Guide d'annotations des espaces interruptifs (interruptions suspensives et disfluentes) en français parlé dans les huit dilogues du CID (Corpus of Interactional Data)

Berthille Pallaud

Guide d'annotations des espaces interruptifs (interruptions suspensives et disfluentes) en français parlé dans les huit dilogues du CID

B. Pallaud, Laboratoire Parole et Langage UMR 6057, Université de Provence, 5 avenue Pasteur, BP 80975 13604 Aix en Provence Cedex 1
Courriel : berthille.pallaud@lpl-aix.fr

Introduction

Les paroles des locuteurs comportent des fluctuations dans le rythme de la fluence verbale qui apparaissent dans le débit de la prononciation des mots eux-mêmes (Pasdeloup, 1992 ; Duez, 2007; Béchet et al. 2013) mais surtout dans la présence plus ou moins nombreuse d'éléments « étrangers » à l'énoncé informatif et insérés dans cet énoncé : pauses *silencieuses* ou *remplies* et insertions parenthétiques diverses (marqueurs de discours, marqueurs phatiques et interjections) suspendent le déroulement syntagmatique. Ces insertions situées à la suite d'une interruption rompent le déroulement du texte de l'énoncé et, par cette pause, définissent un avant et un après dans l'énoncé ce que Shriberg (1994) a nommé *Reparandum* et *Reparans*.

Ainsi, les variations dans le rythme de la fluence verbale peuvent aller jusqu'à l'auto-interruption de mots ou de syntagmes mais elles ne sont pas toujours suivies de ces insertions dans l'espace *Interregnum* (moment potentiellement utilisé par le locuteur entre le *Reparandum* et le *Reparans*[1]) (YM_1506). De plus, ces auto-interruptions et auto-variations dans la fluence verbale peuvent s'accompagner de perturbations qui se traduisent, en particulier, par des effets sur l'organisation morphosyntaxique du flux verbal, les plus fréquemment cités étant les reprises d'énoncés, les auto-réparations (YM_1977) et les inachèvements de syntagme ou de mots (YM_1739) :

YM_1506 un gosse **j'en** // **j'en** // **moi j'en** voulais pas
YM_1739 ça commençait **fé-** // en février ou un truc comme ça
YM_1977 **moi je** // quand il est parti il nous a laissé **tous ses** // euh **tous ses** épices

L'avantage de cette description est qu'elle permet de distinguer ce qui précède l'interruption, ce qui lui succède éventuellement sans renouer le fil du discours (sans fonction syntaxique) et ce qui vient réparer, poursuivre ou laisser inachevée la construction syntaxique interrompue. Notre identification des interruptions dans le flux verbal repose sur la détection de ce qui suit l'instant interruptif c'est-à-dire essentiellement (mais pas toujours, loin de là) sur l'observation des indices d'interruption qui se trouvent dans l'espace Interregnum.

La plupart des travaux consacrés à la disfluence ne distinguent pas les indices d'interruption (pauses remplies ou non, marqueurs de discours, incidentes parenthétiques) des effets de ces interruptions et se sont attachés à décrire les phénomènes acoustiques, phonétiques et prosodiques présents :

- au moment de l'interruption de l'énoncé (déformation de phonèmes, par exemple Shriberg, 1999)
- ou dans le moment qui suit cette interruption avant que l'énoncé ne reprenne ou soit définitivement interrompu (pauses silencieuses ou remplies, Duez, 2001, 2001b et les particules discursives, par exemple).

D'autres études (Pallaud & Henry, 2004 ; Dister, 2007 et 2008) ont observé les effets énonciatifs des auto-interruptions dans des énoncés non préparés ; les disfluences constituant

[1] Nous avons repris la terminologie de Shriberg (1999) : Reparandum, Reparans, Interregnum.

une partie des **effets** de ces auto-interruptions involontaires du flux verbal dans le déroulement syntagmatique:

-reprise avec ou sans modification d'une partie de l'énoncé qui suit l'interruption.

-inachèvement de l'énoncé (mot ou syntagme).

Plutôt que de chercher à quantifier et analyser quelques types de disfluence préalablement définis comme les amorces de mots ou les répétitions, l'option choisie a été d'identifier toutes les auto-interruptions dans les énoncés et de décrire les conséquences morpho-syntaxiques de ces ruptures dans le flux verbal. Cette approche permet d'une part d'identifier de façon exhaustive toutes les auto-interruptions dans un énoncé et d'autre part de décrire les relations existant entre ces auto-interruptions et leurs effets morpho-syntaxiques. Elle permet notamment de préciser la notion de disfluence (Pallaud, 2014).

Les corpus du CID et méthode de description

La méthode d'identification et le système d'annotation des interruptions ainsi que leurs effets sur la suite de l'énoncé ont été élaborés et appliqués sur les huit dilogues du corpus du CID (16 locuteurs). Il s'agit de corpus de paroles non préparées recueillis lors de conversations d'une heure entre deux locuteurs sur un thème proposé (Bertrand *et al*, 2009). Les locuteurs diffèrent quant au nombre de mots prononcés et diffèrent également quelque peu sur le temps de parole occupé.

Les interruptions du flux verbal quelles qu'elles soient ont divers effets possibles, au niveau prosodique et phonétique (de plus en plus étudiés; Shriberg et son équipe) mais aussi sur la séquence verbale qui suit (perturbations dans la linéarité ou la structure syntaxique). Cependant, elles ne sont **pas toutes** sources de disfluence c'est-à-dire de réparation ou d'abandon et peuvent être sans effet sur la poursuite de l'énoncé (interruption simplement suspensive ; voir plus loin).

Trois effets de ces interruptions sont donc observés et étudiés :

1*celles où l'énoncé est simplement poursuivi :

YM _2432 *tu le fais avec **euh** ce que tu veux*

Nous avons nommé cette catégorie d'interruptions « **interruptions suspensives »** puisqu'elles ne provoquent qu'une suspension (elle aussi temporaire) et non une réorganisation de l'énoncé.

2*celles qui provoquent la reprise d'une partie de l'énoncé interrompu avec une éventuelle modification de l'énoncé repris :

CM 884 *mais **les** euh **les** nanas du foyer elles étaient pas au courant*
CM 996 *tu as **des des** feux d'artifices comme ça*
YM_1124 *chaque fois **tu ma-//tu devais ma**rquer*

3*celles qui, l'énoncé ayant été laissé inachevé, sont suivies d'une nouvelle construction ou d'un nouveau syntagme.[2]

CM 1139 *Ah si mais **j'ai** // c'est un truc qui m'avait fait bien rire*
CM 1198 *et* je *// mes fixations sont pas bien réglées*

[2] Toute interruption non suivie d'une reprise ou d'une poursuite de l'énoncé est considérée comme un événement inachevé. Cela peut être dû aux difficultés du locuteur dans l'élaboration de son énoncé mais cela peut aussi être provoqué par l'autre locuteur qui commence à parler. Ces deux sortes d'interruption seront codées et catégorisées comme abandons d'énoncé. Une analyse plus fine permettrait sans doute de distinguer quand l'interruption est vraiment provoquée par l'intervention du locuteur partenaire.

Les auto-interruptions (ainsi définies) et leurs effets morpho-syntaxiques ont été systématiquement annotés dans les huit dilogues du CID (intégralement dans deux des huit dilogues du CID (une heure pour chaque locuteur). Pour ce qui est des six autres dilogues, ils n'ont été annotés que sur 20min pour chaque locuteur.

L'énoncé progresse par séries de mots qui sont séparées les unes des autres par divers phénomènes :

- des pauses silencieuses égales ou supérieures à 200ms
- des pauses remplies (*euh*)
- des incidentes souvent très courtes remplies de marqueurs de discours et d'éléments phatiques (*tu sais, tu vois, bon, alors, donc, mais* etc.)
- Ces phénomènes sont soit **isolés** soit présents dans l'Interegnum en formation **multiple**
  *BX 904.14 des des portes qui ferment pas complètement **euh # enfin bon** y avait **euh***
- **Le point d'interruption** est à la fin du mot qui précède cet espace
- Perturbation morphosyntaxique et espace Interregnum vide
  *BX 278.83 Tout ce qui pouvait se brancher **su-// quelque part** c'était # tout branché sur la multiprise*

II. Méthode d'identification : le système d'annotation des auto-interruptions et des disfluences morphosyntaxiques

II.1. La méthode d'identification des auto-interruptions et disfluences morphosyntaxiques

Afin de décrire **la totalité** des interruptions dans les énoncés (en moyenne 1060 interruptions par locuteur) et les interruptions provoquant des perturbations morphosyntaxiques (en moyenne 492 interruptions disfluentes par locuteur), deux méthodes de détection ont été employées successivement : l'une semi-automatique et l'autre manuelle. Les deux méthodes utilisent le logiciel Praat comme outil d'identification, d'annotation et de description.

Méthode semi-automatique : il s'agit de localiser les interruptions du flux verbal grâce aux indices d'interruption afin de procéder à l'annotation du phénomène qui leur est lié. Les indices d'interruption sont constitués par des événements isolés (par exemple, une pause silencieuse) ou par plusieurs éléments se succédant et dit "associés" (par exemple, une pause remplie suivie d'un marqueur de discours ou d'une interjection). Ces indices d'interruption ont en commun de ne pas avoir de fonction syntaxique.

AB_571 et *euh #* et je *euh bon* moi j'étais dans un état un peu pas très bien
AB_620 C'était *euh tu vois* complètement loufoque comme si- *ouais euh* comme situation

Méthode manuelle : Il reste que certaines interruptions dans la fluence verbale ne sont signalées que par leurs effets morphosyntaxiques (une sur cinq). On n'observe aucun des éléments décrits précédemment ; l'espace potentiel Interregnum entre l'énoncé interrompu et celui qui lui succède est vide. Quelques éléments discursifs comme les insertions parenthétiques qui interrompent un énoncé en provoquant souvent des réorganisations morphosyntaxiques ne peuvent pas non plus être détectée automatiquement :

AB_2187 ce devait être une boutique style mercerie *je sais pas trop* qui // qui euh
CM_2031 et y avait *donc on était aux Etats Unis* y avait des chaînes cryptées

La méthode manuelle consiste donc à faire une lecture-écoute sémantique des transcriptions qui, à l'aide de paramètres prosodiques, sémantiques et/ou syntaxiques, révèle ces ruptures.
Qu'un cinquième des interruptions ne soient signalées que par une discordance syntaxique constitue pour notre méthode une difficulté certaine. Les locuteurs varient quant au nombre d'interruptions dans leurs énoncés mais la moyenne est supérieure à un millier ce qui conduit à identifier, pour chaque locuteur, environ deux cents disfluences non suivies d'un espace Interregnum détectable.
Une solution serait le repérage des discordances syntaxiques ce qui pourrait constituer une méthode de détection automatique de perturbations dans l'énoncé. Elle serait à évaluer par une confrontation avec les résultats de la méthode manuelle décrite ici.

II.2. Le système d'annotation

C'est autour du point d'interruption dans la fluence verbale que l'annotation du phénomène provoqué par cette rupture prend place. L'interruption peut se produire au milieu d'un mot (*word truncation*) ou au milieu d'un syntagme (*phrase truncation*) et délimite trois espaces formels dont la chronologie est la suivante : **le Reparandum, l'espace Interregnum (*Break point*) et le Reparans**. Chacun de ces espaces est défini par l'étendue des items (verbaux ou non).

1° le **Reparandum (R ou I)** est ce qui, avant le point de rupture, contient une perturbation (fragment de mot ou de syntagme) et sera simplement poursuivi, repris, répété, modifié (R) ou abandonné (I), lors du Reparans.

2° l'Interregnum (*Break interval* B) est un moment potentiel avant que n'intervienne le Reparans. Il peut être vide ou contenir **des indices** d'interruption, le plus souvent non lexicalisés (pauses remplies ou silencieuses, répétition de troncation, éléments discursifs, phatiques ou parenthétiques plus ou moins longs, onomatopées, etc.)
3° le Reparans, partie potentielle de l'énoncé prononcé qui peut poursuivre, répéter ou modifier ce qui a été dit lors du Reparandum. Cet élément comporte deux situations selon qu'il est vide ou rempli.
3.1° Reparans vide : énoncé laissé inachevé ; il n'y a pas de Reparans codé.
3.2°Reparans non vide : **Reparans (RA)** qui comporte trois possibilités :

* une simple complétude du syntagme commencé et interrompu[3]
* la reprise partielle de l'énoncé déjà prononcé ce qui correspond à un entassement paradigmatique (le piétinement syntaxique selon Claire Blanche-Benveniste, 1997).
* la reprise de l'énoncé prononcé comporte des modifications.

L'analyse morphosyntaxique de ces interruptions (Pallaud 2002 ; Pallaud & Henry, 2004) a montré que le Reparandum ne peut être identifié qu'à l'aide des éléments qui vont lui succéder et tout particulièrement ce qui va être repris de l'énoncé avant le point de rupture (c'est-à-dire le Reparans). Le nombre d'éléments contenus dans le Reparandum est déterminé par ce qui constitue le Reparans. Lorsqu'il y a inachèvement de l'énoncé le Reparandum est l'item tronqué ou le dernier item du syntagme laissé inachevé. Il en est de même pour les interruptions suspensives.

[3] A noter qu'il n'y a jamais de simple complétude d'un mot tronqué chez un locuteur standard. Un mot tronqué involontairement n'est complété qu'après une reprise au moins du début du mot (parfois du déterminant également ; Pallaud., 2005, 2006 et 2006). En revanche, des personnes bègues peuvent compléter sans reprise un mot tronqué (Pallaud & Xuereb, 2008).

L'annotation des interruptions présentes dans les corpus du CID rend compte de cette structure. Elle est faite sous Praat et portée dans le fichier *TextGrid_Nom_ortho token* où le son est aligné sur les éléments mots, articles, pronoms, etc. Elle est inscrite sur une ou plusieurs Tiers (terminologie de Praat) selon que les disfluences sont ou non enchâssées. Celles-ci sont alignées sur les tokens concernés par les trois sortes d'éléments de la structure : dans l'ordre chronologique, le Reparandum, l'espace Interregnum et le Reparans. Il s'agit donc de définir les intervalles correspondant aux trois axes de la structure. Chaque élément de cette structure comporte, à son tour, plusieurs sortes d'informations qui vont être décrites et codées.

L'annotation peut donc se limiter à identifier les trois éléments de la structure (Reparandum, Interregnum et Reparans) ou décrire également chacun des types d'éléments qui les composent et dont la description peut se résumer par le tableau suivant (Blache *et al*. 2010).

Reparandum		
Reparandum Type	***R***	Temporary interruption
	I	Definitive Interruption
Reparandum_category	***W***	Word reparandum
	P	Phrase reparandum
Lexical_type	***tw***	Tool word
	lw	Lexical word
Break_type B		
	no	no interval
	sp	silent pause (> 200ms)
	fp	filled pause
	dc	discursive connector
	ps	parenthetical statement
	rt	truncation repetition
Reparans RA		
Reparans_position_type	***nr***	no restart
	wr	word restart
	dr	determinant restart
	pr	phrase restart
	or	other restart
Reparans_type	***co***	continuing the item
	wc	repairing word without change
	rp	repairing through repeating
	rc	repair with change in the truncated word
	rm	repair with multiple change

Tableau 1 Système d'annotation des interruptions dans la fluence verbale

Détails sur le codage

I. Les interruptions suivies d'une réparation ou d'un abandon

1. Le Reparandum (R ou I)

Deux sortes de données sont codées:

*L'élément affecté par l'interruption

*Le type de mot

- **1.1. L'élément affecté** par l'interruption : le syntagme **P** (Phrase **P**) ou un mot **W** (word **W**)

CM gpd_56 et on était *p-* (**R,W**,lw) on avait loué une voiture
CM gpd_33 tu perds *un peu* (**R,P,**tw) comment dire euh + des repères
CM gpd_24 *c'est* (**I,P,**tw B,no) ça n'évoque *rien*
AG_1540 normalement on a la *cr-* (**I,W**,lw) enfin (B,dc) *c'est* c'est o- septembre c'est OK

- **1.2. Le type de mot** qui est tronqué ou suivi de l'interruption (tool word **tw**, lexical word **lw**)

Mot lexical tronqué :
CM gpd_56 et on était p- (R,W,**lw**) on avait loué une voiture (B,no RA,pr,rm)
Mot outil tronqué :
CM gpd_55 en *f-* (R,W,**tw**) c'était complètement loufoque
Mot outil non tronqué :
CM gpd_33 tu perds *un peu* (R,P,**tw**) comment dire euh + des repères
Mot lexical non tronqué :
AG_309 je me rappelle qu'à un moment *donné* (R,P,**lw**)euh (B,sp) *on* (RA,nr,co) était en (R,P,tw)donc (B,dc) en(RA,wr,wc) classe

2. Pour l'espace potentiel (B, Break interval) après le Reparandum et avant le Reparans, le type d'espace potentiel est codé :

- Les types d'espace potentiel:
  - Rien (nothing B,**no**)
  - Pause silencieuse (silent pause **B,sp**)
  - Pause remplie (filled pause **B,fp**)
  - Elément discursif (discursive connector **B,dc**)
  - Répétition du fragment (truncation repetition **B,tr**)
  - Enoncé parenthétique (parenthetic statement **B,ps**)

Exemples :
CM gpd_33 oui où tu perds un peu *euh* + **(B,fp,sp)** tu perds un peu *comment dire euh* + **(B,dc,fp,sp)** des repères
CM gpd_46 euh qui était complètement (R,P,tw) ah voilà + **(B,dc,sp)** dés- (RA,nr,co) euh + désynchronisé d'une situation en fait

3. Pour le Reparans (RA), deux types de données sont codés:

- **3.1. la position du Reparans:**
  - **pas de reprise (no restart nr)**

AG_376 je suis resté un an à Dumond d'Urville (R,P,lw) tu sais (B,dc) *à Toulon* (**RA,nr,co**)

  - **reprise au début du mot (word restart wr)**

CM gpd_32 *où* (R,P,tw) *euh oui* (B,fp,dc) *où* (**RA,wr**,rp)
CM gpd_67 on était *complèt(e)ment* (R,P,lw) euh + (**B,fp,sp**) *complèt(e)ment* (**RA,wr,**rp)

  - **reprise au déterminant (determiner restart dr)**

AP gpd_246 918.81 : *un collèg-* (R,W,lw) *enfin* (**B,dc**) *un mec* (**RA,dr,**rc)

o **reprise au début du syntagme (phrase restart pr)**

CM gpd_33 oui où *tu perds un peu* (R,**P,tw**) *euh* + (**B,fp,sp**) *tu perds un peu* (**RA,pr,**rp)
CM gpd_56 et *on était p-* (R,**W,lw B,no**) *on avait loué* une voiture (**RA,pr,rm**)

o **autres types de reprise (other restart or)**

MG_569 il fait une espèce de (R,P,tw B,no) *il est sur une espèce de village* (**RA,or,rm**)
AG_1493 *j'ai ma belle-* (R,W,lw) // (B,no) *il y a ma belle* (**RA,or,rm**) mère qui

- **3.2. le fonctionnement du Reparans**
  - o **simple continuation of the item (co)**

CM gpd_55 en *f-* (R,W,tw B,no) c'était complètement loufoque (RA,nr**,co**)
AP gpd_242 907.934 *le* (R,P,tw) *euh* (B,fp) *mari* (RA,nr**,co**)

o **repairing the truncated word without change (wc)**

AP gpd_249 935.69: la *f-* (R,W,lw B,no) *famille* (RA,wr,**wc**)
CM gpd_46 euh qui était complètement ah voilà + *dés-* (R,W,lw) euh + (B,fp,sp) *désynchronisé* **(**RA,wr**,wc)** d'une situation en fait

o **repairing through repeating (rp)**

AP gpd_247 925.04: *de* (R,P,tw) (B,no) *de* (RA,dr,**rp**) tchatcheur
CM gpd_32 *où* (R,P,tw) euh oui (B,fp,sp,dc) *où* (RA,wr,**rp**) tu perds un peu
CM gpd_33 oui où *tu perds un peu* (R,P,lw) euh + (B,fp,sp) *tu perds un peu* (RA,pr,**rp)**
CM gpd_67 on était *complètement* (R,P,lw) euh + (B,fp,sp) *complètement* (RA,wr,**rp**)

o **repair with changement in the truncated word (rc)**

AP gpd_246 918.81 : *un collèg-* (R,W,lw) enfin (B,dc) *un mec* (RA,d,**rc**)

o **repair with multiple changements (rm)**

CM gpd_25 *de con-* (R,W,lw) (B,no) *de bien connaît(re)*(RA,pr**,rm**)
CM gpd_56 et *on était p-* (R,W,lw B,no) *on avait loué* (RA,pr,**rm**) une voiture

En résumé:
Total des interruptions:

- R,P,_(interruption de syntagme ou de proposition)
  - o RA_nr,co (sans reprise)
  - o RA,dr,wr, pr or (avec reprise)
- R,W_(interruption de mot)
  - o RA, jamais de nr (sauf cas de **dys**fluence comme dans le bégaiement)
  - o RA, wr,dr,pr,or (avec reprise)
- I,P,_(interruption **définitive** de syntagme ou de proposition) pas de Reparans
- I,W_(interruption **définitive** de mot) pas de Reparans

Une analyse supplémentaire (tier supplémentaire) a été faite pour les interruptions de syntagme et de propositions sans reprise (RA,nr). Elles ont été annotées sur la dernière tier (tier 5):

Elle permettrait une comparaison de ses espaces dans le cas de poursuite de l'énoncé avec les abandons et les reprises..

Nr,s: au milieu d'un syntagme.

Nr,s,s (pause)

Nr,s,h (pause remplie, euh)

2432 YM tu le fais avec euh ce que tu veux

2401 YM des trucs insolites **euh #** écossais

Nr, pv: au milieu d'un proposition

2264 YM même l'accent **quoi** c'est dur à comprendre

Nr, pc: suivie d'une coordination

2043 YM c'est des châtaignes **ouais mais** c'est pas

Nr: pq: Que phrase

1833 YM Bernard c'est vers là-bas **non** qu'il habite

Nr,pf: à la fin d'une proposition suivie d'une proposition sans lien avec la première

1734 YM non je crois qu'il commençait plus tard **#** il me semble qu'il commençait

La catégorie Nr,s (interruption de syntagme sans reprise) est suivie d'un intervalle rempli de façon diverse:

Nr,s,s (pause)

561 YM ouais ca t'a **#** assomé quoi

Nr,s,h (pause remplie, euh)

2432 YM tu le fais avec **euh** ce que tu veux

Nr,s,d (élément discursif)

279 YM mais **putain** c'était je pense

Nr,s,p (parenthèse)

786 YM dix heures dix heures et demi **je crois** du soir