# Annotation des disfluences dans les corpus oraux

Marie PIU, Rémi BOVE
Équipe DELIC – Université de Provence
29, Av. Robert Schuman, 13621 Aix-en-Provence Cedex 1
remi.bove@up.univ-mrs.fr, piumarie@yahoo.fr

**Résumé.** Les disfluences (répétitions, amorces, autocorrections, constructions inachevées, etc.) inhérentes à toute production orale spontanée constituent une réelle difficulté en termes d'annotation. En effet, l'annotation de ces phénomènes se révèle difficilement automatisable dans la mesure où leur étude réclame un jugement éminemment interprétatif. Dans cet article, nous présentons une méthodologie applicable à l'annotation des disfluences (ou « phénomènes de production ») que l'on rencontre fréquemment dans les corpus oraux. Le fait de constituer un tel corpus de données annotées, permet non seulement de représenter certains aspects pertinents de l'oral (de manière à servir de base aux observations et aux comparaisons avec d'autres données) mais aussi d'améliorer in fine le traitement automatique de l'oral (notamment l'analyse syntaxique automatique).

**Abstract.** Disfluencies (repeats, word-fragments, self-repairs, aborted constructs, etc.) inherent in any spontaneous speech production constitute a real difficulty in terms of annotation. Indeed, the annotation of these phenomena seems not easily automatizable, because their study needs an interpretative judgement. In this paper, we present a methodology for the annotation of disfluencies (also named "production phenomena") which frequently occur in speech corpora. Constituting such data allows not only to represent some relevant aspects of speech productions (so as to be a basis for observations and comparisons with other data), but also to improve automatic speech processing (particularly for parsing).

**Mots-clés :** corpus oraux, annotation, disfluences, prosodie, XML.

**Keywords:** speech corpora, annotation, disfluencies, prosody, XML.

## 1 Introduction

À l'heure actuelle, les études linguistiques qui basent leurs descriptions sur de vastes corpus électroniques gagnent sans cesse du terrain et les outils d'analyse automatique de plus en plus performants se multiplient. Malgré cela, l'analyse automatisée de l'oral reste marginale car on

ne dispose que de très peu de corpus oraux[1]. La constitution et l'annotation de corpus oraux représentent un enjeu de première importance en vue d'applications telles que reconnaissance vocale, apprentissage des langues, etc.

Par ce travail, nous proposons de constituer des données de référence annotées pour l'oral par le biais d'un schéma d'annotation et d'un formalisme adaptés. L'objectif de ce travail est double : il permet d'une part, d'obtenir des données de référence sur l'oral et d'autre part, de faciliter l'exploitation informatique de ces mêmes données. Pour mener à bien ce projet, nous avons jugé nécessaire de fonder notre analyse sur un cadre théorique existant qui traite des phénomènes de l'oral. En effet, pour garantir sa cohérence, il est indispensable que ce schéma d'annotation soit en adéquation avec un modèle d'analyse de l'oral préalablement défini. Nous nous sommes donc inspirés du modèle d'analyse de « la mise en grille » proposé par (Blanche-Benveniste, 1987), pour annoter les phénomènes de production dans notre corpus.

À notre sens, ce travail doit aussi lier analyse qualitative et quantitative pour rendre compte des multiples aspects du langage que l'on peut observer dans les situations d'oral. L'analyse qualitative se traduit par la mise en place d'un schéma d'annotation et par l'enrichissement des données. Ensuite, les informations quantitatives recueillies permettent de mieux connaître les mécanismes de la langue orale par les informations ponctuelles qui s'en dégagent (fréquence, répartition des phénomènes) mais aussi par les informations de structures (patrons récurrents, contexte étudié).

# 2 Corpus d'étude et phénomènes étudiés

Le corpus à partir duquel nous avons procédé à l'annotation des disfluences est une sous-partie du Corpus de Référence du Français Parlé (CRFP) constitué par l'équipe DELIC (Description Linguistique Informatisée sur Corpus) et spécialement choisie pour son caractère monologique et son hétérogénéité situationnelle.

## 2.1 Corpus de Référence du Français Parlé

Cette sous-partie du CRFP se compose de dix enregistrements (environ 53 minutes de parole, soit plus de 8000 mots) faisant intervenir cinq hommes et cinq femmes. La transcription orthographique a été effectuée entièrement à la main par des experts linguistes avec un cycle de réécoute/validation extrêmement strict. Les conventions de transcription adoptées (cf. Équipe DELIC, 2004) ne contiennent aucun trucage orthographique (du type p'tit, y'a, etc.) ni aucune ponctuation, suivant la tradition de l'équipe. Par ailleurs, un certain nombre de phénomènes de production à l'oral ont été transcrits (sans être annotés spécifiquement) : les répétitions, les amorces, les euh d'hésitation, les allongements, les pauses, les accents ainsi que les mouvements intonatifs majeurs.

[1] On entend généralement par « corpus oraux » les annotations (sous forme de transcriptions orthographiques) et les enregistrements (fichiers sons et/ou vidéo)

## 2.2 Phénomènes étudiés

Nous avons choisi de nous intéresser plus particulièrement à l'annotation des phénomènes de production à l'oral. Voici les phénomènes qui ont constitué notre objet d'étude et pour lesquels nous avons mis en place un schéma d'annotation :

- Les **répétitions** : répétition d'un ou plusieurs mots ou reprise à l'identique d'une syllabe, d'un mot ou d'une amorce de mot, de plusieurs syllabes ou de plusieurs mots, sans aucune valeur sémantique (Candéa, 2000).

  (1) *on entreposait **les: les: les** huiles ↗ (CRFP)*

- Les **autocorrections** : substitution d'un mot ou d'une série de mots à d'autres afin de modifier ou corriger une partie de l'énoncé (Kurdi, 2003).

  (2) *à cette époque **j'avais j'étais** en maîtrise il me restait le mémoire à faire ↗ (CRFP)*

- Les **amorces** : interruption de morphème en cours d'énonciation (Pallaud, 2002)

  (3) *donc je suis restée trois mois en **c-** en camping à peu près hein (CRFP)*

- Les **inachèvements** : énoncés auxquels il manque un ou plusieurs éléments pour qu'ils soient grammaticalement bien formé et interprétable sémantiquement (Kurdi, 2003).

  (4) *j'étais au bord de la mer c'était super ça été <u>un moment de:</u>*

- Les **disfluences combinées** : association simultanée d'au moins deux des phénomènes présentés ci-dessus.

  (5) *je voyais p- il y av- j'avais pas d'autre so- enfin j'avais pas d'autre solution ↘*

# 3 Cadre d'analyse des disfluences

Pour mener à bien notre projet et pour garantir la cohérence de notre méthode d'annotation, nous avons jugé nécessaire d'appuyer notre analyse sur un cadre théorique existant.

L'une des approches les plus répandues concernant la modélisation des disfluences est celle proposée par Shriberg (1994). L'auteur décrit l'organisation interne des disfluences en un ensemble d'espaces distincts délimitant les étapes de la production orale. Le *reparandum* (RM) correspond à la partie qui sera abandonnée au profit de la réparation (*repair*). Le *point d'interruption* (PI) établit la frontière finale du *reparandum* et marque une rupture dans la fluidité du discours. L'*interregnum* (IM) désigne la région comprise entre la frontière finale du *reparandum* et la frontière initiale du *repair*. Enfin, le *repair* (RR) représente la partie corrigée du *reparandum* et marque le retour à la « fluence » du discours. L'exemple suivant illustre ce cadre d'analyse :

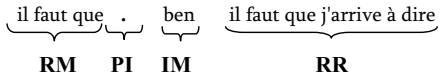


Figure 1 : Structure de la disfluence (Shriberg, 1994)

Cependant, ce modèle révèle un certain nombre de limites. Par exemple, la non-récursivité de ce modèle (*i.e* l'impossibilité d'avoir un schéma **RM/PI/IM/RR** à l'intérieur d'un premier schéma **RM/PI/IM/RR**) empêche de rendre compte de certaines configurations syntaxiques telle que l'imbrication de disfluences.

Il est très fréquent d'observer des imbrications de disfluences : une disfluence s'insère dans une autre avant que la première soit terminée créant ainsi une interdépendance entre les segments disfluents. Il s'agit en fait de plusieurs éléments sur l'axe paradigmatique qui se succèdent et qui se trouvent ainsi sous la dépendance les uns des autres. L'imbrication s'effectue au niveau de la syntaxe où l'on observe les unités syntaxiques fondées sur l'organisation des catégories grammaticales et de leur rection.

**[** on: on parlait souvent **[** du: du fameux euh coq **[** au: au Chambertin ↗ **] ] ]**

A l'inverse du modèle de Shriberg (1994) offrant une vision strictement linéaire de l'organisation des productions disfluentes, le modèle de la mise en grille proposé par (Blanche-Benveniste, 1987) permet de visualiser les configurations du discours grâce à une représentation qui suit deux axes : l'axe syntagmatique (horizontal) et l'axe paradigmatique (vertical). Les phénomènes de production sont ramenés à des « piétinements » sur une même place syntaxique.

| | | |
|---|---|---|
| on:<br>on parlait souvent | | |
| | du:<br>du fameux euh coq | |
| | | au:<br>au Chambertin ↗ |

Figure 2 : Mise en grille de disfluence imbriquée

Dans cet exemple, les éléments *du fameux coq* et *au Chambertin* comportent des piétinements syntaxiques et sont rattachés en cascade au verbe recteur *parler*. Dans nos corpus, nous avons pu relever plusieurs imbrications plus ou moins compliquées. L'imbrication de trois disfluences successives représente le cas le plus complexe. L'intérêt réside ici en une représentation de l'architecture syntaxique des énoncés en suivant un cadre d'analyse unifié. La mise en grille complète la transcription du discours en la rendant à la fois plus lisible et plus compréhensible. Elle permet de traiter les disfluences avec une certaine neutralité (on ne « gomme » pas la disfluence). Grâce à cette représentation tous les essais de lexique sont conservés même s'ils ne font pas avancer le discours.

# 4 Schéma d'annotation

L'annotation des disfluences se révèle difficilement automatisable dans la mesure où l'étude de ces phénomènes réclame un jugement éminemment interprétatif de la part de l'annotateur. Pour cette raison, l'annotation se veut entièrement manuelle et s'effectue à l'aide du logiciel de transcription assistée par ordinateur *Transcriber* qui permet de lier les deux types de ressources nécessaires à l'exploitation des corpus oraux : le fichier son et la transcription.

## 4.1 Principes

Pour réaliser ce travail, il a fallu ensuite trouver une méthode suffisamment « robuste » pour délimiter et coder les disfluences de manière homogène, en limitant les ambiguïtés relatives aux choix d'annotation. Nous nous sommes inspirés du formalisme XML pour la création des balises délimitant les segments disfluents. Le choix d'un codage par balises présente plusieurs avantages : il facilite la hiérarchisation des informations et va dans le sens des normes actuelles qui privilégient ce type de codage dans les projets de normalisation et d'exploitation des corpus oraux (cf. Krul, 2002). De plus, ce type de codage représente un format d'échange standard « universel » et peut être ainsi intégré à des corpus existants utilisant déjà la norme XML.

Dans un premier temps, nous avons donc créé une balise encadrante **<dis>…</dis>** faisant office de délimiteur dans notre schéma d'annotation. Les « piétinements » syntaxiques propres au modèle de (Blanche-Benveniste, 1987) sont représentés par la balise « **<start/>** » placé devant chaque segment.

Exemple :

Segment disfluent initial :

en hiver au Portugal il p- il p- il y a des moments de pluie assez importants des fois ↗

Segment disfluent annoté (1ère passe) :

en hiver au Portugal
**<dis>**
**<start/>** il p-
**<start/>** il p-
**<start/>** il y a des moments de pluie assez importants des fois ↗
**</dis>**

Figure 3 : Première phase d'annotation

Dans un second temps, nous avons attribué à chaque segment disfluent, une étiquette pour qualifier le type de disfluence. Le fait d'affecter un type pour chaque disfluence a pour but de faciliter l'extraction d'informations ponctuelles telle que la répartition des types de disfluences dans le corpus. Là encore, nous nous sommes inspirée du formalisme XML pour décrire les propriétés des segments disfluents. Chaque disfluence possède un attribut « type » qui est défini à l'intérieur de la balise et d'une valeur associée. Nous définissons quatre valeurs possibles pour le type de disfluence : "rep" pour les répétitions, "ac" pour les autocorrections, "am" pour les amorces et "dc" pour les disfluences combinées.

<dis type=**"rep"**>
<start/> il y a:
<start/> il y a une re*mise en question
</dis>

Figure 4 : Deuxième phase d'annotation

Nous avons également annoté les constructions inachevées au moyen d'une marque ponctuelle « **<in/>** ». L'utilisation d'une marque ponctuelle est un choix plus judicieux pour annoter ce type de disfluence car il est très délicat de circonscrire l'inachèvement.

| |
|---|
| là c'était très bien hein j'étais au bord de la mer c'était super ça été un moment de: **<in/>** |

Nous avons également ajouté les informations concernant les marqueurs discursifs (*bon*, *ben*, *voilà*, *donc*, etc. (Chanet, 2004)) sous la forme d'une balise encadrante **<md>…</md>**. Les marqueurs discursifs ne sont pas à proprement parler des disfluences (bien qu'étant étroitement liés à celles-ci), mais leur fréquence élevée dans notre corpus nous oblige à tenir compte de ces unités en les incluant dans le schéma d'annotation. D'un point de vue syntaxique, les marqueurs peuvent être définis comme des mots qui, dans le discours, n'entrent dans aucune construction syntaxique, tout en étant attachés prosodiquement au syntagme dans lequel ils prennent place.

| |
|---|
| **<md>**bon**</md>** là c'était très bien **<md>**hein**</md>** j'étais au bord de la mer c'était super |

## 4.2 Problèmes rencontrés

Même en suivant un modèle théorique préalablement défini, l'annotation se heurte à des cas problématiques. La principale difficulté lorsque l'on annote les disfluences réside dans le fait qu'il n'est pas toujours évident de circonscrire le segment disfluent. En effet, l'absence de ponctuation dans les transcriptions peut poser problème pour délimiter le début et la fin de la disfluence. Un autre problème récurrent est celui des disfluences imbriquées où la difficulté principale réside dans l'application d'un balisage correct des disfluences les unes à l'intérieur des autres.

### *4.2.1 Délimitation du segment disfluent*

Comme nous avons pu le voir dans les exemples précédents, la frontière gauche de la disfluence débute au premier « piétinement » syntaxique, ce qui ne pose pas de difficultés puisque l'on applique la même stratégie d'annotation pour chaque élément disfluent. En revanche, la borne à droite est beaucoup plus difficile à identifier. En l'absence de toute ponctuation, il nous semble que le rôle des indices prosodiques (mouvements intonatifs majeurs) et des autres marques comme les pauses et les allongements ainsi que les marqueurs discursifs est essentiel pour déterminer le début et la fin du segment disfluent. Nous nous sommes donc appuyés sur un certain nombre d'indices pour délimiter la frontière droite de la disfluence. Nous nous sommes basée en premier lieu, sur les éléments prosodiques présents dans le *CRFP*. Plusieurs auteurs ont fait mention du lien entre l'intonation et l'organisation syntaxique de l'énoncé :

> *« La présence de tons finals dominants, avec effet de regroupement, aux frontières syntaxiques majeures, indique une correspondance entre structure syntaxique et structure intonative. »* (Blanche-Benveniste, 1990 : 173)

La figure ci-après montre l'utilisation des marques intonatives comme délimitation de frontière droite de la disfluence.

```
<dis type="ac">
        <start/> mon
        <start/> un vieux collègue de sciences naturelles m'avait dit sur*tout ↗ + pas* d'histoire avec
        les filles ↘
</dis>

on a
<dis type="rep">
        <start/> des:
        <start/> des bons clients ↗
</dis>
```

Figure 5 : Délimitation de la frontière droite à partir de la prosodie

Cependant, les corpus oraux existants ne bénéficient pas systématiquement d'une annotation prosodique. Si l'on souhaite par exemple élargir notre méthodologie à d'autres corpus, il est donc nécessaire, à notre sens, de nous appuyer en parallèle sur d'autres indices plus largement codés dans les corpus oraux à savoir les pauses (silencieuses et remplies), les allongements et les marqueurs discursifs.

### *4.2.2 Disfluences imbriquées*

Il est très fréquent d'observer des imbrications de disfluences : une disfluence s'insère dans une autre avant que la première soit terminée créant ainsi une interdépendance entre les segments disfluents. Il s'agit en fait de plusieurs éléments sur l'axe paradigmatique qui se succèdent et qui se trouvent ainsi sous la dépendance les uns des autres. L'imbrication s'effectue au niveau de la syntaxe où l'on observe les unités syntaxiques fondées sur l'organisation des catégories grammaticales et de leur rection.

```
<dis>
<start/> on:
<start/> on parlait souvent
                        <dis>
                        <start/> du:
                        <start/> du fameux euh coq
                                                <dis>
                                                <start/> au:
                                                <start/> au Chambertin ↗
                                                </dis>
                        </dis>
</dis>
```

Figure 6 : Imbrication de disfluences

Dans cet exemple, les éléments *du fameux coq* et *au Chambertin* comportent des piétinements syntaxiques et sont rattachés en cascade au verbe recteur *parler*. Dans nos corpus, nous avons pu relever plusieurs imbrications plus ou moins compliquées. L'imbrication de trois disfluences successives représente le cas le plus complexe.

### *4.2.3 Raccordement syntaxique impossible*

Notre méthode d'annotation permet de rendre compte de la plupart des configurations de disfluences. Cependant, quelques énoncés disfluents ne peuvent pas être annotés en utilisant

cette méthode. En effet, les relations entre les éléments peuvent être à une distance tout à fait notable et le raccordement syntaxique s'effectue trop loin dans l'énoncé pour être annoté et mis en grille. C'est le cas notamment de l'exemple suivant pour lequel il n'a pas été possible de mettre en place notre schéma d'annotation :

ben parce que le: Charlemagne euh + paraît-il ↗ **ne buvait que des rouges** ↗ + parce qu'
<dis type="dc">
<start/> il voulait pas ↗
<start/> il
<start/> il se tâchait ↗ +
<dis type="dc">
<start/> sa:
<start/> sa: + </in>
</dis>
</dis> il **ne buvait que* des Blancs** pardon ↗

Figure 7 : Autocorrection non annotée

La correction *ne buvait que des Blancs* remplace *ne buvait que des rouges* initialement produite par le locuteur. Cependant, cet énoncé se situe à une trop grande distance pour être annoté en tant qu'autocorrection.

# 5 Aspects quantitatifs

Après avoir annoté les corpus, nous avons réalisé une étude quantitative dont l'objectif est d'illustrer la typologie des phénomènes de l'oral. L'approche quantitative permet d'accéder plus facilement à la description des phénomènes qui présentent de l'intérêt et dont il aurait été difficile de cerner les contours *a priori*. A l'aide de scripts permettant d'automatiser les décomptes (scripts en *Perl* et *Bash*), nous avons quantifié les types de disfluences, les constructions inachevées ainsi que les marqueurs discursifs.

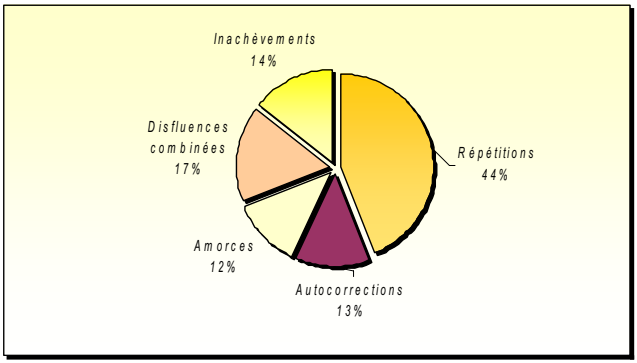


Figure 8 : Répartition des types de disfluences

Nous avons relevé 293 disfluences : les répétitions constituent le type le plus largement représenté (44%). Les autres types sont repartis de manière plus homogène : leur pourcentage varie entre 17% pour les disfluences combinées et 12% pour les amorces. Les comptages

permettent également d'effectuer quelques constats sur la fréquence des marqueurs discursifs, et permettent de dégager des hypothèses sur le fonctionnement de ces unités.

| **TÊTE DE LISTE : Fréquence des marqueurs discursifs** | | | |
|---|---|---|---|
| Rang | Forme | Fréquence | Fréquence Relative |
| 1 | donc | 67 | 18,61% |
| 2 | hein | 55 | 15,28% |
| 3 | bon | 36 | 10% |
| 4 | quand même | 15 | 4,17% |
| 5 | enfin | 15 | 4,17% |
| 6 | ben | 13 | 3,61% |
| 7 | alors | 13 | 3,61% |
| 8 | là | 9 | 2,50% |
| 9 | quoi | 8 | 2,22% |
| 10 | mais | 8 | 2,22% |

Tableau 1: Fréquence des marqueurs discursifs (tête de liste)

# 6 Conclusion et perspectives

L'étude de l'oral est aujourd'hui un thème de recherche très riche, même s'il reste encore de nombreux progrès à faire pour permettre d'automatiser complètement son traitement. L'une des premières phases dans l'optique du développement d'applications en TAL dans ce domaine, peut passer, notamment, par la constitution et l'annotation de corpus oraux.

Cet article rend compte du travail d'annotation effectués de ces phénomènes à partir de corpus oral (*Corpus de Référence du français parlé*) en suivant un modèle d'analyse précis (la mise en grille) et en utilisant une norme générique d'annotation de textes (XML). De plus, l'étude quantitative et qualitative menées conjointement dans notre travail nous ont permis d'avoir une connaissance plus précise des caractéristiques des phénomènes de l'oral même s'il reste encore beaucoup de cas à étudier, et ce, sur de plus grands volumes de données. Nous avons pu, grâce aux observations sur corpus, dégager un certain nombre de régularités qui nous renseignent sur le fonctionnement des disfluences (qui peut être enrichi par l'observation des différents patrons syntaxiques des segments disfluents (Piu, 2006)) et peuvent servir de base pour l'amélioration des outils de traitement automatique (par exemple l'analyse syntaxique automatique) qui butent encore sur les données orales.

# Références

BLANCHE-BENVENISTE, C., & JEANJEAN, C. (1987). *Le français parlé – Édition et transcription.* Paris : Didier-Érudition.

BLANCHE-BENVENISTE, C. (1990). *Le français parlé – Études grammaticales.* Paris : CNRS.

CANDÉA, M. (2000). *Contribution à l'étude des pauses silencieuses et des phénomènes dits « d'hésitation » en français oral spontané.* Thèse de doctorat. Université Paris III.

CHANET, C. (2004). *Fréquence des marqueurs discursifs en français parlé : quelques problèmes de méthodologie.* Recherches sur le français parlé, 18, 83-105.

ÉQUIPE DELIC. (2004). Présentation du Corpus de référence du français parlé. *Recherches sur le français parlé* 18, 11-43.

HENRY, S. (2002). Étude des répétitions en français parlé spontané pour les technologies de la parole. Actes de la *6ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, 467-476. Nancy (France).

KRUL, A. (2002). *Annotation structurelle de corpus oraux avec XML*. Mémoire de Maîtrise. Université Paris III Sorbonne Nouvelle.

KURDI, M. Z. (2003). *Contribution à l'analyse du langage oral spontané.* Thèse de doctorat. Université de Grenoble I.

PALLAUD, B. (2002). Les amorces de mots comme faits autonymiques en langage oral. *Recherches sur le Français parlé*, 17, 79-102.

PIU, M. (2006). *Annotation des disfluences dans les corpus oraux*. Mémoire de Master. Université de Provence, Aix-en-Provence.

VÉRONIS, J. (1998). *Annotation automatique de corpus : état de la technique*. Colloque International « Questions de méthode dans la linguistique de corpus ». Perpignan (France).