorrection immédiate et l'amorce de morphème. Nous entendons par **répétition** la reprise à l'identique, dans le contexte direct ou après une séquence de mots spécifiques (onomatopées, *oui*, etc.), d'un mot ou d'un groupe de mots, comme c'est le cas de *sans* et de *la* dans l'exemple suivant :

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

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

2 Une barre oblique indique une pause brève, une double barre oblique une pause longue.

Les **autocorrections immédiates** (Dister, 2008) constituent une variante de la répétition. Dans les autocorrections, l'un des traits morphologiques de l'élément répété varie, comme l'illustre l'exemple suivant où le déterminant défini est répété, en changeant le trait 'singulier' par le trait 'pluriel' :

> **ileFN1** et le journalisme et puis euh **le** les études de journalisme en soi ne me plaisaient pas [ileFN1r]

Nous traitons également l'**amorce**, phénomène langagier qui consiste en « une interruption de morphèmes en cours d'énonciation » (Pallaud, 2002 : 79). L'exemple suivant est un cas typique d'amorce. Le morphème interrompu – symbolisé par une barre oblique collée directement à la droite de celui-ci, à l'endroit de l'interruption – est corrigé plus loin dans l'énoncé, où il est repris sous sa forme pleine :

> **ilrPC1** (…) j'aimerais bien moi ouvrir un **ma/** un petit magasin (…) [ilrPC1r]

Ces marques peuvent se présenter seules, mais également imbriquées les unes dans les autres, comme l'illustre l'exemple suivant, où s'entremêlent autocorrection, répétition et amorce de morphème :

> norHJ1 (…) mais bon on je je j/ je prends des décisions (…) [ilrPC1r]

3. Reconnaissance automatique des disfluences

L'idée sur laquelle se fonde notre système est de repérer et de classer les séquences disfluentes, pour ne conserver dans le texte que la séquence réparée afin de les soumettre à un analyseur linguistique de données plus standard. La classe et les positions initiales des séquences disfluentes sont sauvegardées afin de pouvoir les réutiliser et les réinsérer dans le texte une fois que ce dernier a été analysé. Ainsi, après passage dans notre système, l'exemple suivant

> **ilePA2** or une trémie **euh** grammaticalement c'est une chose **qui s'en/** qui s'enfonce plutôt dans la terre

devient

> **ilePA2** or une trémie grammaticalement c'est une chose qui s'enfonce plutôt dans la terre

Les positions initiales de *euh* (type euh) et *qui s'en/* (type amorce) sont sauvegardées afin de les réintégrer par la suite.

Une disfluence simple est relativement facile à repérer à l'aide de patrons simples. Cependant, les combinaisons de plusieurs disfluences entrelacées les unes dans les autres complexifient la tâche. Une solution est d'écrire différents patrons correspondant à chacune des combinaisons possibles, avec l'inconvénient de devoir écrire de multiples règles. Une autre, celle que nous avons choisie, consiste à n'appliquer qu'un seul patron de manière itérative jusqu'à obtenir un point fixe. Le patron appliqué est composé de trois parties : une séquence *w* de mots, une séquence d'insertions *I* (mots d'éditions, pauses silencieuses, …) et une séquence *c* de mots (potentiellement vide) qui correspond à une correction de *w*. Chaque mot de *c* est une correction

du mot correspondant de *w* (quand il en a un[3]) : soit le mot lui-même, soit un mot appartenant à sa classe d'équivalence définie par l'utilisateur (ex. {*le*, *la*, *les*}), soit un mot dont *w* est le préfixe dans le cas d'une amorce (ex. *j'* pour *je*). Lors de l'application de ce patron à une position donnée du texte, s'il correspond, la séquence *wI* est supprimée pour ne conserver que la correction *c*. L'algorithme consiste à faire plusieurs applications glissantes du patron de reconnaissance sur le texte jusqu'à ce que ce dernier ne soit plus modifié. Soit l'exemple « *le le chien euh les le chien dort* ». La première application glissante du patron donne successivement :

le le chien euh les le chien dort

le le chien euh les le chien dort

le chien euh les le chien dort

le chien les le chien dort

le chien le chien dort

La deuxième application du patron produit :

le chien le chien dort

le chien le chien dort

le chien dort

La troisième application ne modifie plus le texte en entrée. La procédure s'arrête donc. Une fois toutes les séquences disfluentes repérées, celles-ci sont typées à l'aide de règles simples utilisant le type de corrections détectées. L'ensemble de la procédure est implanté dans l'outil *Distagger* (Constant et Dister, 2010).

L'évaluation de notre outil de détection des disfluences a consisté à l'appliquer sur deux transcriptions[4] extraites de nos données, qui ont la caractéristique d'avoir des locuteurs dont le taux de disfluences est supérieur aux autres (Dister, 2007). Le résultat de cette application a été confronté à l'annotation de référence validée manuellement[5]. Ce corpus de référence comprend au total 1297 tours de parole, 22476 mots graphiques, 5817 méta-étiquettes et 1280 disfluences. Les séquences « à supprimer » comptent 1945 mots. La distribution des types de disfluences est indiquée dans la ligne #disfl du Tableau 1.

Nous avons mesuré la qualité de l'annotation des disfluences en calculant précision, rappel et f-score : la précision (p) est la proportion de disfluences bien repérées (et typées) parmi l'ensemble des disfluences automatiquement repérées ; le rappel (r) est la proportion de disfluences bien repérées (et typées) parmi l'ensemble des disfluences du corpus de référence ; le f-score (f) est la moyenne pondérée de la précision et du rappel. Nous avons également

3 Les amorces n'ont pas forcément de mot correspondant. Le mot *euh* n'a pas de mot correspondant.

4 L'outil a été développé et testé à l'aide d'un corpus de développement de 4 transcriptions comportant 34 624 mots, disjoint du corpus d'évaluation.

5 Le corpus de référence a été construit semi-automatiquement. L'outil de prétraitement a d'abord été appliqué. Le résultat a alors été vérifié manuellement. Les erreurs ont été corrigées et les disfluences manquantes annotées.

calculé ces mêmes mesures pour chaque type de disfluences. Les résultats sont synthétisés dans le Tableau 1.

Dans l'analyse des résultats, nous n'avons pas fait de distinction entre les répétitions disfluentes et les répétitions intensives, afin d'avoir un traitement unifié de ce phénomène qui relève dans tous les cas d'un entassement paradigmatique. Par contre, nous avons considéré comme une erreur le fait de reconnaitre comme une répétition une cooccurrence de forme exigée par la syntaxe, comme dans l'exemple suivant où les deux occurrences de *Martinet* assument des fonctions syntaxiques clairement différentes :

> **norHJ1** et qui vient de la part de Martinet // Martinet m'en est très reconnaissant d'ailleurs [norHJ1r]

Les erreurs de reconnaissance ont différentes causes, que nous ne pouvons détailler faute de place. Certaines ne sont pas possibles à corriger, comme c'est le cas dans l'exemple précédent, ou dans le suivant où nous manquons d'indice nous permettant de repérer automatiquement que nous ne sommes pas dans le cas d'une amorce de morphème complétée directement, ce que le format de la disfluence peut faire penser, avec un mot qui commence par a- (*aucun*) et qui suit immédiatement une amorce de morphème :

> **norHJ1** (…) euh eh bien ces Noirs donc qui étaient là euh (xxx) et qui qui // qui affirmaient le / leur attachement au français // il y a aucun qui l'**a/** aucun ne rêvait de s' en affranchir // et à la fin de la biennale je le vois encore ils sont venus vers moi à plusieurs en me disant / monsieur le président // ne nous abandonnez pas [norHJ1r]

Nous pourrions aussi corriger notre système en y ajoutant des règles, notamment pour le repérage de séquences comme la suivante, actuellement non reconnue comme une autocorrection avec un changement de déterminant après une préposition répétée, mais comme une simple amorce de morphème non corrigée :

> **norFA0** mais oui oui mais ce que je reprocherais un petit peu à à Walter et à Martinet c'est que dans leurs **in/** dans le choix de leurs informateurs / ils ont pris dix-sept personnes qui étaient parisiennes [norHJ1r]

	TOUS (non typé)	TOUS (typé)	amorce	correction	euh	répétition	mélange
#disfl (%)	1280 (100%)	1280 (100%)	158 (12%)	48 (4%)	194 (15%)	727 (57%)	153 (12%)
p	95,3	94.9	87	91	93	97	98
r	95,8	95.4	94	60	95	98	92
f	95,5	95.2	90	73	94	98	95

Tableau 1 : Evaluation de la détection automatique de disfluences

4. Les mots composés

Les expressions à mots multiples regroupent un grand nombre de phénomènes linguistiques, pour lesquels la terminologie n'est pas unifiée et dont les propriétés syntaxiques et sémantiques

ne se recoupent que partiellement : les expressions figées et semi-figées, les collocations, les entités nommées, les verbes à particule, les constructions à verbe support, les mots composés, les termes, etc. (Sag *et al.*, 2002). Elles ont la caractéristique d'avoir un comportement particulier d'un point de vue lexical, syntaxique, sémantique, pragmatique ou/et statistique. Dans cet exposé, nous nous intéressons aux mots composés, et en particulier ceux répertoriés dans deux corpus annotés : le corpus arboré de Paris 7 formé d'articles journalistiques (Abeillé *et al.*, 2003), ainsi que le corpus d'Orléans, corpus oral annoté en parties du discours (Eshkol *et al.*, 2010). Les mots composés annotés sont, dans les deux corpus, des séquences contigües[6] de mots avec de fortes contraintes lexicales, syntaxiques et/ou sémantiques. Ils comportent tous un certain degré de non-compositionnalité. Par exemple, les sens de l'adverbe temporel *tout de suite* et de la conjonction de subordination *bien que* ne peuvent pas être déduits du sens de leurs composants internes simples. Le nom *femme au foyer* a un certain degré de compositionnalité sémantique car cette séquence identifie une femme. Cette femme a cependant une caractéristique qui est très difficilement calculable à partir des composants *au* et *foyer* : elle ne travaille pas. Cette notion de non-compositionnalité n'a pas de définition stricte et forme un continuum. Afin de rendre reproductible leurs annotations dans nos corpus de travail, leurs auteurs de ces corpus ont donc utilisé des critères syntaxiques comme le figement lexical ou structurel du mot composé. Par exemple, *à cause de* est une préposition composée car sa structure de surface est figée : **à la cause de*. La séquence *femme au foyer* est un nom composé car elle n'admet pas de variations lexicales telles que **femme à domicile* ou **femme à la maison*. Les mots composés appartiennent aux différentes parties du discours : les noms communs (*bonne sœur, tire-bouchon*), les noms propres (*Electricité de France, André Salem*), les prépositions (*à cause de, à part*), les conjonctions (*parce que*), les adverbes (*en effet*, *à peu près*), les déterminants (*beaucoup de*) …

5. Reconnaissance des unités polylexicales

La reconnaissance d'unités polylexicales est un vaste sujet d'étude. De nombreuses voies ont été explorées : extraction automatique par analyse syntaxique (partielle ou profonde) ou/et filtrage statistique au moyen de mesures associatives (voir, entre autres, Seretan *et al.*, 2004 ; Ramisch *et al.*, 2010 ; Watrin et François, 2011), par apprentissage d'un modèle probabiliste sur corpus annoté (Constant *et al.*, 2011 ; Green *et al.*, 2011), par alignement bilingue (Caseli *et al.*, 2009), etc.

Nous avons décidé de nous baser sur une approche par apprentissage d'un modèle CRF (champs aléatoires markoviens) sur un corpus annoté en parties du discours couplé à un lexique morphosyntaxique à large couverture (Constant *et al.*, 2011). Les auteurs rapportent d'excellents résultats sur l'écrit. Des expériences ont également montré de bonnes performances sur l'oral (Eshkol *et al.*, 2010). Le modèle appris combine segmentation en mots composés et étiquetage morphosyntaxique. Il intègre différents traits provenant des propriétés des mots du corpus (valeur, suffixes, préfixes, présence d'une majuscule, etc.), les étiquettes des mots trouvées dans un lexique externe (incorporant des mots composés), leurs contextes lexicaux (ex. bi-grammes de mots), leurs contextes grammaticaux (bi-grammes d'étiquettes). Nous utilisons l'outil *lgtagger* librement disponible (Constant et Sigogne 2011). Les ressources lexicales et les

6 Les mots composés acceptent très rarement des insertions, le plus souvent des modifieurs tels que les adverbes.

patrons de traits y sont définis par l'utilisateur dans un fichier de configuration. Le modèle CRF est appris au moyen du logiciel Wapiti (Lavergne *et al.*, 2010) et les ressources lexicales sont appliquées à l'aide du logiciel Unitex (Paumier, 2011). Cet outil a la caractéristique d'avoir une bonne précision de détection des mots composés (environ 90 %) et un silence de l'ordre de 30 % sur un texte écrit (sur le corpus arboré de Paris 7 (Abeillé *et al.* 2003)).

Pour notre tâche, nous avons associé à l'outil les dictionnaires de langue générale DELA (Courtois, 1990 ; Courtois *et al.*, 1997) et Lefff (Sagot, 2010), ainsi que le dictionnaire des toponymes Prolex (Piton *et al.*, 1999). Les patrons de traits sont les mêmes que ceux utilisés dans Constant et Sigogne (2011). Le corpus d'apprentissage est formé de deux corpus : une partie du corpus arboré de Paris 7 (Abeillé, 2003) transformé automatiquement en corpus annoté en parties du discours (455 264 mots) ; une partie du corpus annoté de transcriptions orales qu'ont utilisé Eshkol *et al.* (2010) pour des expériences d'étiquetage morphosyntaxique (40 559 mots). Le jeu d'étiquettes a été modifié pour les deux corpus afin d'être conforme à celui décrit dans (Candito et Crabbé, 2009). Les mots composés sont marqués dans les deux corpus. Le corpus journalistique comporte 5,6 % de mots composés (parmi tous les mots) ; le corpus oral en comporte 6,5 %.

Nous avons évalué la précision du segmenteur-étiqueteur appris dans cette configuration sur les 30 000 premiers mots graphiques de notre corpus de travail où les disfluences ont été préalablement supprimées. Nous avons manuellement validé le corpus d'évaluation annoté automatiquement par *lgtagger*. Ce corpus comportait 1 000 mots composés, dont 776 appartenant à nos ressources de mots composés (soit 77,6 %). Nous avons observé un taux de précision de 91 % dans la détection des mots composés. Lorsque le mot composé appartient à nos ressources lexicales, ce taux grimpe à 98 %. Lorsque c'est une unité inconnue de nos ressources lexicales, c'est-à-dire devinée par l'outil, ce taux baisse dramatiquement à 68 %. Dans ce cas-là, nous observons quelques phénomènes très intéressants. L'outil a tendance à reconnaitre des expressions avec des structures syntaxiques moins fréquentes. En particulier, il repère correctement de nombreuses structures verbales figées comme *couler de source* ou *il y a eu*. Or, malgré l'élimination des disfluences dans le corpus par notre prétraitement avec *Distagger*, il subsiste encore des structures syntaxiques inachevées et/ou non linéaires, et qui peuvent passer pour des structures non standards donc potentiellement polylexicales. Dans l'exemple ci-dessous, la séquence *étais au avant* est reconnue car c'est une structure particulière qui est causée par l'amorce d'une construction (*j'étais au*) non achevée et non marquée dans la transcription[7].

> **ilrLD3** et je suis passé de deuxième rénové je suis passé en troisième professionnelle / et comme à l'école où ce que j'**étais au avant** parce que je suis je ne suis pas ici depuis longtemps ça fait que deux ans [ilrLD3r]

L'erreur ci-dessous (séquence *serait ce serait*) correspond en fait à une autocorrection non repérée (*ça ça serait → ce serait ça*).

> **ilrDT1** bè c'est pf / il y en a à Marche / il y en a à Bastogne qui parlent bien // oui ça ça **serait ce serait** ça // où je où où le le le niveau [ilrDT1r]

7 La transcription ne donne une indication que pour les amorces de morphèmes (notées avec une barre oblique à la fin du mot amorcé), pas pour les amorces de constructions syntaxiques.

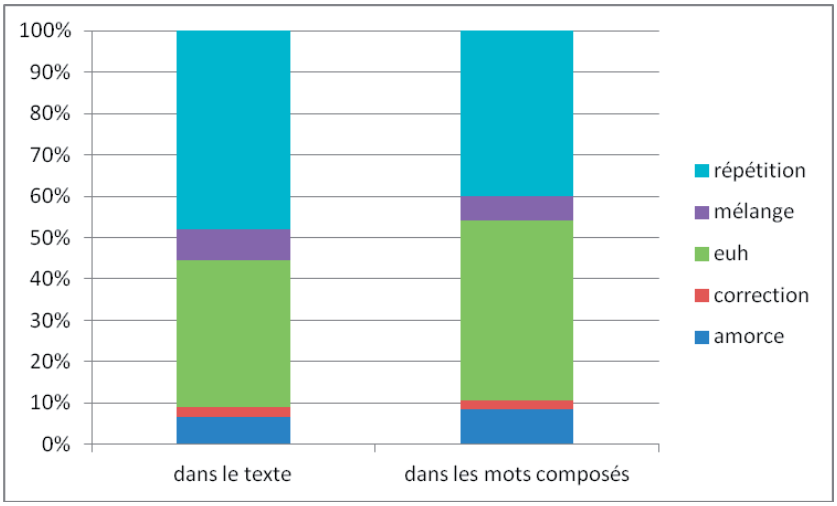
Par ailleurs, les onomatopées ou les mots qui reçoivent une analyse particulière dans les énoncés oraux (*quoi* qui fonctionne comme particule discursive, p. ex.) sont sources d'erreurs comme dans les séquences suivantes erronément reconnues : *bénéfice quoi*, *ben pf* ou *mm eh bé*. L'outil reconnait également de nombreuses séquences *adverbe composé* + *adjectif* (ex. *tout à fait parfait*) qui ne forment pas une expression multi-mots à proprement parler, mais plutôt un chunk syntaxique. Rappelons qu'un chunk est un constituant syntaxique non récursif simple.

6. Analyse statistique

Bien que notre procédure automatique de détection des disfluences ne soit pas parfaite, nous utilisons telles quelles les annotations résultantes, car les quelques erreurs et silences observés nous semblent répartis de manière homogène. Par ailleurs, étant donné la très mauvaise précision de reconnaissance des mots composés inconnus de nos ressources lexicales, nous ne tenons compte que des mots composés reconnus et présents dans celles-ci, afin de ne pas fausser les calculs.

Notre corpus de travail comprend 478 084 unités dont 22 205 unités disfluentes. Une unité est soit une séquence disfluente soit un mot graphique. On constate ainsi que 4,6 % des unités du corpus sont disfluentes. Par ailleurs, le corpus comprend 15 350 mots composés appartenant à nos ressources, ce qui correspond à 38 400 unités. Par contre, seules 2,7 % des unités des mots composés sont des séquences disfluentes (soit 40 % de moins que la distribution sur le texte entier). Ce résultat confirme notre hypothèse de départ de manière claire. Nous constatons également qu'un mot composé avec au moins une disfluence comporte en moyenne un petit peu plus d'une disfluence. Par ailleurs, nous avons examiné les positions des disfluences dans les mots composés. Plus précisément, nous avons regardé deux positions particulières. Les disfluences qui se trouvent en position initiale du mot composé : ex. *ca/ carte bancaire*. Les disfluences en position interne : *carte euh bancaire*. Nous observons qu'elles ont tendance à se trouver en position initiale (89 % d'entre elles contre 11 % en position interne). Ceci revient à dire qu'une fois le mot composé bien amorcé, son énonciation tend à être linéaire.

Nous avons ensuite analysé la distribution des disfluences selon leur type dans le cas général et à l'intérieur des mots composés (cf. Graphique 1). On constate une augmentation de la proportion de *euh* et d'amorces, ainsi qu'une diminution des répétitions et corrections dans les mots composés.

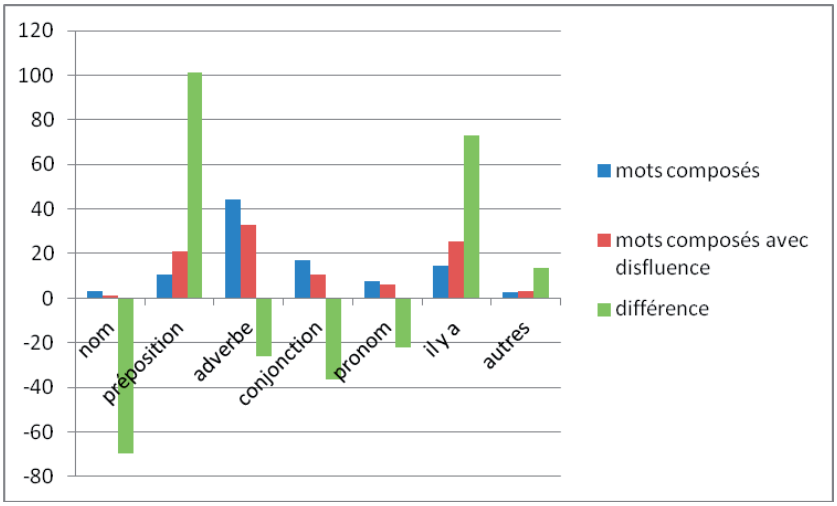


Graphique 1 : Distribution des disfluences selon leur type

Ensuite, nous avons étudié la distribution des mots composés selon leur catégorie grammaticale (cf. Tableau 2 illustré par le Graphique 2). Nous avons en particulier comparé la distribution générale des mots composés avec ceux possédant au moins une disfluence. Nous avons calculé leur différence relative de distribution pour chaque catégorie grammaticale. On observe que les prépositions et la locution *il y a* sont particulièrement sujettes aux disfluences. Ceci peut peut-être s'expliquer par le fait qu'elles jouent le plus souvent le rôle d'introducteur de chunk nominal et se présentent donc dans l'énoncé à un moment où le locuteur est à la recherche de la dénomination (Blanche-Benveniste, 1984). Par contre, les adverbes, les noms et les conjonctions ont tendance à être moins propice à une disfluence.

Taille	nom	préposition	adverbe	conjonction	pronom	il y a	Autres
mots composés	3,3	10,4	**44,1**	16,9	7,7	14,7	2,9
mots composés avec disfluence	1	20,9	**32,7**	10,7	6	25,4	3,3
Différence relative	-69,7	**101**	-25,9	-36,7	-22,1	72,8	13,8

Tableau 2 : Distribution des mots composés

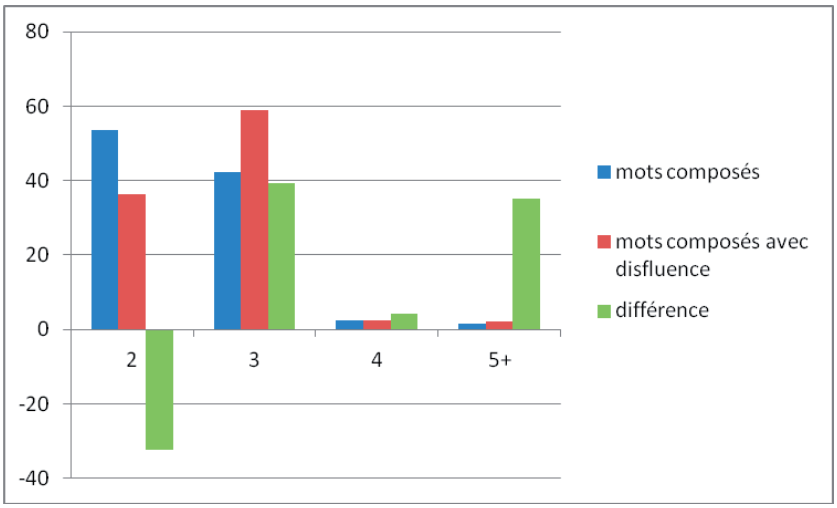


Graphique 2 : Distribution des mots composés selon leur catégorie grammaticale

Si l'on regarde maintenant la distribution des mots composés selon leur taille (en nombre de mots graphiques) et que l'on compare leur comportement général par rapport à celui des mots composés avec disfluence (cf. Tableau 3 illustré par le Graphique 3), on constate que les petits mots composés (taille 2) sont moins propices à l'apparition de disfluences que les mots composés plus longs. Ceci s'explique aisément par le fait que plus une séquence est longue, plus elle donne prise au piétinement syntaxique.

longueur	2	3	4	5+
mots composés	53,7	42,2	2,4	1,7
mots composés avec disfluence	36,4	58,8	2,5	2,3
différence relative	-32,2	39,3	4,2	35,3

Tableau 3 : Distribution des mots composés selon leur longueur



Graphique 3 : Distribution des mots composés selon leur longueur

7. Conclusions et perspectives

Nous avons présenté une procédure automatique permettant de reconnaitre les mots composés dans un corpus de transcriptions de français parlé, tout en détectant quatre types de disfluences. Les outils utilisés ont permis de mettre clairement en évidence que les disfluences apparaissent moins fréquemment dans les mots composés que dans les autres parties du texte. Ceci s'explique par le fait qu'un mot composé forme une unité syntaxique et sémantique, et est donc, de ce fait, moins propice à toute interruption.

Cette recherche demanderait à être améliorée en augmentant la couverture des mots composés reconnus, ainsi qu'en affinant davantage la procédure de reconnaissance des disfluences.

Dans de futures études, nous aimerions étendre l'analyse faite ici aux chunks, partant de la même hypothèse qu'une séquence disfluente apparaitra préférentiellement entre des chunks qu'à l'intérieur de ceux-ci.

Remerciements

Nous tenons à remercier Iris Eshkol pour nous avoir mis à disposition son corpus oral annoté.

Bibliographie

Abeillé A., Clément L. et Toussenel F. (2003). Building a treebank for French, in Abeillé A. editor. *Treebanks*. Kluwer, Dordrecht.

Blanche-Benveniste Cl. (1985). La dénomination dans le français parlé : une interprétation pour les "répétitions" et les "hésitations". *Recherches sur le français parlé*, vol. (6) : 109-130.

Caseli H., Ramisch C., Nunes M. et Villavicencio A. (2009). Alignment-based extraction of multiword expressions. *Language resources and evaluation*. Springer.

Constant M. et Dister A. (2010). Automatic detection of disfluencies in speech transcriptions. *Spoken Communication*. In Pettorino M., Giannini A., Chiari I., Dovetto F. M. editors. Cambridge Scholars Publishing :259-272.

Constant M. et Sigogne A. (2011). MWU-aware Part-of-Speech Tagging with a CRF model and lexical resources. *Proc. of MWE'11 (ACL Workshop on Multiword Expressions: from Parsing and Generation to the Real World)*.

Constant M., Tellier I., Duchier D., Dupont Y., Sigogne A. et S. Billot. (2011) Intégrer des connaissances linguistiques dans un CRF : application à l'apprentissage d'un segmenteur-étiqueteur du français. *Actes de TALN'11 (Conférence sur le traitement automatique des langues naturelles)*.

Courtois B. (1990). Un système de dictionnaires électroniques pour les mots simples du français. *Langue Française*, vol. (87).

Courtois B., Garrigues M., Gross G., Gross M., Jung R., Mathieu-Colas M., Monceaux A., Poncet-Montange A., Silberztein M. et Vivès R. (1997). Dictionnaires électronique DELAC : les mots composés binaires. Rapport technique 56, LADL, Université Paris 7.

Candito M.-H. et Crabbé B. (2009). Improving generative statistical parsing with semi-supervised word clustering. Proc. of IWPT'09 (11th International Conference on Parsing Technologies), Paris, France.

Delic (2004). Présentation du *Corpus de Référence du Français Parlé*. *Recherches sur le français parlé*, vol. (18) : 11-42.

Dister A., Francard M., Geron G., Giroul V., Hambye Ph., Simon A. C. et Wilmet R. (2006). *Conventions de transcription régissant les corpus de la banque de données VALIBEL* (http://valibel.fltr.ucl.ac.be, corpus oraux, conventions de transcription).

Dister A. (2007). *De la transcription à l'étiquetage morphosyntaxique. Le cas de la banque de données textuelles orales Valibel*. Thèse de doctorat, Université de Louvain.

Dister A. (2008). L'autocorrection immédiate en français parlé : le cas des déterminants. *Actes des JADT 2008 (5es Journées internationales d'Analyse statistique des Données Textuelles)*. Lyon : Presses universitaires de Lyon.

Eshkol I., Tellier I., Taalab S. et Billot S. (2010), Étiqueter un corpus oral par apprentissage automatique à l'aide de connaissances linguistiques. *Actes des JADT'10* (*10th International Conference on statistical analysis of textual data)*.

Green S., de Marneffe M.-C., Bauer J. et Manning C. D. (2010). Multiword Expression Identification with Tree Substitution Grammars: A Parsing *tour de force* with French. Proc. of *EMNLP'11 (Conference on Empirical Methods in Natural Language Processing)*.

Lavergne T., Cappé O. et Yvon Fr. (2010). Practical Very Large Scale CRFs. Proc. of *ACL'10 (48 Annual Meeting of the Association for Computational Linguistics)*.

Pallaud B. (2002). Les amorces de mots comme faits autonymiques en langage oral. *Recherches sur le français parlé,* vol. (17): 79-101.

Pallaud B. (2004). La transgression et la variation. *Marges Linguistiques*, vol. (8) : 76-87.

Paumier S. (2011). *Unitex user manual*. http://igm.univ-mlv.fr/~unitex.

Piton O., Maurel D. et Belleil C. (1999). The Prolex Data Base : Topopnyms and gentiles for NLP. *Proc. of NLBD'99 (Third International Workshop on Applications of Natural Language to DataBases)*

Ramisch C., Villavicencio A. et Boitet Chr. (2010). mwetoolkit: a Framework for Multiword Expression Identification. *Proc. of LREC'10 (7th International Conference on Language Resources and Evaluation)*.

Rhapsodie, http://rhapsodie.risc.cnrs.fr

Sag I. A., Baldwin T., Bond F., Copestake A. et Flickinger D. (2002). Multiword expressions: a pain in the neck for NLP. *Proc. of CICLING'02 (3rd International Conference on Intelligent Text Processing and Computational Linguistics)*.

Sagot B. (2010). The Lefff, a freely available, accurate and large-coverage for French. *Proc. of LREC'10 (7th international conference on Language Resources and Evaluation)*.

Seretan V., Nerima L. et Wehrli E. (2004). Multi-word collocation extraction by syntactic composition of collocation bigrams. *Recent Advances in Natural Language Processing III: Selected Papers from RANLP 2003*. Amsterdam & Philadelphia: John Benjamins : 91-100

Shriberg E. (1994). *Preliminaries to a Theory of Speech Disfluencies*, Université de Berkeley, Thèse non publiée.

Watrin P. et François Th. (2011). An N-gram Frequency Database Reference to Handle MWE Extraction in NLP Applications. *Proc. of MWE'11 (ACL Workshop on Multiword Expressions: from Parsing and Generation to the Real World)*.