SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801386

« parler sans accent pour moi c'est sans sans sans bafouiller » Quelles répétitions de formes en français parlé ?

Dister, Anne

Université Saint-Louis – Bruxelles
anne.dister@usaintlouis.be

1 Introduction

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

Connaissant un intérêt récent auprès des linguistes étudiant l'oral, les disfluences ont principalement fait l'objet d'études de la part des psycholinguistes (les pionniers Maclay et Osgood 1959, Levelt 1989), des psychanalystes (on sait l'intérêt de Freud pour le lapsus, où l'attention est portée davantage sur le lexique) ou encore des psycho-cliniciens. C'est l'aspect « raté » qui a d'abord intéressé, ainsi que les mécanismes d'encodage et de décodage des énoncés (cf. Duez 2001 pour un historique de ces notions).

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

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

La disfluence de certains énoncés est donc opposée à la fluence de certains autres, avec les jugements de valeur que cela implique bien souvent. Comme le constate Habert (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.*

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

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

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

Article available at http://www.shs-conferences.org or http://dx.doi.org/10.1051/shsconf/20140801386

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801386

On assiste pourtant aujourd'hui à un retour en grâce du terme *disfluence*. Dans la littérature anglo-saxonne, on rencontre également le terme *non-fluencies* (Hindle 1983, Neslon 1996).

De notre point de vue, le terme *hésitation* a ceci de gênant qu'il fait référence à un processus cognitif. Ainsi, malgré les critiques que l'on peut émettre sur le choix du terme *disfluence*, nous le préfèrons à celui de *hésitation*. En effet, nous ne nous intéresserons pas ici aux fonctions des disfluences en termes cognitifs ou selon leurs éventuelles fonctions énonciatives – qu'elles soient des marques du travail de formulation (Morel et Danon-Boileau 1998), signe d'hésitation, signe d'un malaise ou d'une difficulté d'encodage, recherche du mot juste dans la mémoire du locuteur, marque signifiant la volonté de poursuivre l'énoncé, etc. – mais à leur fréquence et leur régularité en corpus. Notre but est en effet de cerner le phénomène du point de vue de la morphosyntaxe et d'en voir les éventuelles régularités afin de les formaliser dans la perspective d'une analyse automatisée ou semi-automatisée.

2 Modélisation des disfluences

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

Shriberg (1994 : 7-9), à la suite notamment de Levelt (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*** : *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*.

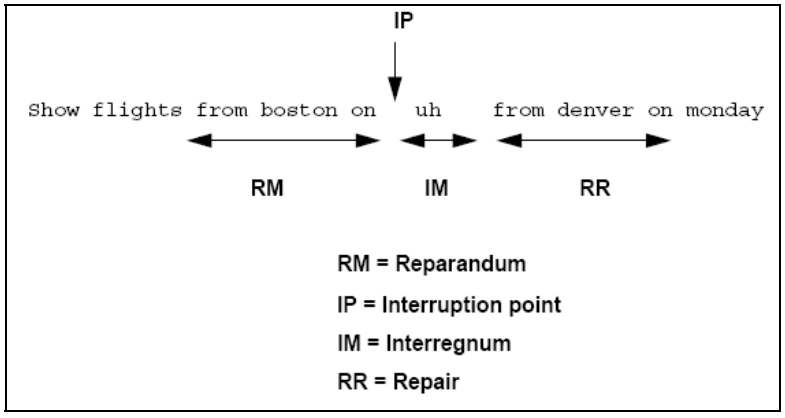

Figure 1. Modélisation de la séquence disfluente (Shriberg 1994 : 8)

Depuis longtemps, on a tenté de montrer certaines régularités qui affectent les disfluences, notamment pour l'anglais (Blankenship et Kay 1964, Cook 1971). L'équipe autour de Claire Blanche-Benveniste, le

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801386

Groupe Aixois de Recherche en Syntaxe (GARS), s'est attachée elle aussi, depuis ses débuts, à mettre en évidence les régularités des modes de production de l'oral – y compris les disfluences –, et à les intégrer dans une description unifiée de la langue. Parmi les différents types de disfluences, nous allons nous intéresser plus particulièrement à la répétition.

3 La répétition

3.1 Définition

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

La rhétorique classique, déjà, faisait usage du concept de la répétition, terme générique qui comprenait des figures différentes (assonance, allitération, épanaphore, épistrophe, épanalepse, anadiplose, épanadiplose, homéotéleute, etc.).

Parmi toutes ces figures qui relèvent de la répétition, certaines concernent une identité de contenu, d'autres une identité de forme. Dans cette étude, nous nous intéressons aux répétitions de forme, et reprenons à notre compte la définition de la répétition proposée par Candéa (2000a : 315) :

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

Une répétition est donc une séquence totale qui comprend un répétable et un ou plusieurs répété(s). Pour le dire autrement : **répétition = répétable + répété(s)**

La longueur du répétable est variable. Celui-ci peut être constitué soit d'un seul mot, soit de plusieurs mots, au sens typographique du terme. Le répétable peut être repris une seule fois ou plusieurs fois. On parlera d'un mot répété une fois (répétition simple), s'il apparait deux fois (répétable + répété), répété deux fois (répétition multiple) s'il apparait trois fois (répétable + répété1 + répété2), etc.

3.2 Les types de répétitions formelles

Parmi les répétitions formelles, certaines concernent la reprise d'un élément qui ne relève pas d'un choix du locuteur mais des exigences de la syntaxe française. Il en est ainsi des pronoms aux 1re et 2e personnes du pluriel dans la conjugaison des verbes réciproques et réfléchis, par exemple (*nous nous levons*). D'autres ont un effet oratoire. On trouve fréquemment l'interprétation intensive de la répétition (*c'est pas joli joli, il est un peu fou fou, tiens tiens !*). D'autres encore permettent une modification de sens (**magPM1** (…) *et puis alors un gâteau mais un* ***gâteau gâteau*** *pas un gâteau de glace* [magPM1r]), ou une mise en évidence (*vous, vous êtes belle*). Toutes ces répétitions relèvent des règles de la langue. A côté de cette 1re catégorie, on trouve les répétitions « faits de parole » qui relèvent de la performance et participent des achoppements propres aux productions orales : à la différence des répétitions « faits de langue » pour lesquelles le discours poursuit son déroulement sur l'axe syntagmatique, on constate que les répétitions « faits de parole » « nous oblige[nt] en fait à "piétiner" sur le même emplacement syntaxique » (Blanche-Benveniste 1985 : 113) ; la linéarité du discours est interrompue et un entassement se fait sur un même point de l'axe syntagmatique. Voici un exemple typique de répétition « fait de parole » : **blaAD0** *ça va il a l'air* ***de de*** *te reconnaitre* [blaNB1l]. Ce type de répétition n'a pas de

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801386

fonction sémantique : le sens de l'énoncé ne change pas à cause de la répétition ; il n'est pas inscrit dans la syntaxe de la langue. Il est donc imprévisible, même si nous allons voir qu'il obéit à certaines régularités.

3.3 Un partage difficile entre disfluence et faits de langue

Les exemples donnés jusqu'ici sont relativement clairs quant à leur statut : inscrits dans la grammaire de la langue, ayant un effet oratoire ou relevant d'un achoppement dans la linéarité de la parole. Dans notre corpus, de nombreux cas ne peuvent être distingués avec autant de certitude, et les indices prosodiques n'aident pas nécessairement à la catégorisation. En effet, disfluences et figures de style sont parfois très proches. Entre les deux : la notion d'intention du locuteur (Martinie 2001). Pour notre travail, nous évacuons cette notion d'intention du locuteur non seulement parce qu'elle échappe à nos instruments, mais parce qu'elle implique un jugement. Nous faisons donc un traitement analogue pour tous les types de répétitions formelles.

4 Les répétitions dans notre corpus

4.1 Le corpus

Le corpus n'est pas détaillé dans cette proposition de communication afin de ne pas identifier l'auteure. Il s'agit de près de 450.000 mots graphiques de français parlé non planifié.

4.2 Repérage des répétitions

Le repérage des répétitions dans le corpus s'est fait de manière totalement automatisée, en recherchant par programme un mot (ou une suite de mots, toujours au sens typographique du terme) répété immédiatement dans son contexte droit. La longueur du répétable (*l*), en termes de nombre de mots, n'a pas été prédéfinie, de même que le nombre de répétés (*n*).

Néanmoins, il est des répétitions qui ne suivent pas le schéma formel d'une contigüité stricte entre répétable et répété dans la chaine sonore. En effet, il arrive que le répété soit séparé du répétable par l'insertion d'un élément, de nature et/ou de fonction variées. On parlera de *répétition associée* quand un élément apparait entre le répétable et le répété, par opposition à la *répétition directe* où répétable et répété sont contigus (Henry 2001).

Afin de voir quelles sont les séquences possiblement présentes entre le répétable et le répété, nous avons extrait les segments insérés, de maximum quatre mots, entre deux mots ou deux suites de mots identiques.

Nous avons ainsi identifié 21 marques principales insérées entre le répétable et le répété, ou entre deux répétés : *ah, bè, ben, boh, bon, comment, disons, enfin, euh, ha, hein, hum, m, mh, mm, non, oh, oui, ouais, pf, pff*. À cette liste s'ajoutent encore l'amorce de morphème (*le cha/ le chapeau*), la pause brève (/), la pause longue (//) et le silence (*silence*), ainsi que les indications métalinguistiques *(rire), (rires), (toux), (soupir),* et les marques de non-compréhension du transcripteur.

Nous avons exclu de cette recherche automatique les répétitions qui répondent pourtant à notre critère formel mais qui sont trans-tour de parole, qu'il s'agisse d'une répétition qui est le fait d'un même locuteur (comme tous les cas que nous prenons en considération ici), soit de deux locuteurs différents.

Cette méthode de reconnaissance automatique a évidemment ses lacunes. En effet, certaines suites ainsi reconnues comme des répétitions ne le sont pas en termes d'analyse syntaxique. C'est le cas dans l'exemple suivant, où le hasard de la cooccurrence fait reconnaitre des séquences de mots identiques qui se suivent :

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801386

> **norKJ1** (…) je crois / que nous aurions sur le plan // sociolinguistique avantage à décrire // toutes les performances // euh / allant du dialecte / au **français le français le** plus no/ normé le plus proche de la norme (…) [norKJ1r].

Notons d'emblée que ces phénomènes sont peu fréquents relativement au nombre d'occurrences de répétitions au sein du corpus et n'invalident donc pas les données chiffrées que nous allons présenter. De plus, si le repérage, tel que nous le faisons ici, reconnait de mauvaises suites et est inévitablement source d'erreurs pour l'analyse syntaxique, il n'en est pas de même dans le cas de l'étiquetage morphosyntaxique. En effet, l'étiquetage qui consiste à assigner un lemme et une catégorie grammaticale ne sera pas remis en question par ces faux repérages.

La conséquence principale de l'automatisation de la reconnaissance des séquences répétées est que notre collecte regroupe, sous le terme général de *répétition*, des identités formelles de séquences qui sont aussi bien exigées par la syntaxe ou voulues par le locuteur pour des effets oratoires que des disfluences. Certains auteurs (Jeanjean 1984, qui analyse les « ratés » de la communication ; Blanche-Benveniste 1985) voient dans le phénomène de la répétition (disfluent ou intentionnel) un phénomène régulier. C'est donc la régularité de ce procédé général que nous voudrions étudier ici, qu'il relève de la disfluence ou non.

4.3 Résultats globaux : répartition des répétitions dans le corpus

4.3.1 Fréquence générale

Le repérage automatique tel que décrit ci-dessus nous a permis de récolter 12 192 séquences répétitions. Comme nous l'avons dit, celles-ci peuvent être aussi bien des répétitions de langue que des répétitions de parole, ces dernières étant nettement majoritaires dans notre corpus.

Si l'on considère le phénomène de façon générale, sur l'ensemble du corpus, on a 3,19 répétitions tous les 100 mots. Le calcul a été effectué en divisant le nombre de mots du texte par le nombre de mots de la séquence répétition dont on a soustrait le nombre de mots du répétable. On a ainsi considéré que le dernier répété fait partie du texte et non de la séquence disfluente, suivant en cela le schéma proposé par Schriberg (1994) et reproduit ci-dessus.

Henry (2005) obtient environ 17 répétitions tous les 1000 mots, soit une répétition toutes les 23 secondes. L'auteure ne donne pas d'indication sur la manière dont elle a effectué le comptage. On ne sait donc pas si chaque mot compte pour un ou si c'est la séquence formée par la répétition qu'elle considère comme une unité, par exemple. On est devant la même incertitude pour le décompte effectué par Henry *et al.* (2004).

Chez Candéa (2000a), la répétition est la troisième marque la plus fréquente de disfluence (après les *euh* et les allongements). En fait, nos résultats montrent une fréquence du phénomène beaucoup plus élevée que chez Henry *et al.* (2004) ou que chez Candéa (2000a), principalement à cause du fait que nous avons un grand nombre de répétitions qui ne sont sans doute pas prises en compte chez ces auteurs (outre les répétitions non disfluentes, les formes répétées de *oui* et de *non* par exemple, dont on verra ci-dessous la fréquence particulièrement élevée dans notre corpus).

4.3.2 La répartition par locuteur

Comme la plupart des phénomènes relevant de la disfluence, la répétition connait une grande variation inter-locuteurs (Pallaud 2004).

Dans notre corpus, la répartition de la fréquence par locuteur va de 10,18 séquences répétitions tous les 100 mots à 0,10 séquence tous les 100 mots pour le locuteur répétant le moins souvent. La moyenne observée sur l'ensemble des locuteurs du corpus est de 3,19 %.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801386

Si l'on cherche à établir une corrélation entre les différentes données que nous possédons, on n'obtient aucun résultat significatif. On n'observe ainsi aucune corrélation entre la longueur moyenne des tours de parole d'un locuteur et son taux de répétitions : un locuteur ayant un nombre élevé de séquences répétitions n'a pas nécessairement des tours de parole dont la longueur moyenne est relativement courte ou relativement longue, et inversément. Par ailleurs, le paramètre qui consiste à prendre en compte le degré d'expertise du locuteur comme professionnel de la parole donne des résultats peu probants : on a des locuteurs peu scolarisés qui obtiennent un score très bas et des journalistes de la presse télévisuelle avec un taux de répétitions élevé. D'après les résultats que nous obtenons, force est de constater que tombe le mythe qui voudrait qu'un locuteur bafouille, cafouille ou se répète parce qu'il est peu sûr de lui et/ou a une mauvaise maitrise de la langue.

4.3.3 Nombre de répétés (n)

Si l'on observe nos données en fonction du nombre de répétés, la répartition dans le corpus est la suivante :

Nombre de répétés	Nombre d'occurrences	Pourcentage
1	10661	87,44 %
2	1217	9,98 %
3	241	1,98 %
4	44	0,36 %
5	18	0,15 %
6	6	0,05 %
7	3	0,02 %
8	0	0 %
9	1	0,01 %
10	1	0,01 %

Tableau 1. Répartition selon le nombre de répétés

Avec 87,44 % de l'ensemble, les répétitions à un seul répété sont largement majoritaires. Elles sont suivies par celles constituées de 2 répétés, les pourcentages d'occurrences allant décroissant – et devenant pratiquement insignifiants – jusqu'à 10 répétés (n=10). On peut s'étonner de trouver des séquences où le nombre de répétés est si élevé.

Pourtant, comme le fait remarquer Blanche-Benveniste (2000c) : « On peut aller jusqu'à sept, sans que cela se remarque. ». Mais il semble néanmoins exister des seuils au-delà desquels on passe dans la *dysfluence* pathologique (cf. Zellner 1992).

Ainsi, si à la lecture de l'oral transcrit on est inévitablement attiré par ces achoppements qui ne correspondent pas à nos habitudes de lecteur, dont l'œil est aguerri à de l'écrit standard, notre oreille semble quant à elle ignorer totalement ces marques qui passent donc la plupart du temps inaperçues tant elles sont communes dans l'oral spontané.

En fait, les répétés de longueur $n \geq 7$ concernent exclusivement les adverbes *oui* et *non*. Dans notre corpus, ces deux adverbes sont des répétables particulièrement fréquents, puisque le *oui* est l'objet de 824 répétitions (c'est d'ailleurs le répétable le plus fréquent de notre corpus) et le *non* de 296 répétitions. Cette forte fréquence est évidemment liée au fait qu'une grande partie de nos textes sont des entrevues semi-dirigées, dans lesquelles les locuteurs répondent à une série de questions. *Oui* et *non* sont donc des formes très fréquentes en corpus, et le lieu privilégié de répétitions. On est en droit de se demander si

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801386

celles-ci relèvent véritablement du phénomène de la disfluence. C'est sans doute là l'un des facteurs qui explique les écarts entre nos résultats et ceux des études antérieures (notamment Henry *et al.* 2004), en termes de fréquence du phénomène. On peut faire l'hypothèse que si le nombre de répétés est élevé, la répétition n'est pas disfluente. Cela se vérifie pour les séquences où $n \geq 7$.

Dans les cas où le répétable est également un mot grammatical monosyllabique, la répétition est clairement disfluente. C'est en fait moins le nombre de répétés seul que la classe grammaticale du répétable qui permet de formuler des hypothèses quant à la disfluence ou non de la répétition. Nous reviendrons ci-dessous sur le problème de la classe grammaticale.

4.3.4 Longueur du répétable (l)

La longueur du répétable (l), en termes de nombre de mots graphiques, est variable : elle va de l=1 à l=8. Les répétitions les plus fréquentes sont celles où le répétable est composé d'un seul mot ; elles représentent près de trois-quarts des occurrences. Ici aussi, on observe une régularité indéniable : le nombre d'occurrences des séquences répétées va en décroissant selon que la longueur du répétable augmente.

longueur du répétable	nombre d'occurrences	pourcentage
1	9045	74,19 %
2	2141	17,56 %
3	694	5,69 %
4	217	1,78 %
5	58	0,48 %
6	24	0,20 %
7	10	0,08 %
8	3	0,02 %

Tableau 2. Répartition selon la longueur du répétable

Si l'on croise maintenant les deux paramètres de la longueur du répétable et du nombre de répétés, on obtient le tableau suivant qui présente la répartition en pourcentages.

Longueur du répétable	Nombre de répétés										
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Total**
1	63,18 %	8,60 %	1,85 %	0,32 %	0,15 %	0,05 %	0,02 %	0,00 %	0,01 %	0,01 %	74,19 %
2	16,26 %	1,14 %	0,12 %	0,03 %	0	0	0	0	0	0	17,55 %
3	5,48 %	0,20 %	0,01 %	0,01 %	0	0	0	0	0	0	5,70 %
4	1,74 %	0,04 %	0	0	0	0	0	0	0	0	1,78 %
5	0,48 %	0	0	0	0	0	0	0	0	0	0,48 %
6	0,20 %	0	0	0	0	0	0	0	0	0	0,20 %
7	0,08 %	0	0	0	0	0	0	0	0	0	0,08 %
8	0,02 %	0	0	0	0	0	0	0	0	0	0,02 %

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801386

Total	87,44 %	9,98 %	1,98 %	0,36 %	0,15 %	0,05 %	0,02 %	0,00 %	0,01 %	0,01 %	

Tableau 3. Répartition selon la longueur du répétable et le nombre de répétés

À la lecture de ce tableau, on peut faire trois remarques principales, qui reprennent partiellement ce qui a déjà pu être observé pour chaque paramètre séparément :

- en croisant nos deux paramètres, on constate que certaines configurations sont totalement absentes. Elles sont mêmes majoritaires sur l'ensemble du tableau puisque 57 des 80 possibilités ne sont pas actualisées dans notre corpus, qui comprend pourtant un nombre de séquences répétitions non négligeable ;

- les répétitions les plus fréquentes concernent majoritairement des répétables constitués d'un seul mot, répété une seule fois, puisque cette configuration représente 63,18 % du total des occurrences. Viennent ensuite les répétables composés de deux mots répétés également une seule fois avec 16,26 %. Ces séquences sont suivies par les répétables d'un mot répétés deux fois (8,60 %) puis par les répétables de trois mots répétés une seule fois (5,48 %). Ces quatre configurations constituent à elles seules 93,52 % de toutes nos occurrences de séquences répétitions ;

- les 18 autres configurations représentées dans le corpus totalisent chacune un nombre d'occurrences pratiquement insignifiant.

On le voit donc ici, apparait dans ces données croisées une très grande régularité dans la forme que prennent les séquences répétitions. Nous allons voir que cette régularité concerne également les répétables.

4.3.5 Distribution des répétables

Dans notre corpus, 1757 formes différentes de répétables sont concernées par le phénomène de la répétition, indépendamment du nombre de répétés.

4.3.5.1 Formes les plus fréquentes

Les 30 formes les plus fréquentes dans notre corpus sont les suivantes, présentées par ordre de fréquence décroissant :

oui (824), *de* (732), *c'est* (613), *je* (424), *des* (382), *les* (335), *et* (334), *ça* (317), *qui* (309), *non* (296), *le* (291), *à* (281), *un* (265), *on* (253), *euh* (233), *dans* (204), *mais* (187), *la* (177), *en* (159), *que* (157), *il* (151), *vous* (142), *il y a* (140), *une* (121), *ou* (119), *très* (105), *ils* (88), *si* (78), *du* (73), *ouais* (71).

On retrouve dans cette liste les adverbes *oui*, *non* et *ouais* déjà mentionnés ci-dessus. Ce qui saute aux yeux à la lecture de cette liste est que tous les items sont des monosyllabes appartenant à la métacatégorie des mots grammaticaux : pronom, déterminant, préposition, conjonction et adverbe. Si le *c'est* peut être analysé comme une suite PRO + V, on voit bien son fonctionnement comme une forme unique de présentatif. De la même manière, *il y a* n'est qu'une exception apparente à la catégorie des monosyllabes. Transcrit en une suite de trois mots graphiques, *il y a* est bien souvent prononcé par le locuteur en une seule syllabe [ja].

Ces 30 items les plus fréquents totalisent 7861 occurrences, soit près de 65 % des répétitions de notre corpus.

4.3.5.2 Les hapax

On a une majorité de répétables qui ne sont l'objet d'une répétition (quel que soit le nombre de répétés) qu'à une seule reprise dans le corpus.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801386

En effet, sur les 1757 répétables différents, 1303 sont des hapax, ce qui représente 74,1 % des répétables recensés. Ce chiffre peut étonner, mais en fait il va dans le sens de la distribution des vocables en corpus, et ceci indépendamment du phénomène de la répétition, objet qui nous intéresse ici. En effet, comme nous l'avons déjà mentionné, on a dans tout corpus 1) un ensemble composé d'un nombre restreint de formes qui apparaissent très fréquemment, 2) un second ensemble composé d'un nombre important de formes « rares », dans le sens où elles apparaissent une seule fois dans le corpus et 3) un 3e ensemble qui comprend des vocables dont la fréquence d'apparition est entre ces deux groupes (cf. la loi de Zipf, 1949). La distribution des répétables ne semble donc pas déroger à cette règle générale, qu'on peut vérifier quel que soit le type de corpus, même si le pourcentage d'hapax est ici plus élevé que celui des hapax calculé sur l'ensemble des formes du corpus.

Les hapax concernent aussi bien des répétables longs que des répétables courts. Néanmoins, chaque catégorie est représentée en proportions variables : comme l'on pouvait s'y attendre, tous les répétables les plus longs sont des hapax ($l \geq 6$) ; ensuite, la répartition entre hapax et non-hapax varie selon la longueur du répétable, mais quelle que soit *l*, les hapax sont toujours plus nombreux.

On a vu ci-dessus que les répétables les plus fréquents étaient des mots grammaticaux. Toutefois, on constate dans nos données que cette catégorie des mots grammaticaux n'est pas absente des hapax, qu'il s'agisse de prépositions (*devant*), d'adverbes (*absolument, aussi*), de déterminants (*ces, cet, tel*), d'onomatopées (*bah, bang, bouf, ouh*), de conjonctions (*quoique, puisque*) ou encore de pronoms (*cela, celui, ceux, eux, t', te*).

Parmi les hapax où le répétable est de longueur *l=1*, les mots lexicaux sont largement majoritaires. En effet, ils représentent 84,4 % des occurrences de cette catégorie.

Néanmoins, il ne faudrait pas penser que les répétables qui sont des mots lexicaux seuls (c'est-à-dire sans déterminant d'aucune sorte) appartiennent à la catégorie des hapax. En effet, nous avons 21 répétables mots lexicaux seuls qui apparaissent dans des séquences répétitions à plus d'une reprise. Parmi eux : 2 adjectifs (*dur* et *normal*), un nom commun (*monsieur*), un nom propre (*Charleroi*), et 11 verbes (13 conjugués et 4 à l'infinitif). Pour ce qui est des verbes conjugués, on observe une constante dans la liste qui suit : *dis, doit, donne, es, font, parlent, vont, attends, avait, peut, ont, faut et sont*. En effet, tous ces verbes sont des verbes très fréquents en corpus, indépendamment de leur mise en place dans une séquence répétition. De plus, hormis *attends*, ils sont tous monosyllabiques. En cela, ils rejoignent d'une certaine manière les mots grammaticaux.

Le chiffre de 74,1 % des hapax est élevé, mais ramené à la fréquence effective des répétitions dans le corpus, et non à l'ensemble des formes de répétables différentes, ces hapax constituent seulement 10,69 % des occurrences.

Le tableau suivant synthétise ces données. La première colonne est la longueur *l* du répétable, la deuxième donne le nombre de répétables hapax, la troisième indique le nombre de répétables différents faisant au moins deux fois l'objet d'une répétition[1], la quatrième colonne donne le pourcentage des occurrences d'hapax en corpus, la dernière colonne reprend la fréquence des occurrences de formes qui ne sont pas des hapax.

longueur *l* du répétable	**hapax n=**	**répétables apparaissant au moins 2 fois n=**	**Pourcentage des occurrences d'hapax en corpus**	**Pourcentage des occurrences de répétables apparaissant au moins 2 fois en corpus**
1	218	177	1,79 %	72,40 %
2	482	210	3,95 %	13,61 %

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801386

3	358	49	2,94 %	2,76 %
4	152	17	1,25 %	0,53 %
5	56	1	0,46 %	0,01 %
6	24	0	0,20 %	0 %
7	10	0	0,08 %	0 %
8	3	0	0,02 %	0 %
TOTAL	1303	454	10,69 %	89,31 %

Tableau 4. Tableau synthétique des hapax

4.3.6 Analyse morphosyntaxique des répétitions

Tout mot, appartenant à n'importe quelle classe grammaticale, peut théoriquement être l'objet d'une répétition. De plus, comme nous avons eu déjà l'occasion de le voir, le répétable peut aussi bien être un mot unique qu'une suite de plusieurs mots. Cette séquence peut constituer un groupe syntaxique complet ou celui-ci peut être inachevé.

Or, si tout mot et si tout type de séquence peut être répété, nous avons vu dans ce qui précède que certaines régularités émergeaient, qui concernent aussi bien le répétable que le(s) répété(s). Nous avons notamment mis en évidence à plusieurs reprises la forte fréquence des mots grammaticaux concernés par les répétitions.

Nous avons donc voulu analyser plus en détail les répétitions, en ce qui concerne leur catégorie grammaticale. Pour cela, nous avons analysé de manière approfondie un corpus plus petit, un peu plus de 50 000 mots, composé de sept enregistrements. Nous avons obtenu 1151 séquences répétitions. La distribution de celles-ci correspond à celle que nous avons observée pour la totalité de notre corpus :

- les répétables d'un seul mot sont de loin les plus fréquents, avec 73,3 % des occurrences ;
- les occurrences les plus nombreuses sont celles où le nombre *n* de répété égale 1. On obtient effectivement 88,8 % des cas sur ce schéma.

On constate donc une indéniable régularité dans le format général de la répétition, puisque ces pourcentages sont pratiquement identiques à ceux que nous avons obtenus sur notre corpus, où l'on avait 74,19 % de répétables d'un seul mot et 87,44 % de séquences répétitions où n=1. Cette régularité correspond également à celle observée sur le français par Blanche-Benveniste (2003) Candéa (2000a), Grosjean et Deschamps (1972), Henry (2001 et 2005) ou encore Jeanjean (1984).

Nous voudrions maintenant analyser de façon plus approfondie cette classe de répétables composés d'une seul mot.

4.3.6.1 Catégorie grammaticale des répétables

Le sous-corpus comprend 727 répétables où l=1. La ventilation de ces formes en fonction de leur classe grammaticale peut être synthétisée dans le tableau suivant :

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801386

	n=1	n=2	n=3	n=4	Total
Préposition	19,38 %	2,14 %	0,12 %	0,24 %	21,64 %
Prepdet	1,66 %	0,24 %	0,48 %	0,12 %	2,50 %
Déterminant	17,00 %	2,50 %	0,48 %	0,12 %	19,98 %
Pronom personnel	10,23 %	0,95 %	0,00 %	0,00 %	11,18 %
Pronom relatif	4,04 %	0,48 %	0,24 %	0,00 %	4,76 %
Pronoms divers	3,69 %	0,12 %	0,12 %	0,00 %	3,92 %
Conjonction de subordination	4,99 %	0,36 %	0,00 %	0,00 %	5,35 %
Conjonction de coordination	5,35 %	0,83 %	0,12 %	0,00 %	6,30 %
Onomatopée	0,48 %	0,12 %	0,12 %	0,00 %	0,71 %
Adverbe	15,46 %	3,45 %	0,24 %	0,12 %	19,26 %
Nom	0,71 %	0,00 %	0,00 %	0,00 %	0,71 %
Adjectif	0,12 %	0,00 %	0,00 %	0,00 %	0,12 %
Verbe conjugué	2,50 %	0,12 %	0,12 %	0,00 %	2,73 %
Verbe à l'infinitif	0,24 %	0,00 %	0,00 %	0,00 %	0,24 %
amorces	0,59 %	0,00 %	0,00 %	0,00 %	0,59 %
	86,44 %	11,30 %	2,04 %	0,60 %	100,00 %

Tableau 5. Catégorie grammaticale des répétables de longueur 1

Les données ci-dessus confirment la tendance (règle ?) que nous avons déjà mise en évidence à plusieurs reprises : lorsque le répétable est constitué d'un seul mot, il appartient dans la majorité des cas à la métacatégorie des mots grammaticaux. En effet, si on laisse les amorces de côté (dont il est parfois incertain de « prédire » la catégorie grammaticale du terme entier, cf. chap. 1 sur les conventions de transcription), seules 23,06 % des occurrences appartiennent à la catégorie des mots lexicaux. Il s'agit majoritairement, dans ce groupe, de répétables qui appartiennent à la classe des adverbes (19,26 %).

En ce qui concerne les mots grammaticaux, on voit que les positions de tête, en termes de fréquence, obtiennent pratiquement des scores équivalents : les prépositions sont légèrement les plus fréquentes (21,64 %), suivies par les déterminants (19,98 %).

Les autres catégories obtiennent des scores plus faibles. Néanmoins, les pronoms, que nous avons répartis dans le tableau en trois sous-catégories, totalisent un score équivalent aux deux catégories précédentes, avec 19,86 % des occurrences.

Nos résultats vont tendanciellement dans le même sens que ce que relevaient déjà certains chercheurs. Des divergences apparaissent néanmoins liées notamment à des facteurs déjà évoqués plusieurs fois qui tiennent à une manière différente d'effectuer les décomptes ou à une définition différente du phénomène. Ainsi, Blanche-Benveniste (1987) cite une étude de Jeanjean dans laquelle 97 % des répétitions concernent des pronoms clitiques sujets (50%), des prépositions (23 %), des articles (17 %) et des conjonctions (7 %).

Grosjean et Deschamps (1975) comptabilisent quant à eux 70 % de « mots grammaticaux ou groupes de mots grammaticaux ». On regrette que les auteurs ne donnent pas le détail chiffré dans la répartition des

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801386

mots grammaticaux, ni la distinction de catégorie grammaticale, entre déterminant et pronom par exemple pour *le, la* et *les*.

Candéa (2000a) montre que le phénomène concerne quatre fois plus souvent un mot outil (MO) qu'un mot plein (MP). Cette constatation l'incite à penser que

> les répétitions de MO en français oral ont un caractère systématique, alors que les répétitions de MP ont un caractère plutôt accidentel. (Candéa 2000a : 316)

Henry (2005), dans un travail centré exclusivement sur les répétitions, a analysé un corpus d'un million de mots qui fait intervenir 1200 locuteurs. Son étude quantitative sur 15 786 répétitions montre que les répétitions impliquant des mots outils sont 5 fois plus nombreuses que celles comportant un mot plein[2].

Henry (2002) obtient des scores différents des nôtres, avec à l'intérieur de la classe des mots grammaticaux, 41,5 % de déterminants, 26 % de pronoms et 13 % de prépositions. En fait, cette différence s'explique notamment par le fait que l'auteure ne tient pas compte uniquement des séquences où la longueur du répétable est de 1, mais aussi des séquences où le répétable est composé de plusieurs mots.

Grosjean et Deschamps, dans leur étude de 1972, relèvent :

> (…) le rôle majeur joué par les mots grammaticaux. Il faut aussi mentionner leur fréquence d'occurrence (où l'on relève par ordre d'importance les mots suivants : le/la/les, de, je, des, un, et, à, il[s])[3] est fortement corrélée (r = 0,80) avec la fréquence d'usage des mots français (…). (Grosjean et Deschamps 1972 : 154)

Cette remarque va dans le sens de celle de Morel et Danon-Boileau (1998 : 84) qui constatent :

> On ne sera pas surpris de constater que ceux-ci [les mots-outils les plus fréquemment répétés] se classent parmi les soixante-quinze premiers mots de la liste de fréquence du français fondamental (établie en 1950 par G. Gougenheim *et alii*).

Nous avons déjà remarqué ce fait lors de l'analyse des répétables les plus fréquents du corpus.

En ce qui concerne ces formes très fréquentes, nous avons voulu voir dans notre corpus comment se fait la répartition entre les formes possiblement ambigües dans les catégories grammaticales déterminant et pronom. Ainsi, nous obtenons des résultats comparables à ceux de Henry (2001) où les répétables *les, le* et *la* sont beaucoup plus fréquemment des déterminants que des pronoms, ce qui correspond également à la tendance observée pour ces formes ambigües lorsqu'elles ne sont pas répétées. En effet, Vergne et Giguet (1998) dénombrent seulement 2,24 % de pronoms sur 1054 occurrences de *le, l', la* et *les*.

4.3.6.2 Le répétable est une séquence de mots grammaticaux

Nous venons d'analyser les répétables composés d'un seul mot, majoritairement constitués de mots grammaticaux, principalement des prépositions et des déterminants.

Quand on y regarde de près, on constate que la plupart des répétables de deux mots contiennent également au moins un mot grammatical.

Si l'on prend l'exemple de la préposition *dans*, on a 204 occurrences où *dans* est répétable seul. À cela s'ajoutent les occurrences où il est suivi d'un déterminant : *dans ce, dans certains, dans cette, dans des* (4), *dans l'* (2), *dans la* (4), *dans le* (6), *dans les* (13), *dans leurs, dans ma* (2), *dans mes, dans mon* (2), *dans nos, dans notre, dans son* (3), *dans un* (5) et *dans une* (6). On a aussi des répétables plus longs comme *dans d'autres (pays) ou dans le même (billet).*

Pour l'anglais, Blankenship et Kay (1964), ainsi que Cook (1971), ont montré qu'en général, la répétition se fait à l'initiale du syntagme. C'est cette régularité de reprise à l'inititiale du syntagme qu'illustrent les exemples de *dans*.

On peut citer le même type de fonctionnement avec les autres prépositions *à, avec, entre, pour, etc.*

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801386

On le répète encore, le phénomène de la répétition est soumis à certaines contraintes : la répétition disfluente n'apparait pas n'importe où dans l'énoncé, et elle concerne de manière privilégiée les mots grammaticaux monosyllabiques (cf. Blanche-Benveniste 2003). Cette manière de reprise à l'initiale, qui semble relever de la systématicité, explique le petit nombre d'occurrences qui ne sont pas des mots grammaticaux que l'on observait dans les répétables de longueur 1. En fait, en général, un substantif n'est pas répété seul. S'il fait l'objet d'un problème de sélection lexicale, ce problème se manifeste par la répétition des mots-outils qui le précèdent. C'est donc tout le groupe syntaxique qui est repris à l'initiale, comme l'illustrent les exemples suivants :

> **ilcCF1** l'accent c'est la la tonalité que l'on met **sur les / sur les syllabes** [ilcCF1r]

> **ileGG1** **sur une / sur une classe** de de de quinze ou vingt que nous étions il y en avait peut-être un ou deux ou trois et encore c'était en pleine Ardenne (…)[ileGG1r]

Le phénomène est le même pour les verbes, où l'on remarque la reprise de la séquence à gauche de celui-ci :

> **ilpMJ1** (…) moi je je / je connais peu de jeunes qui qui parlent wallon de A à à Z / par contre qui utilisent des termes wallons ou des expressions et cetera je crois que **ça se ça se transmet** / exactement de la même manière (…) [ilpMJ1r]

Cela explique le nombre réduit de mots lexicaux seuls que l'on a dans nos données. Ce n'est pas sur le lexème seul que se manifeste le travail de formulation, mais sur toute l'unité dont il est la tête.

4.3.6.3 Continuité ou rupture syntaxique de l'énoncé

La répétition des mots-outils constitue-t-elle une rupture dans la construction syntaxique de l'énoncé ? En d'autres mots, la place syntaxique qu'ils ouvrent est-elle actualisée dans le discours ou la répétition du mot-outil marque-t-elle une auto-interruption de l'énoncé ? Selon Morel et Danon-Boileau (1998 : 85),

> La répétition multiple [n>1] et sans pause du mot-outil semble relever d'une gestion à très court terme de la formulation. Elle aboutit toujours à l'énoncé d'un mot.

Dans ce cas, il n'y aurait donc pas de rupture syntaxique, la construction ouverte par le mot-outil étant au final achevée. Pour ces cas, Henry (2005 : 85) parle d'un « vide lexical » momentané, et d'un « remplissage lexical qui s'effectue dans un second temps pour aboutir à un énoncé achevé ».

À l'analyse de notre corpus, nous avons pu observer que 5,7 % des répétitions étaient le lieu d'une rupture dans la construction syntaxique de l'énoncé.

Ainsi donc, dans la plupart des cas, la place ouverte par la répétition est achevée. Mais on constate des cas, comme celui mentionné ci-dessus, où la place est achevée, mais pas immédiatement. Dans le cadre de l'étiquetage morphosyntaxique, cela pose évidemment problème.

Le plus fréquemment, la rupture se produit après la répétition, comme dans l'exemple suivant :

> **ilrMS1** (…) je vois pas pourquoi pour euh pour aller **à à / chez** les Flamands ben il faudrait plus avoir son accent liégeois (…) [ilrMS1r]

Néanmoins, nous avons beaucoup d'occurrences où la rupture précède la répétition. En fait, la répétition marque le début d'une nouvelle construction syntaxique, laissant la séquence antérieure inachevée, comme dans :

> **norKJ1** (…) euh pf je je sais bien qu'il vaudrait mieux que mes enfants aient l'orthographe **de // qu'on / qu'on** souhaite voir // euh dans les journaux et sur les murs // hein [norKJ1r]

Dans cet exemple, la rupture est marquée par la pause longue. Mais ce n'est pas toujours le cas, comme le montre l'énoncé suivant :

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801386

> **norKJ1** alors là évidemment -| oui / hein oui bien sûr // euh les la société est pour le moment ainsi faite / que / que les les manquements // tant sur le plan grammatical // euh / à ces normes qui ne sont donc pas encore très très bien étudiées mais dont l'évidence s'impose néanmoins // **on les les** manquements sont sanctionnés // (…) [norKJ1r]

Ici, on a bien une pause longue dans le contexte gauche de la répétition, mais celle-ci ne la précède pas directement comme dans l'exemple précédent. La rupture de construction se produit sans aucune marque dans le texte : la construction entamée avec le pronom sujet *on* est laissée inachevée (la séquence verbale ne sera pas actualisée, en tout cas pas suite à un pronom sujet), pour repartir sur une nouvelle construction qui débute par la répétition du déterminant. Le sujet n'est plus ici pronominal mais nominal, et il débute par une répétition à l'initiale du syntagme.

On voit donc bien, à l'analyse d'un cas comme celui-ci, que les indices manquent pour déterminer sur des bases formelles les cas d'inachèvements de construction.

5 Conclusions

Si l'on s'est intéressée un peu longuement aux répétitions, c'est parce que ces disfluences sont particulièrement fréquentes dans le corpus ; les laisser telles quelles dans le texte diminuerait considérablement les performances de l'étiquetage morphosyntaxique.

Nous avons constaté que la répétition, telle que nous l'avons définie ici en tout cas, obéit à d'indéniables régularités, qu'il s'agisse du nombre de répétés, de la longueur du répétable ou encore des séquences mixtes, plus nombreuses que les répétitions contigües. Mais malgré les régularités observées et les tendances qui se dégagent, on est face à des structures complexes et à de nombreuses variations dans notre corpus.

Nous avons vu également que la répétition affecte principalement les mots grammaticaux (des déterminants, des prépositions et des pronoms) ou des suites de ceux-ci.

Néanmoins, malgré les régularités observées, nous échouons à trouver des indices qui nous permettraient de détecter les cas où la répétition est le lieu d'une rupture syntaxique dans la construction de l'énoncé.

Références bibliographiques

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

Blanche-Benveniste Cl. (1987). Syntaxe, choix de lexique et lieux de bafouillage. *DRLAV* 36-37, 123-157.

Blanche-Benveniste Cl. (2003). La naissance des syntagmes dans les hésitations et répétitions du parler. J.-L. Araoui (Éd), *Le sens et la mesure. Hommages à Benoît de Cornulier*. Paris : Honoré Champion, 40-55.

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

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

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

Candéa, M. (2000b). Les *euh* et les allongements dits "d'hésitation" : deux phénomènes soumis à certaines contraintes en français oral non lu. *Actes des 23es Journées d'Étude sur la Parole (JEP'2000)* (19-23 juin, Aussois).

Constant, M., Dister, A. (2012). Mots composés et disfluences. Dans Dister Anne, Longrée Dominique et Gérald Purnelle (Eds), *Actes des 11es JADT*, Liège, 269-280.

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

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801386

Duez D. (2001). Signification des hésitations dans la production et la perception de la parole spontanée. *Revue Parole* 17-18-19, 113-137.

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

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

Henry, S. (2001). Étude des répétitions en français parlé spontané pour les technologies de la parole. *Actes de RÉCITAL* (Nancy, 24-27 juin 2001).

Henry, S. (2005). Quelles répétitions à l'oral ? Esquisse d'une typologie. G. Williams (Éd.), *La Linguistique de corpus*. Rennes : Presses universitaires de Rennes, 81-92.

Henry, S., Campione, E., Véronis, J. (2004). Répétitions et pauses (silencieuses et remplies) en français spontané. *Actes des 25es Journées d'Étude sur la Parole (JEP'2004)* (19-22 avril, Fès), 261-264.

Henry, S., Pallaud, B. (2003). Word fragments and repeats in spontaneous spoken French. R. Eklund (Éd.), *Proceedings of DISS'03. Disfluency in Spontaneous Speech Workshop, (5-8 Septembre 2003, Göteborg University, Sweden), Gothenburg Papers in Theorical Linguistics* 90, 77-80.

Hindle, D. (1983). Deterministic parsing of syntactic non-fluencies. *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics*. Cambridge : Massachussetts, 123-128.

Jeanjean, C. (1984). Les ratés c'est fa fabuleux. Étude syntaxique des amorces et des répétitions. *LINX* 10, 171-177.

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

Maclay, H., Osgood, C**.** (1959). Hesitation Phenomena in Spontaneous English Speech. *Word*, 15 19-44.

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

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

Nelson, G. (1996). The Design of the Corpus. S. Greenbaum (Éd.) *Comparing English worldwide. The International Corpus of English*. Oxford : Clarendon Press, 27-35.

Nelson Gerald (1996b). Markup systems. S. Greenbaum (Éd.) *Comparing English worldwide. The International Corpus of English*. Oxford : Clarendon Press, 36-53.

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

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

Vergne J., Giguet, E. (1998). Regards théoriques sur le tagging. *Actes de TALN 1998* (10-12 juin, Paris).

Zellner, B. (1992). Le bé bégayage et *euh*… l'hésitation en français spontané. *Actes des 19es Journées d'Étude sur la Parole (JEP'92)* (19-22 juin, Bruxelles), 481-487.

---

[1] Les chiffres des colonnes 2 et 3 donnent le nombre de répétables différents (qui apparaissent une seule fois ou plusieurs fois) et non le nombre d'occurrences en corpus. Ces données ne font pas de distinction en fonction du nombre de répétés.

[2] L'auteure ne s'attache qu'aux répétables qui totalisent aux moins 100 occurrences. Nous avons vu que les hapax étaient plus nombreux parmi les mots lexicaux, ce qui explique peut-être ses résultats.

[3] Dans leur étude de 1973, les mots grammaticaux les plus répétés sont, par ordre de fréquence décroissante : *qui, le, la, les, de, un(e), sur, sa, au, à.*