Etude des répétitions en français parlé spontané pour les technologies de la parole

Sandrine Henry

Équipe DELIC – Université de Provence
29, Av. Robert Schuman, 13621 Aix-en-Provence Cedex 1
sandrine_henry@hotmail.com

Mots-clefs – Keywords

Répétitions, français parlé spontané, « disfluences », phénomènes de performance, étude quantitative, reconnaissance de la parole, étiquetage morpho-syntaxique.

Repetitions, spontaneous French speech, disfluencies, performance phenomena, quantitative study, speech recognition, part-of-speech tagging.

Résumé – Abstract

Cet article rapporte les résultats d'une étude quantitative des répétitions menée à partir d'un corpus de français parlé spontané d'un million de mots, étude réalisée dans le cadre de notre première année de thèse. L'étude linguistique pourra aider à l'amélioration des systèmes de reconnaissance de la parole et de l'étiquetage grammatical automatique de corpus oraux. Ces technologies impliquent la prise en compte et l'étude des répétitions de performance (en opposition aux répétitions de compétence, telles que *nous nous* sujet + complément) afin de pouvoir, par la suite, les « gommer » avant des traitements ultérieurs. Nos résultats montrent que les répétitions de performance concernent principalement les mots-outils et apparaissent à des frontières syntaxiques majeures.

This article is a report of a quantitative study of repetitions based on a corpus of a one-million-word spontaneous spoken French, conducted during the first year of our PhD thesis. This linguistic study can contribute to the improvement of speech recognition and spoken French part-of-speech tagging. Improvement of these technologies requires taking into account and studying performance repetitions (such as complement + complement *nous nous*) in order to be able to "erase" them before further processing. Our results show that repetitions mainly involve function words and take place at major syntactic boundaries.

1 Introduction

Les technologies du T.A.L. sont difficiles à transposer directement à l'oral spontané, non préparé. Ainsi, par exemple, l'étiquetage morpho-syntaxique est relativement bien maîtrisé sur l'écrit pour des langues telles que l'anglais ou le français, puisque des résultats supérieurs à 95% d'étiquettes correctes sont couramment publiés. Cependant, l'étiquetage de corpus oraux pose des problèmes bien plus épineux, et l'on ne dispose à l'heure actuelle de pratiquement aucun corpus oral de taille significative morpho-syntaxiquement étiqueté pour le français. En effet, tout énoncé oral spontané, conserve les traces de son élaboration à travers des phénomènes de performance, tels que la répétition, l'autocorrection, l'allongement de la finale, etc., qui constituent de précieux indices susceptibles d'éclairer le fonctionnement de la langue, mais constituent autant de points d'achoppement pour les technologies dérivées de l'écrit (cf. Valli, Véronis, 1999). Par ailleurs, les technologies de reconnaissance de la parole (dictée vocale, etc.) ont essayé de façon constante au fil des années de permettre de plus en plus de naturel et de souplesse aux locuteurs. On est ainsi passé progressivement de la reconnaissance de mots isolés mono-locuteur à la reconnaissance de parole continue multi-locuteurs (avec pour l'instant des contraintes : ambiance non-bruitée, etc.). Pour autant, les systèmes de reconnaissance de la parole sont encore bien loin de la reconnaissance d'un flot continu de parole faisant intervenir des phénomènes de performance nombreux, caractéristiques de l'oral non contraint et non préparé. Pourtant, l'accès à ce type d'élocution serait une valeur ajoutée non négligeable pour les technologies concernées, avec, à la clé, un marché potentiel extrêmement important.

Pendant nombre d'années, la tradition grammairienne, ainsi que celle du T.A.L., semblent avoir délaissé, voire dévalorisé au profit de l'écrit normatif, l'étude des phénomènes propres à l'oral, ne développant aucun cadre syntaxique pour les analyser. Les travaux les plus anciens sur l'oral appartiennent donc aux psycholinguistes ((Maclay, Osgood, 1959), (Levelt, 1983), (Shriberg, 1994), etc.) qui ont envisagé ces « disfluences » comme un moyen privilégié pour délimiter les étapes de la production langagière, dans le but d'établir des modèles de performance dans la problématique de l'encodage/décodage.

Dans le cadre de notre thèse, nous nous intéressons plus particulièrement à l'étude et à la détection automatique d'un phénomène de performance particulier, celui de la répétition. (Blankenship & Kay, 1964) ont démontré que certains types de reprises tendent à redémarrer (du moins en anglais) à partir de l'initiale du syntagme, validant ainsi la pertinence des unités abstraites de la linguistique dans le processus d'encodage. Des travaux plus récents (Candéa, 2000) semblent démontrer que (sur le français) la fréquence de répétition est directement corrélée au type d'unité impliquée : il arrive presque six fois plus souvent qu'un mot-outil (MO) soit répété qu'un mot plein (MP). Ces résultats sont toutefois obtenus à partir d'un corpus restreint (environ 10 000 mots) et d'un type particulier : celui des corpus d'enfants.

Il nous a donc semblé intéressant de réaliser une étude quantitative des répétitions en prenant appui sur un corpus oral de plus grande taille (environ 1 million de mots) faisant intervenir de très nombreux locuteurs (environ 1200), ce qui nous a permis de dégager des tendances générales, c'est-à-dire d'échapper aux spécificités langagières individuelles.

Notre but est de répondre aux interrogations suivantes :

1. La répétition touche-t-elle indifféremment les mots-outils et les mots pleins ?

2. En explorant plus avant chacune de ces deux catégories, comment sont alors réparties les répétitions ? Sont-elles distribuées de manière aléatoire ?

3. Est-ce qu'un examen plus minutieux sur certaines formes ambiguës, pouvant appartenir à deux classes syntaxiques différentes, permet de dégager une quelconque tendance ?

4. Sachant que les répétitions peuvent aussi bien renvoyer à des cas à répété unique (*le le lapin*) qu'à des cas à répétés multiples (*la la la la fleuriste*), comment ces répétitions directes sont-elles réparties en fonction du nombre de répétés ?

5. Dans quelle proportion les répétitions associées, c'est-à-dire celles qui intègrent entre le répétable et le répété d'autres marques du « travail de formulation » [1], ou encore n'importe quel mot, apparaissent-elles ?

2 Précisions terminologiques

2.1 Les répétitions de performance : vers une définition

Du point de vue de la stricte séquence de surface des unités linguistiques, on peut observer des mots ou des séquences identiques qui se succèdent, mais leur statut linguistique peut être différent. La répétition résulte parfois des règles propres à la langue (rencontres syntaxiques, reprise pronominale, emphase, etc.), comme le montrent les exemples suivants :

ensuite je vous dis **nous nous** sommes vus le jour de la visite de l'école [BUSabcd][2]

il a fallu passer la ligne de démarcation -[3] et on l'a passée euh ma mère est passée toute seule avec un de mes frères - et **nous nous** avons pris le train [EVACUAT]

Nous parlerons dans ce cas de **répétition de compétence**. Dans d'autres cas, et ce sont justement ceux que nous cherchons à discriminer, la répétition est une marque du « travail de formulation » :

tout à l'heure tu disais que la syntaxe n'était pas la même à l'oral qu'à l'écrit et donc cela peut **nous nous** poser un problème pour les exposés parce que on s'en tient à l'écrit euh à nos notes quoi [33VOIL]

Nous nommerons cette dernière **répétition de performance**[4]. La discrimination entre les

---

[1] Terme emprunté à (Morel, Danon-Boileau, 1998).

[2] L'ensemble des exemples fournis sont extraits de *Corpaix*. Nous avons précisé le nom du corpus dont est issu l'exemple entre crochets.

[3] Les conventions de transcription établies par le G.A.R.S. prévoient le tiret pour marquer la présence d'une pause silencieuse.

[4] Pour simplifier, lorsque nous utilisons dans la suite le terme « répétition » seul, nous entendons par là « répétition de performance ».

deux types de répétitions est délicate d'un point de vue automatique. Il suffit par exemple pour s'en convaincre d'avoir utilisé des correcteurs orthographiques qui ont une tendance caractérisée à présenter comme erronées des séquences répétées quelles qu'elles soient. Le problème est rendu encore plus complexe par le fait que la répétition n'est pas toujours *verbatim,* et concerne parfois plusieurs unités consécutives :

il faut arriver à réaliser une harmonie entre les deux euh sans cela **le la le** dialogue ne peut exister [INTERWY]

on payait déjà assez donc euh je vois pas pourquoi ils **nous ils nous** demandaient encore ça [GENVE]

Les psycholinguistes ((Levelt, 1983), etc.) ont proposé un cadre d'analyse qui est celui d'un « gommage » d'un phénomène de performance, qui permettrait de réduire les énoncés à un oral « propre », proche de l'écrit[5]. Nous reprenons ici la terminologie proposée par (Shriberg, 1994) afin de présenter ici un cas de répétition de performance (**Erreur ! Source du renvoi introuvable.**).

- Le reparandum (RM) :désigne une partie ou la totalité de la séquence qui sera abandonnée au profit du *repair*.
- L'interruption point (IP) :établit la frontière finale du *reparandum*.
- L'interregnum (IM) : désigne la région comprise entre la frontière finale du *reparandum* et la frontière initiale du *repair*. L'interregnum peut contenir un terme d'édition (*editing term*) qui peut être réalisé par une pause remplie ou encore un commentaire épilinguistique.
- Le repair (RR) :représente la partie réparée/corrigée du *reparandum*.

c'est pour ça que je vois pas l'intérêt d'ailleurs de la	•	< >	de la réforme orthographique	
(31TYT, 3, 3-4) RM	▲	IM	RR	
	IP			

Figure 1 : Exemple de répétition de performance

La répétition correspond au cas où la « zone achevée » de l'énoncé (le *repair*) ne présente **aucune correction** par rapport à la « zone en devenir » qu'est le *reparandum*.

Un autre modèle de représentation du phénomène, bien connu des syntacticiens mais encore peu exploité dans le domaine du T.A.L., est la « mise en grille » développée par (Claire Blanche-Benveniste, 1987). Contrairement au modèle précédent qui implique une prise de décision (gommage du phénomène jugé « disfluent »), la représentation en grilles présente l'avantage d'être plus neutre en ce sens qu'elle permet de visualiser à la fois les répétitions de performance (Figure 2) mais également celles qui semblent délibérées : répétitions intensives, etc. (Figure 3). L'intérêt réside ici en une représentation de l'architecture syntaxique des énoncés en suivant un cadre d'analyse unifié.

---

[5] Ce cadre, qui ramène finalement l'oral à une norme (écrite) hypothétique est très discutable, et sa critique (preuves empiriques à l'appui) fait l'objet du travail futur dans le cadre de notre thèse.

Les figures 2 et 3 montrent bien que la répétition, par opposition à d'autres marques du « travail de formulation », brise l'avancée syntagmatique, créant ainsi un « piétinement » sur un même emplacement syntaxique, au profit d'une ouverture de l'axe paradigmatique.

mais euh là je me suis inscrite	**à**	
	à l'école de voile	**de**
		de mon entreprise [CLAIRE]

Figure 2 : « Mise en grille » de répétitions de performance

alors il y a des oliviers qui sont **très**		
	très	
	très vieux [JERUSALEM]	

Figure 3 : « Mise en grille » d'une répétition « stylistique »[6]

2.2 La classification des unités lexicales

Un autre aspect méthodologique qu'il nous faut également préciser avant d'aller plus loin dans notre réflexion est la classification des unités lexicales. Il nous faut distinguer en effet les mots qui ont une charge lexicale pleine (les mots pleins) de ceux qui participent simplement de la structuration de la langue (les mots-outils). Cette distinction mot-plein/mot-outil est largement utilisée en T.A.L. (recherche d'information, etc.), mais il est important de noter que les catégories proposées pour les mots-outils de l'écrit ne recouvrent que très imparfaitement les unités de l'oral. Nous avons retenu la méthode de classification initialement forgée par (Morel, Danon-Boileau, 1998), mais nous ne l'avons reprise que partiellement en excluant la classe des ligateurs. Nous ne sommes malheureusement pas en mesure, étant donné la taille de notre corpus, de traiter individuellement tous les connecteurs et conjonctions. Précisons aussi que nous avons exclu les interjections de type *euh*, *bé*, *mm*, etc., qui correspondent à des marques du « travail de formulation ». La Figure 4 présente notre classification.

MO
Déterminants
prépositions et locutions prépositionnelles
verbes auxiliaires, attributifs, opérateurs, supports et copulatifs
pronoms personnels, démonstratifs, possessifs, relatifs et indéfinis
conjonctions et locutions conjonctives
adverbes de négation *ne* et *pas*
introducteur de rhème *c'est*

MP
noms
adjectifs
verbes
adverbes et locutions adverbiales
onomatopées et interjections
syntagmes et débuts de syntagmes nominaux ou verbaux

Figure 4 : Classification des unités

[6] Concernant cet exemple, seul l'accès au son nous permet d'affirmer qu'il s'agit d'une répétition « stylistique ».

2.3 Vers la réalisation d'une typologie des répétitions de performance

Afin d'établir une classification des répétitions, nous proposons de distinguer, d'une part, la répétition dite « simple » de la répétition « multiple », et d'autre part, la répétition « directe » de la répétition « associée » à la pause silencieuse ou remplie, ou encore à un mot quelconque. La répétition « simple » correspond au bloc formé du *répétable* suivi d'un seul et unique *répété*. *A contrario*, toute répétition dont le répétable est repris au moins deux fois sera qualifiée de « multiple ». Lorsque le répétable est suivi immédiatement par son répété sans aucune autre marque du « travail de formulation » ou de tout autre élément, la répétition est dite « directe », sinon elle est dite « associée ». Dans la schématisation que nous proposons (Figure 5), R_0 représente le répétable et R_1 R_2 … R_n les répétés. Les parenthèses, optionnelles, indiquent que la pause peut apparaître indifféremment entre le répétable et le répété ou entre les répétés. Nous envisageons, dans notre thèse, d'intégrer également une étude des contextes gauche et droit des répétitions.

	Directe	**Associée**		
		pause silencieuse	pause remplie	mot
Simple	R_0 R_1	R_0 + R_1	R_0 *euh* R_1	R_0 m R_1
Multiple	R_0 R_1 R_2 … R_n	R_0 (+ ou *euh* ou m) R_1 (+ ou *euh* ou m) R_2 … (+ ou *euh* ou m) R_n		

Figure 5 : Schématisation des différents types de répétitions

3 Méthode de travail

3.1 Présentation de notre corpus de travail

Notre corpus de travail comporte 1 000 382 mots. Nous l'avons constitué à l'aide de *Corpaix*, un corpus établi au cours de ces vingt-cinq dernières années par le G.A.R.S.[7] (actuellement équipe D.E.L.I.C.[8]), qui correspond à un recueil de paroles spontanées, auquel nous avons ajouté de nouveaux enregistrements personnels. Contrairement à certains corpus « orientés » tels que AMEX, SWBD, British National Corpus, etc., où la situation de communication avait été déterminée au préalable (échanges téléphoniques avec le personnel de compagnies aériennes pour obtenir des renseignements, ou encore exécution par le locuteur de certaines consignes préalablement fournies), nous proposons des situations d'enregistrement extrêmement variées où le locuteur adulte produit un oral totalement spontané.

Les enregistrements ont été transcrits, pour la plupart, par des étudiants en linguistique, puis vérifiés par des transcripteurs expérimentés. Ces transcriptions ont été réalisées en suivant les conventions de transcription initialement établies par le G.A.R.S., puis reprises par le groupe D.E.L.I.C. Pour l'instant, notre corpus ne présente pas un alignement texte/son.

---

[7] *Groupe Aixois de Recherche en Syntaxe*.

[8] *DEscription Linguistique Informatisée sur Corpus*.

3.2 Détection automatique du phénomène

Dans un premier temps, nous avons écrit un programme en langage *Perl* qui détecte les phénomènes de répétition « brute », de « surface », ce qui correspond, du point de vue de la programmation, à tout mot[9] suivi immédiatement par lui-même ou bien séparé de sa répétition par un autre mot. Nous avons ensuite appliqué ce programme à notre corpus de travail à l'intérieur duquel nous avons supprimé tous les séparateurs de corpus, les notes de bas de page, les signes typographiques marquant les allongements vocaliques ainsi que toutes les notations indiquant les tours de parole entre les différents locuteurs (notées L1, L2, etc.).

Le résultat constitue un réservoir de 29 948 répétitions brutes contenant à la fois des répétitions qui apparaissent comme des faits de langue (par exemple, ***peu à peu***, ou encore ***nous nous*** *sommes baignés*) dans lesquelles nous incluons les répétitions « stylistiques », mais également des répétitions de performance, aussi a-t-il été nécessaire de trier les données obtenues afin de ne garder que ces dernières.

3.3 Filtrage manuel des données

Etant entendu que le traitement informatique des données n'a réalisé qu'une détection de surface du phénomène ici étudié, il nous a donc fallu procéder à une analyse plus fine de ces 29 948 cas. Nous ne pouvons bien évidemment pas détailler ici l'ensemble des cas de figure rencontrés (Henry, 2001), mais simplement pointer quelques difficultés.

A la lecture des énoncés suivants, il apparaît clairement qu'on ne peut retenir de telles répétitions comme phénomènes de performance puisqu'il s'agit ici de dislocations :

la population très certainement va se rendre euh pfff euh nombreuse lors des grandes fêtes qui vont-avoir lieu à Marseille comme dans toutes les villes pour le bi-centenaire bien entendu mais **nous Marseillais nous** resterons attachés particulièrement à nos marins-pompiers [BIRNIE]

moi je préfère les garçons et **elle elle** préfère les filles [CIRILL1]

ou de la reprise obligatoire du pronom dans les constructions pronominales :

ah ben **nous nous** sommes baignés dans la mer Morte [JERUSALEM]

ou encore de la succession de deux unités syntaxiques différentes :

alors la danse fait beaucoup pour **nous nous** aimons danser [APOSTROPHE]

Dans ces cas, le déroulement syntagmatique n'est aucunement interrompu.

---

[9] Dans notre programme, un mot est défini comme un caractère (une lettre, un chiffre ou un souligné) ou une suite de caractères séparé(s) par un espace.

4 Résultats

Premier point, nous avons pu ainsi dégager de notre corpus de travail 15 786 répétitions de performance. Précisons qu'afin d'éviter de traiter les *hapax* et les cas quasiment isolés, nous avons écarté, dans notre étude, les formes dont le nombre total d'occurrences est inférieur ou égal à 100 ; notre seuil d'intégration est donc fixé à 10^{-4}. Le tableau de la Figure 6 rapporte les fréquences relatives des différentes unités lexicales (MO, MP, MO/MP) entrant dans une répétition de performance. Ajoutons qu'étant donné la taille de notre corpus, il ne nous a pas été possible de définir la catégorie d'appartenance pour certaines formes qui, alternativement, en fonction du contexte, se comportent comme des MO ou des MP. Il s'agit principalement de verbes copules (avoir, être…) et de certaines conjonctions qui peuvent aussi parfois être des noms ou des adverbes.

	%	**Fréquence Absolue**	**Nombre total d'occurrences**
MO	3,26	14 534	446 378
MP	0,72	430	59 914
MO/MP	1,24	822	66 375

Figure 6 : Fréquences relatives des unités lexicales faisant l'objet d'une répétition de performance

Nos résultats corroborent dans l'ensemble ceux donnés par Candéa (2000) qui a montré que les MO font plus fréquemment l'objet d'une répétition (2,94%) que les MP (0,56%). Effectivement, on remarque, d'après nos fréquences, qu'il arrive presque 5 fois plus souvent qu'un MO soit répété par rapport à un MP et Candéa a trouvé, quant à elle, un rapport de 6.

Deuxième point, les répétitions ne touchent pas indifféremment les différentes classes de mots. La répartition (en %) par classe des 14 534 répétitions impliquant des MO et des 430 répétitions impliquant des MP est présentée dans le tableau suivant (Figure 7).

	Classe	**%**		**Classe**	**%**
MO	déterminants	41,5	**MP**	adverbes et loc. adv.	52,0
	pronoms	26,0		adjectifs	25,0
	prépositions et loc. prép.	13,0		Verbes	9,0
	introducteur de rhème	9,5		Noms	8,5
	conjonctions et loc. conj.	7,0		syntagmes nominaux	3,5
	blocs de MO	2,0		syntagmes verbaux	2,0
	verbes auxiliaires, copulatifs, etc.	1,0			

Figure 7 : Répartition par classe des répétitions

Nous pouvons vérifier à la lumière de ces résultats que l'ouverture de l'axe paradigmatique n'a pas lieu sur n'importe quelle place de la structure syntaxique puisqu'on observe que ce sont les mots-outils se trouvant à l'initiale de syntagmes qui sont majoritairement répétés. Il

s'agit des déterminants (41,5%), pronoms (26%), prépositions (13%). Dans le cadre de notre étude, l'introduction d'une fonction syntaxique et l'apparition de la répétition de performance semblent donc être étroitement liées.

Troisième point, pour certaines formes, il nous a semblé intéressant de fournir une estimation plus fine. Ainsi, nous proposons ci-dessous (Figure 8) une nouvelle répartition des taux de répétition pour les formes ambiguës *le*, *la*, *les* et *leur* selon qu'elles renvoient à des déterminants ou à des pronoms personnels compléments.

Formes	**Déterminants**	**Pronoms Personnels**
les	6,34	1,16
le	5,64	1,33
leur	4,62	0,88
la	2,76	0,75

Figure 8 : Taux de répétition des formes *les*, *le*, *leur*, *la* en fonction de leur classe

En comparant ces taux de répétition, il apparaît très nettement que les déterminants font beaucoup plus fréquemment l'objet d'une répétition que les pronoms clitiques compléments. Quelle hypothèse avancer ? La psycholinguistique pourrait certainement nous aider. Il faudrait admettre que la répétition est un moyen de suspendre son discours jusqu'à ce que l'on soit capable de le poursuivre, ou encore de ne pas rompre l'acte de communication. Dans le cas de la répétition du déterminant, c'est l'accès à l'information lexicale qui semble poser problème. On ne peut pas rencontrer cette difficulté pour les clitiques compléments, mis en place très tôt dans la programmation du syntagme verbal. Quatrième point, il nous a également semblé pertinent de proposer une répartition des répétitions en prenant en compte le nombre de répétés (Figure 9).

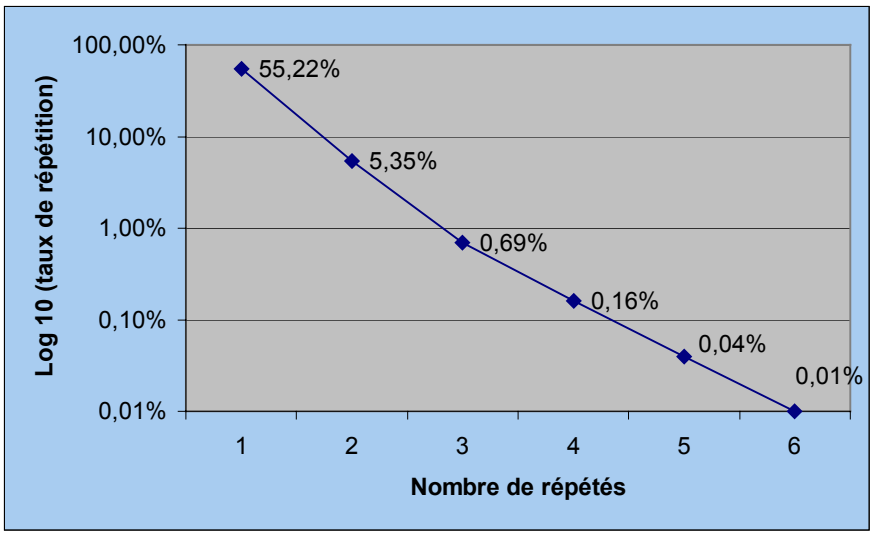

Figure 9 : Répartition des répétitions en fonction du nombre de répétés

On constate d'après ce graphique que les répétitions simples (répété unique) et directes sont largement prépondérantes et que la probabilité d'avoir une répétition diminue pour un nombre

de répétés élevé. On peut donc affirmer qu'il existe en définitive très peu de répétitions incluant plus de deux répétés. Ces résultats acquièrent une totale validité dans le cadre de notre étude. Cinquième point, si l'on s'intéresse maintenant à la répartition des répétitions associées (Figure 10), là encore, une tendance particulière se dégage : les répétitions contenant n'importe quel mot entre le répétable et le répété sont plus fréquentes que celles incluant une pause silencieuse ou remplie.

	RmR	**R + R**	**R *euh* R**
%	28,35	6,56	3,61

Figure 10 : Répartition des répétitions associées

5 Perspectives

Dans le contexte technologique actuel, nous ne pouvons procéder qu'à une détection de surface des phénomènes de répétition de performance, dont on ne peut évidemment pas se satisfaire. L'élaboration d'outils informatiques, qui permettraient une détection plus fine intégrant une composante syntaxique, fait partie de notre plan de travail futur, ainsi que l'étude de l'interaction avec les autres paramètres de la parole, pauses, pauses remplies, mouvements mélodiques, qui, dans certains cas, sont nécessaires à la détermination du type de répétition.

Références

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

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

Candéa M. (2000), Contribution à l'étude des pauses silencieuses et des phénomènes dits « d'hésitation » en français oral spontané. Etude sur un corpus de récits en classe de français, Thèse d'Etat, Université Paris III (Sorbonne Nouvelle).

Henry S. (2001), Etude quantitative des répétitions marques du travail de formulation en français oral spontané, D.E.A., Université de Provence (Aix-Marseille 1).

Levelt W.J.M. (1983), Monitoring and self-repair in speech. *Cognition,* Vol. 14, pp.41-104.

Maclay H., Osgood C.E. (1959), Hesitation phenomena in spontaneous English speech, *Word*, Vol. 15, pp.19-44.

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

Shriberg E.E. (1994), *Preliminaries to a Theory of Speech Disfluencies, Unpublished Ph.D. Thesis*, Department of Psychology, University of California, Berkeley.

Valli A., Véronis J. (1999), Etiquetage grammatical de corpus oraux : problèmes et perspectives, *Revue Française de Linguistique Appliquée,* Vol. *IV*(2), pp. 113-133.