L'autocorrection immédiate en français parlé : le cas des déterminants

Anne Dister

Université de Louvain – Centre de recherche VALIBEL

Abstract

Transcriptions of spoken French contain many features, called disfluencies, that are not present in standard written French (euh, repetitions, repairs,…). These disfluencies constitute an important problem for the Natural Language Processing. In the aim of part-of-speech tagging of spoken corpora, we analyse in-depth these "disfluencies." In this article, we will focus on the self-correction of the determiners. We will make a classification by sub-categories, and we will analyse their configuration, frequency and interaction with the repetition. We will show that the self-correction of determiners rarely breaks up the construction of the sentence.

Résumé

Les transcriptions de français parlé notent une série de phénomènes qui les distinguent du français écrit standard, phénomènes, qualifiés traditionnellement de *disfluences* (répétitions, euh, etc.), perturbent la linéarité de l'énoncé et constituent un problème pour le traitement automatique des langues. Dans le cadre de l'étiquetage morphosyntaxique de données orales, nous avons analysé de manière approfondie certaines *disfluences*, notamment les auto-corrections immédiates. Dans cet article, nous nous centrons sur les autocorrections de déterminants définis. Nous les classons selon leur sous-catégorie, en analysant leur format, leur fréquence et leur interaction avec le phénomène de la répétition. Nous montrons que leur apparition constitue rarement une rupture syntaxique de l'énoncé.

Mots-clés : *disfluences*, autocorrection immédiate, déterminant, corpus de français parlé, transcription de l'oral, annotation morphosyntaxique.

1. Introduction

Cet article s'inscrit dans le cadre d'une recherche plus vaste sur l'annotation morphosyntaxique de transcriptions de français parlé. L'objectif final de notre travail est d'enrichir des transcriptions de l'oral, en ajoutant à chaque mot (ou groupe de mots), une étiquette qui indique son lemme, sa catégorie grammaticale et des informations flexionnelles.

Cette tâche, relativement bien maitrisée pour des données standard (Véronis, 2000 ; Clément, 2001), n'a guère été tentée sur des corpus oraux retranscrits (voir cependant Mertens, 2002 ; Valli et Véronis, 1999).

C'est qu'en effet les transcriptions de l'oral notent des particularités qui les distinguent de l'écrit standard. Ainsi, les transcriptions sur lesquelles nous travaillons, réalisées au centre de recherche VALIBEL[1], valorisent l'oralité des données : *euh*, ponctuants, répétitions, amorces de morphèmes, interruptions, chevauchements de parole, etc. sont notés strictement dans une

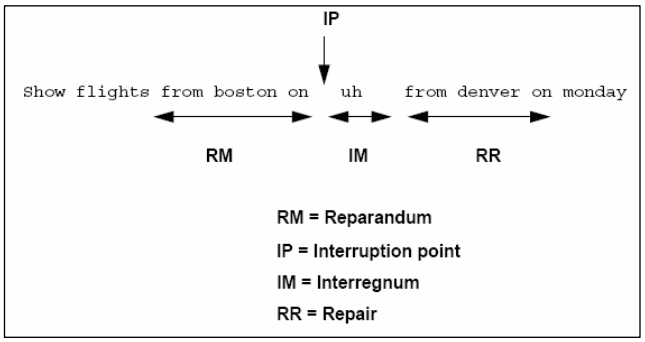
[1] http://valibel.fltr.ucl.ac.be/

transcription non ponctuée qui adopte l'orthographe standard (Dister *et al.* 2006). Ces phénomènes sont traditionnellement qualifiés de *disfluences*[2].

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

Shriberg (1994 : 7-9), à la suite notamment de Levelt (1989), a modélisé la séquence disfluente en la décomposant en quatre éléments distincts, qui correspondent à trois régions :

- ***reparandum*** : le *reparandum* (*RM*) est la partie produite par le locuteur qui ne sera pas conservée et sera remplacée ultérieurement au profit du *repair* ;
- ***interrupting point*** : le *point d'interruption* (*IP*) est le moment de l'énoncé qui coïncide avec la fin du *reparandum*. Ce point d'interruption est vide ;
- ***interregnum***[3] : *l'interregnum* (*IM*) est la région qui commence à la fin du reparandum et s'achève au début du *repair*. *L'interregnum* peut ou non contenir un terme d'édition (*editing term*), c'est-à-dire une pause silencieuse, une pause remplie, etc., ou plusieurs autres tentatives de formulation inachevées également ;
- ***repair*** : le *repair* (*RR*) indique le retour à la « fluence », par la réparation, la correction du *reparandum*.



Modélisation de la séquence disfluente (Shriberg 1994 : 8)

Depuis longtemps, on a tenté de montrer certaines régularités qui affectent les disfluences, notamment pour l'anglais (Blankenship et Kay, 1964 ; Cook, 1971)[4].

La question qui sous-tend notre travail est double : quel est le format de ce piétinement ? Celui-ci constitue-t-il, selon le format adopté, une rupture de la structure syntaxique de l'énoncé ? En effet, notre méthode d'étiquetage analyse le contexte local de la chaine graphique (Laporte et Monceaux, 1998 ; Dister 2006) ; si le déroulement syntagmatique est stoppé (momentanément ou définitivement), la bonne application des grammaires locales est entravée. Nous avons donc analysé de manière approfondie 4 types de disfluences, afin de les intégrer dans le processus d'annotation (Dister, 2007).

---

[2] Comme le dit Habert (2005 : 57) : « (…) on manque ainsi de termes positifs pour décrire les régulations de l'oral, parfois fâcheusement dénommées *disfluences* par transfert de l'anglais *disfluencies.* » Nous conservons néanmoins ce terme, sans y voir de connotation péjorative.

[3] L'*interregnum* correspond à l'*editing phase* de Levelt.

[4] Voir aussi les travaux réalisés par l'équipe autour de Claire Blanche-Benveniste, le Groupe Aixois de Recherche en Syntaxe (GARS), devenu DELIC.

2. L'autocorrection parmi les disfluences

Nous appelons *autocorrection (immédiate)*[5] le phénomène langagier qui consiste, pour un locuteur, à énoncer un morphème suite à un autre morphème différent qui appartient à la même catégorie grammaticale. Ce deuxième morphème vise à corriger le premier morphème énoncé.

Voici un exemple type d'autocorrection :

> **ilePA2** (…) je me dis après tout l'accent c'est un peu **le la** carte d'identité c'est un peu l'appartenance à une région (…) [ilePA2r]

Le corpus sur lequel nous avons travaillé comprend 443 047 mots graphiques, ce qui correspond *grosso modo* à 40 heures de parole.

Dans cet article, par manque de place, nous nous penchons exclusivement sur les autocorrections qui concernent les déterminants[6].

3. Les types de déterminants

Nous avons recherché de manière automatique toutes les séquences où cooccurrent des déterminants définis. Dans la recherche, nous avons prévu l'insertion d'éléments, à savoir : *ah, bè, ben, boh, bon, comment, disons, enfin, euh, ha, hein, hum, m, mh, mm, non, oh, oui, ouais, pf, pff,* la pause brève (/), la pause longue (//), le silence (*silence*) et l'amorce de morphème.

3.1. Les déterminants définis : le, la, les, l'

Les 208 séquences d'autocorrection immédiates du déterminant défini se ventilent comme suit dans le tableau ci-dessous. La colonne en abscisse indique la forme initialement énoncée par le locuteur (= la forme au début de la séquence disfluente), la colonne en ordonnée représente la forme qui constitue la correction (= la fin de la séquence disfluente avec la reprise du déroulement sur l'axe syntagmatique).

Nous présentons les résultats en pour cent, et également, entre parenthèses, en nombre d'occurrences.

[5] La terminologie est loin d'être unifiée, et l'on rencontre dans la littérature le terme d'*autocorrection* pour n'importe quelle séquence dans laquelle le locuteur se corrige, indépendamment de la forme que peut prendre cette correction. Fornel et Marandin (1996) utilisent quant à eux le terme *auto-réparation* ; Bénard (2005) emploie le terme *révision* ; Grosjean et Deschamps (1972) préfèrent *faux-départ*, équivalent de l'anglais *false-start* ; Martinie (2001) parle d'énoncés réparés. Candea (2000) distingue quant à elle *autocorrections immédiates* et *faux départs*, ceux-ci étant plus complexes et concernant aussi des mots pleins ou des structures syntaxiques inachevées.

[6] Parmi les 4 types de disfluences auxquelles nous nous sommes intéressée, les autocorrections sont les moins fréquentes : on a en effet 12 192 répétitions, 9423 *euh* et 3612 amorces de morphèmes. Cela tient à notre définition des différents phénomènes, et au fait que nous nous sommes limitée aux autocorrections repérables formellement, sans ambigüité. En effet, notre volonté est d'automatiser leur repérage afin d'intégrer les résultats obtenus dans une phase de prétraitement, elle aussi totalement automatisée (voir Dister, 2007). La collecte opérée ici peut donc sembler bien faible, au regard du nombre d'occurrences obtenu pour les autres disfluences. Il nous semble néanmoins qu'elle permet d'éclairer le phénomène.

	l’	le	la	les	TOTAL
l’	0,48 % (1)	1,44 % (3)	0 % (0)	1,44 % (3)	3,37 % (7)
le	19,71 % (41)	0,48 % (1)	20,19 % (42)	13,46 % (28)	53,85 % (112)
la	8,17 % (17)	13,94 % (29)	0,48 % (1)	4,81 % (10)	27,40 % (57)
les	2,88 % (6)	6,25 % (13)	5,77 % (12)	0,48 % (1)	15,38 % (32)
TOTAL	31,25 % (65)	22,12 % (46)	26,44 % (55)	20,19 % (42)	100 % (208)

Autocorrections du déterminant défini

Des données présentées dans ce tableau, nous pouvons dégager certaines tendances générales. En effet, mise à part la correction du *l’* en *la*, le corpus atteste toutes les possibilités de corrections du déterminant défini par un autre déterminant défini. Nous verrons ci-dessous que ce n’est pas le cas pour les autres sous-catégories de déterminants dans notre corpus.

Si toutes les possibilités sont représentées, c’est néanmoins en fréquence variable. La forme *le* est celle sur laquelle porte le plus fréquemment l’autocorrection, avec près de 54 % des occurrences. Vient ensuite le *la* (27,40 %), le *les* (15,38 %) et enfin le *l’*, concerné par seulement 7 occurrences. Le petit nombre de cas de *l’* s’explique selon nous par les choix de transcription. En effet, nous formulons l’hypothèse que dans ces cas, le transcripteur a noté l’élision (ou ce qui pourrait en tout cas être interprété raisonnablement comme une élision) non pas en tant que telle mais comme une amorce de morphème – ce qui explique notamment la forte représentation de ce type d’amorce, corrigée immédiatement par la forme considérée comme pleine (Dister, 2007 : 198 et sv.). Ceci explique aussi sans doute qu’on n’a pas de répétitions du *l’* dans nos données.

On a ainsi :

> **accBF1** non / mais / de par les Français qui / qui imitent **l/ l’**accent bruxellois // ils disent toujours euh une fois mais si/ sinon [accBF1r]

En fait, toutes les occurrences de *l’* prises en compte ici sont suivies d’une amorce de morphème, comme le montre l’exemple suivant :

> **norGA1** donc **l’en/ le** maître est important mais il n’est pas tout seulement un mauvais maître ben on perd son temps [norGA1r]

Le contexte droit du son [l] orienterait ainsi les choix de transcription[7] : forme élidée quand une amorce de morphème suit, amorce de morphème dans tous les autres cas.

[7] Notons qu’on n’a aucune indication de ce type dans les conventions de transcription de VALIBEL, mais il semble bien que tous les transcripteurs aient procédé de la même manière au fil des années.

On a aussi de nombreux exemples où, à la seule lecture de la transcription, on peut se demander si un autre choix n’aurait pas été possible (autre choix peut-être invalidé par des indices prosodiques que nous ne possédons pas).

En ce qui concerne le *repair*, la répartition des données en termes de fréquence est moins inégale que pour la forme qui subit la correction (le *reparandum*). On constate néanmoins une préférence pour la forme élidée, avec 31,25 % des séquences qui s'achèvent finalement par cette forme. Viennent ensuite *la* (26,44 %), *le* (22,12 %) et enfin *les* (20,19 %).

Si l'on se centre non plus maintenant sur les formes en tant que telles mais bien sur le nombre du déterminant, on constate que le singulier est largement dominant, aussi bien en ce qui concerne la forme de départ (84,62 % des occurrences) que la forme choisie au final (77,88 %). Néanmoins, lorsqu'il y a changement de nombre dans la réparation, on se rend compte que celui-ci concerne plus souvent un passage du singulier vers un pluriel (19,34 %) que l'inverse (14,9 %).

L'autocorrection peut également consister, pour le locuteur, à ajuster le genre de son déterminant au substantif qu'il va énoncer.

Plusieurs hypothèses ont été avancées en ce qui concerne les hésitations sur le genre. Pour Sauvageot (1962 et 1972), le locuteur hésite sur le genre du mot qui va suivre. Cappeau (1998) remet en cause cette hypothèse :

> Le genre propre au nom impose la forme du déterminant qui le précède. D'où la nécessité de se livrer à une « analyse anticipative » et les risques de conflit qui peuvent conduire à ces ratés. (…) [les exemples cités] concernent des termes courants qui ont peu de chance d'être mal maîtrisés par les locuteurs. L'explication par une méconnaissance du genre serait donc peu plausible. (Cappeau 1998 : 305)

Au vu des substantifs concernés dans nos exemples, on peut difficilement suivre l'hypothèse de Sauvageot.

Outre les formes *le* corrigées en *la* (20 % des occurrences) et inversément (près de 14 %), on a les corrections qui concernent la forme élidée. Sous cette forme, le genre du déterminant est indistinct. Dans certaines corrections, il s'agit non pas d'un changement de genre mais simplement d'une adaptation phonétique : le locuteur a énoncé un déterminant, dont le genre correspond bien à celui du substantif qu'il anticipe, mais qui doit en fait être élidé puisque celui-ci commence soit par une voyelle, soit par un *h* dit aspiré :

> **ilpCM1** (…) je vais pas faire ici **le l'**hypocrite (…) [ilpCM1r]

Mais dans d'autres cas, le déterminant élidé n'est pas du même genre que celui du substantif qui le précède :

> **ileGF0** mm mais vous êtes quand même si si je vous suis bien donc **le l'**écriture est un un moyen de traduction euh de la réalité [ileGG1r]

On a un processus comparable pour la forme corrigée *les*. Le locuteur corrige le nombre, passant d'un singulier à un pluriel, sans qu'il y ait nécessairement correction du genre. Dans l'énoncé suivant, on passe d'un masculin singulier à un masculin pluriel :

> **ilrDT1** je ne sais pas moi (silence) mais tous tous l/ les coins de Belgique où **le / les** revenus sont un peu pluS élevés [ilrDT1r]

Ici, la séquence *la la l'accent* aurait peut-être pu être transcrite *l'a/ l'a/ l'accent*. Tout au long de ce travail, nous avons adopté le parti pris de faire confiance à la transcription. La séquence est donc comptabilisée ici dans les autocorrections.

> **accPH1** parce que **la la l'**accent de la mer belge // je comprends // si vous allez / tout doucement je comprends mais ici elle pouvait parler tout doucement comme tu |- veux [accPH1r]

Mais la correction peut porter non seulement sur le nombre, mais également sur le genre. Dans l'énoncé suivant, le locuteur entame le syntagme avec un déterminant féminin singulier pour le corriger en un masculin pluriel :

> **ileAO1** vous verrez que **la les** trois quarts de la rédaction ont déjà quitté les lieux [ileAO1r]

Dans notre corpus, les résultats pour les corrections en *l'* et *les* peuvent être synthétisés comme suit :

la → l' : 10 adaptations phonétiques, 7 changements de genre ;

le → l' : 23 adaptations phonétiques, 18 changements de genre ;

la → les : 3 féminins, 5 masculins (5 changements de genre) ;

le → les : 16 masculins, 9 féminins (9 changements de genre).

On le voit, hormis pour le passage du *la* en *les*, les corrections qui aboutissent au déterminant élidé ou à *les* consistent plus souvent à conserver le même genre qu'à le modifier. Néanmoins, les résultats entre ces deux « stratégies » ne sont pas véritablement tranchés puisqu'on a au total 57,14 % d'occurrences pour lesquelles on garde le même genre, contre 42,86 % des cas dans lesquels le genre est modifié.

C'est pour le masculin que le genre est le plus souvent conservé, avec 59 % des occurrences, tandis que pour le féminin, le score tombe à une conservation du genre dans 52 % des cas.

Avec ces cas de changement de genre, ajoutés aux changements déjà repérés dans le tableau ci-dessus, on comptabilise 110[8] occurrences dans lesquelles la correction consiste au moins en un changement de genre (avec parfois aussi un changement de nombre), soit près de 53 % du total des occurrences d'autocorrections. Ces changements de genre montrent une nette préférence pour le changement qui va du masculin vers le féminin (62,72 %), les corrections aboutissant au déterminant *la* étant moins nombreuses (37,27 %).

On voit donc ici une nette préférence pour le masculin (au départ comme à l'arrivée). On a également plutôt un passage du pluriel au singulier que l'inverse. Néanmoins, cette forte tendance n'est pas une généralité puisque de nombreux cas attestent le mouvement inverse. Nous ne pouvons donc aller dans le sens de la conclusion à laquelle aboutit Candea (2000 : 357), qui affirme que

> toutes ces autocorrections [à part une exception] vont du *plus neutre* au *plus spécifié* (masculin → féminin, présent → non présent, singulier → pluriel).

L'hypothèse sous-jacente est que la première apparition de la séquence serait en quelque sorte automatique, rectifiée ensuite pour aller du plus général (*grosso modo* le masculin singulier), au plus spécifique (le féminin et/ou le pluriel). Cela va dans le sens de Cappeau (1998 : 307) :

> La prédominance du masculin pourrait être due au caractère non marqué de ce genre : le locuteur qui cherche un terme l'utiliserait donc en priorité.

Cela dit, le total de 53 % de changement de genre que l'on obtient est finalement faible en regard de ce à quoi on pouvait s'attendre. Le résultat qui nous semble finalement intéressant ici est le nombre important d'occurrences qui participent d'une adaptation morphologique/phonétique.

[8] 42+29+7+18+5+9=110.

Dans le tableau donné ci-dessus, nous avons considéré la forme de départ et la forme d'arrivée de la séquence. On a donc pu s'étonner de trouver dans les autocorrections un *le reparandum* et un *le repair* au final de la séquence, comme s'il s'agissait d'une répétition. C'est en fait parce que nous avons recherché les séquences disfluentes maximales, c'est-à-dire qui s'achèvent avec la réparation. Or, il arrive que cette réparation ait la même forme que le terme de départ, avec entre les deux un changement de déterminant. C'est le cas de l'exemple suivant :

> **ileGG1** euh / finalement **le / la le** contenu est quand même servi par la forme [ileGG1r]

Néanmoins, ces cas de retour au déterminant initial sont rares, puisque l'on n'en rencontre que 4 occurrences dans le corpus.

Nous avons ici examiné les formes *le, la, les* et *l'* appartenant à la catégorie des déterminants. Nous avons donc fait un tri manuel puisque ces formes peuvent aussi être des pronoms.

En fait, on constate qu'on n'a que deux occurrences d'autocorrections qui concernent ces formes lorsqu'elles sont pronoms. On a dans notre corpus la même tendance pour les autocorrections que pour les répétitions, avec une écrasante majorité en faveur de la catégorie des déterminants par rapport à celle des pronoms pour les formes potentiellement ambigües (Dister 2007 : 165).

Les deux occurrences recensées dans le corpus suivent la même structure : le pronom *le* est corrigé immédiatement en un *l'* suivi d'un verbe à l'infinitif qui débute par une voyelle (*le l'entretenir* et *le l'enseigner*).

3.2. Les déterminants indéfinis

Pour les déterminants indéfinis, nous sommes directement confrontée au problème de l'ambigüité virtuelle : une suite de deux formes qui sont possiblement deux déterminants indéfinis en langue ne sont pas nécessairement des déterminants indéfinis en discours. Si l'on cherche ainsi une suite de deux déterminants indéfinis, dans un texte non désambiguïsé (qui est le format de texte sur lequel nous travaillons dans cette phase de la recherche), on recense des exemples comme les suivants :

> **chaGG0** et vous / vous souvenez **d'un** problème d'héritage qu'il y aurait eu [chaSR1r]
> **chaBR1** et voilà voilà **un des** enterrements où il y avait beaucoup des gens à cet enterrement-là (…) [chaBR1]

puisque *un, d'* et *des* sont possiblement des déterminants.

En fait, les exemples ci-dessus ne participent pas du tout de la disfluence. Afin de travailler sur des autocorrections certaines, nous avons décidé de nous centrer uniquement sur les autocorrections de *un* en *une* et inversément. Les résultats sont les suivants :

	un	**une**
un	1	41
une	21	2

Autocorrections des déterminants indéfinis

Ici, la tendance que l'on observait pour le déterminant défini, en ce qui concerne le genre, est encore accentuée. En effet, 63 % des occurrences montrent une correction du masculin en féminin, alors que 32,3 % des items manifestent le passage inverse.

On a trois cas où l'autocorrection aboutit finalement à la forme de départ : une occurrence concerne le masculin et deux le féminin. On ne peut tirer de conclusion d'un échantillon si réduit[9].

Si avec les suites *un(e) des* on ne peut garantir que l'on a affaire au phénomène de l'autocorrection, ce n'est pas le cas des séquences *des un* et *des une* (le deuxième déterminant étant au singulier). On obtient les scores suivants :

- *des un* : 9 occurrences
- *des une* : 1 occurrence

N'avant pas relevé les occurrences d'autocorrections qui vont du singulier vers le pluriel, on ne peut faire de comparaison.

3.3. Les déterminants possessifs

Les cas d'autocorrections qui portent sur la sous-catégorie des déterminants possessifs sont beaucoup plus rares puisqu'on n'a que 16 occurrences. Aucune de ces formes n'apparait d'ailleurs ni dans les 100 formes les plus fréquentes du corpus, ni dans les 30 mots grammaticaux qui sont le plus fréquemment l'objet d'une répétition.

Le détail de ces déterminants corrigés est le suivant : *ma – mon* (2*), mes – ma* (4), *mon – ma* (3), *mon – mes* (1), *sa – ses* (1), *ses – son* (1), *son – sa* (2), *tes – ton* (2).

On ne peut guère tirer de conclusions sur un si petit nombre d'occurrences. On ne peut non plus faire de comparaison avec les autres déterminants du point de vue du genre, étant donné que, à part *ma, ta* et *sa*, les autres formes sont épicènes.

Néanmoins, on peut faire trois remarques :

- forme initiale et forme finale appartiennent toutes à la même sous-catégorie : on ne change pas le possesseur (pas d'exemples du type **mon ton*) ;
- on n'a pas de déterminant dont le possesseur est pluriel (*notre, votre, leur, nos, vos, leurs*). On ne voit pas bien comment nos données auraient pu déterminer cela, et selon nous il ne s'agit pas d'un « effet du corpus ». Néanmoins, ceci nous semble un point qui mériterait d'être vérifié sur d'autres données ;
- on n'a aucune rupture syntaxique ; après le déterminant corrigé, le syntagme s'achève normalement.

3.4. Les déterminants démonstratifs

Les autocorrections de déterminants démonstratifs sont à peine plus nombreuses que celles des possessifs, avec 18 occurrences, qui se répartissent comme suit : *ce – cet* (3), *ce – cette* (5), *ce – ces* (4), *ces – cette* (1) et *cette – ce* (5).

[9] Nous recueillons peu d'occurrences, mais rappelons que le corpus sur lequel s'effectue notre recherche se compose tout de même de 443 047 mots, ce qui n'est pas rien.

Pour les démonstratifs, on a aussi des autocorrections liées à un aménagement phonétique, lorsque le *ce* est corrigé en *cet.*

En ce qui concerne le changement de genre, on a 7 occurrences pour le passage du masculin au féminin, et 5 dans l'autre sens. Les occurrences sont trop peu nombreuses pour tirer des conclusions, mais on voit néanmoins que ces résultats montrent une tendance comparable à celle observée pour les déterminants définis.

Dans les occurrences recensées ici, on constate que la réparation n'est pas le lieu d'une coupure syntaxique de l'énoncé, mais que celui-ci poursuit son déroulement syntagmatique après le *repair*.

3.5. Les changements de sous-catégorie de déterminants

On a vu pour les déterminants possessifs que la réparation se faisait à l'intérieur d'une même sous-catégorisation en ce qui concerne le possesseur : un syntagme initié par un déterminant dont le possesseur est à la première personne du singulier est réparé par un déterminant dont le possesseur est également à la première personne du singulier, et ainsi de suite.

En fait, nous avons ici cherché les autocorrections à l'intérieur d'une même sous-catégorie (définis, indéfinis, démonstratifs et possessifs), sans autoriser la possibilité de croisement. En faisant des recherches de ce type sur le corpus, on se rend compte que la récolte est très médiocre, comme le font remarquer Morel et Danon-Boileau (1998 : 86) :

> (…) dans la classe des déterminants du nom, seules sont possibles des corrections de genre ou (plus rarement) de nombre. Dans les cas exceptionnels où l'on finit par changer de classe, ce changement nécessite la présence de marques supplémentaires.

Nous n'avons donc pas exploré davantage cette piste, qui aurait donné une série de séquences dont on n'aurait pu tirer un enseignement sur le fonctionnement global, nous limitant aux mêmes sous-catégories de déterminants.

4. Empilement de disfluences : interaction de l'autocorrection et de la répétition

Nous avons analysé les cas d'interaction de la répétition et de l'autocorrection, ces 2 phénomènes étant, à bien des égards, relativement proches. Nous avons 59 cas où autocorrection et répétition cooccurrent, soit 28 % des cas d'autocorrections recensés. On assiste donc à un piétinement sur l'axe syntagmatique, piétinement renforcé par un empilement des disfluences.

Trois grands cas de figure peuvent se présenter :

1. la répétition apparait avant l'autocorrection :

   norGA1 (…) moi je ne vois pas **la la l'**intérêt d'imiter la prononciation parisienne (…) [norGA1r]

2. la répétition apparait après l'autocorrection :

   ileDC1 euh parce que d' abord pour le d'abord sur le plan de la forme ils ont une langue assez claire / claire bien bien découpée / un peu chantante sans sans en avoir sans avoir **la le le disons le le** ton liégeois par exemple (…) [ileDC1r]

3. l'autocorrection se place au milieu de la répétition :

   accCT0 et n/ / est-ce que pour vous l'accent c'est une euh // c'est ça a à voir avec euh **la le euh la** catégorie sociale // je vous ai demandé ça mais vous n'avez pas été euh [accPH11r]

Ces cas sont ceux pour lesquels la séquence débute et finit par le même déterminant.

Dans le corpus, les deux premières structures sont le plus fréquemment actualisées, à parts égales, avec 24 occurrences chacune. On n'a que 5 séquences qui suivent le troisième format.

Outre ces trois grands cas de figure, on a une séquence où l'on a une répétition puis une autocorrection répétée (une suite de deux répétitions donc) :

> **norHJ1** (…) c'était donc le le best-seller comme on dit |- n'est-ce pas <norFA0> mm -| donc c' était vraiment **la la l/ l'ob/ l'**ouvrage est resté en vedette [norHJ1r]

Nous recensons aussi 4 suites plus complexes, où se mêlent plus intimement répétition et autocorrection[10] :

> **ileDC1** beaucoup par clichés / la réunion de la dernière chance euh les euh les carottes sont cuites **euh les comment euh la / le le** feu couvre ou bien euh la comment dirais-je la réunion de la dernière chance ça (…) [ileDC1r]

Selon Candea (2000 : 323),

> (…) toutes les combinaisons de la répétition avec l'autocorrection font partie de marques de TdF [travail de formulation] très complexes comportant des combinaisons de 4 ou plusieurs phénomènes ; ces combinaisons occasionnent toujours une rupture importante de l'énoncé.

Or, au vu de nos données, on ne peut tirer une telle conclusion. En effet, parmi ces séquences d'empilement, nous n'avons que 7 cas où la disfluence marque le lieu d'une rupture syntaxique dans la construction de l'énoncé : le discours, s'il se poursuit sur l'axe syntagmatique, laisse inachevée la place syntaxique ouverte par le déterminant initialisant le syntagme. Par comparaison, nous avons pour les autocorrections de déterminants définis également 7 autres cas de ruptures syntaxiques, dans des séquences où l'on n'a pas d'empilement avec la répétition. Nous sommes confrontée au problème de la reconnaissance de la rupture syntaxique. Dans notre corpus, l'empilement, qui pourrait donner à penser à une plus grande complexité (et donc à un « raté » plus important en un point précis), n'est pas un indice certain de rupture : autant de cas de rupture avec l'autocorrection se font sans empilement avec la répétition qu'avec empilement ; plus de cas d'empilements ne donnent pas lieu, comme le suggère Candea, à une rupture syntaxique. On a d'ailleurs des séquences complexes, où plusieurs marques se succèdent, pour lesquelles le déroulement syntagmatique se poursuit régulièrement passé le point de disfluence, même si celui-ci est long et que s'additionnent divers phénomènes :

> **iljNJ1** pff non // il faut dire que // moi le wallon que je connais n'est pas vraiment le wallon pur tel que le parlent **le le enfin le les** vrais wallingants c'est plutôt vraiment un mélange de patois de de certaines expressions qui sont un peu mélangées (…) [iljNJ1r]
>
> **norKJ1** (…) il y a évidemment **le le le // euh l'**aspect extérieur hein qui // qui qui qui a maquillé cette réalité-là (…) [norKJ1r]

Dans le premier énoncé, on a d'une part une séquence composée d'une répétition, à l'intérieur de laquelle on a un ponctuant (*enfin*), suivie d'une autocorrection ; la disfluence du second exemple voit s'empiler une répétition avec deux répétables, suivie d'une pause longue et d'un *euh*, avant d'arriver à l'autocorrection du déterminant. Ces séquences disfluentes, pourtant longues, n'empêchent pas le déroulement syntaxique de se poursuivre, avec l'achèvement du groupe nominal initié par le déterminant.

[10] Les 3 autres séquences ont la forme : *le la le le la ; les la la l' ; les le le la.*.

5. Conclusion

L'autocorrection de déterminant est relativement peu fréquente dans nos données. Nous avons néanmoins pu mettre en évidence certaines régularités :

- les changements de genre et/ou de nombre vont préférentiellement du général vers le particulier, c'est-à-dire du singulier vers le pluriel et du masculin vers le féminin, mais l'on ne peut en faire une règle générale ;
- l'autocorrection pour les déterminants se fait à l'intérieur d'une même sous-catégorie ;
- répétition et autocorrection sont souvent cooccurrentes dans une même séquence, et il est des cas où les phénomènes sont entremêlés, au point qu'après une autocorrection, le *reparandum* est le même que le *repair* ;
- les formes les plus souvent concernées par l'autocorrection sont celles qui sont les plus fréquentes dans le corpus général, et aussi celles qui font le plus souvent l'objet d'une répétition ;
- la fin de l'autocorrection donne lieu en général à la reprise du déroulement syntagmatique de l'énoncé, sans rupture. Lorsque celle-ci a lieu, on a bien du mal à trouver des indices formels qui indiqueraient qu'on laisse l'énoncé inachevé à l'endroit de l'autocorrection.

Dans notre système de levée d'ambigüités, nous avons donc considéré que l'autocorrection ne constituait pas une rupture dans l'énoncé, et les séquences concernées ont été balisées automatiquement, afin de ne garder pour l'analyse par les grammaires locales que le *repair*.

Références

Bénard F. (2005). *Normalisation de corpus oraux : des métadonnées à l'annotation des transcriptions*. Université Paris-3, Sorbonne Nouvelle, Mémoire de maîtrise.

Blankenship J., Kay C. (1964). Hesitation phenomena in English Speech: a study in distribution. In *Word*, 20, 360-372.

Blanche-Benveniste C., Bilger M., Rouget C., van den Eynde K. (1990). *Le Français parlé. Études grammaticales.* CNRS Éditions.

Candea M. (2000). *Contribution à l'étude des pauses silencieuses et des phénomènes dits « d'hésitation » en français oral spontané*, Université de Paris-3 Sorbonne-nouvelle, Thèse non publiée.

Cappeau P. (1998). Quelques mots sur quelques bribes liés au genre. In Bilger M., van den Eynde K. et Gadet F. (Éds.), *Analyse linguistique et approches de l'oral. Recueil d'études offert en hommage à Claire Blanche-Benveniste.* Orbis Supplementa, 10, Peeters, 301-311.

Clément L. (2001). *Construction et exploitation d'un corpus syntaxiquement annoté pour le français*, Université Paris-7, Thèse non publiée.

Cook M. (1971). The Incidence of Filled Pauses in Relation to Part of Speech. *Language and Speech* 14, 135-150.

Dister A. (2006). *Les grammaires ELAG pour le français*. http://www-igm.univ-mlv.fr/~unitex/.

Dister A. (2007). *De la transcription à l'étiquetage morphosyntaxique. Le cas de la banque de données textuelles orales VALIBEL*, thèse non publiée, Université de Louvain.

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

de Fornel M., Marandin J.-M. (1996). L'analyse grammaticale des auto-réparations. *Le gré des langues*, 10, 8-68.

Grosjean F., Deschamps A. (1972). Analyse des variables temporelles du français spontané. *Phonetica*, 26, 129-156.

Habert B. (2005). *Instruments et ressources électroniques pour le français*. Ophrys.

Laporte É., Monceaux A. (1999). Elimination of lexical ambiguities by grammars: the ELAG system. In Fairon C. (Éd.), *Analyse lexicale et syntaxique : le système INTEX, Lingvisticæ Investigationes*, 22 (1-2). John Benjamins, 341-367.

Levelt W. J. M. (1989). *Speaking: from intention to articulation*. MIT Press.

Martinie B. (2001). Remarques sur la syntaxe des énoncés réparés en français parlé. *Recherches sur le français parlé*, 16, 189-206.

Mertens P. (2002). Les corpus de français parlés ELICOP : consultation et exploitation. In J. Binon, P. Desmet, J. Elen, P. Mertens, L. Sercu (Éds.), *Tableaux Vivants. Opstellen over taal-en-onderwijs aangeboden aan Mark Debrock.* Leuven Universitaire Pers.

Morel M.-A., Danon-Boileau L. (1998). *Grammaire de l'intonation. L'exemple du français.* Ophrys.

Sauvageot A. (1962). *Français écrit, français parlé.* Larousse.

Sauvageot A. (1971). L'articulation du discours. In Rigault A. (Dir.), *La Grammaire du français parlé.* Hachette, 137-147.

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

Valli A., Véronis J. (1999). Étiquetage grammatical des corpus de parole : problèmes et perspectives. *Revue française de linguistique appliqué,* 4 (2), 113-133.

Véronis J. (2000). Annotation automatique de corpus : panorama et état de la technique. In J.-M. Pierrel (Éd.), *Ingénierie des langues.* Hermès, 111-129.