

Annotation des marqueurs de fluence et disfluence dans des corpus multilingues et multimodaux, natifs et non natifs

Version 1.0

bigskip *Crible L., Dumont A., Grosman I., Notarrigo I.*

29 janvier 2015

Citer ce papier : Crible L., Dumont A., Grosman I., Notarrigo I. 2015. Annotation des marqueurs de fluence et disfluence dans des corpus multilingues et multimodaux, natifs et non natifs. Version 1.0. *Working paper*. Université catholique de Louvain et Université de Namur.

Cette recherche est le fruit d'une collaboration entre
l'Université catholique de Louvain et l'Université de Namur

Fluency and disfluency markers. A multimodal contrastive perspective



FÉDÉRATION
WALLONIE-BRUXELLES

Action de Recherche Concertée financée par la Fédération Wallonie-Bruxelles
(n° de financement Bourse 12/17-044)



Table des matières

1	Introduction	4
1.1	Contexte scientifique	4
1.2	Objectif du document	4
2	Aperçu du protocole : design et applications	5
2.1	Définitions et terminologie	5
2.2	Aspects techniques de l’annotation	6
3	Catégories de fluencèmes couvertes par le protocole d’annotation	8
3.1	Fluencèmes simples	8
3.1.1	Pauses (UP, FP, S1, S2, S3)	8
3.1.2	Palm-up en LS (PU)	10
3.1.3	Marqueur de discours (DM)	10
3.1.4	Terme explicite d’édition (ET)	12
3.1.5	Faux départ (FS)	12
3.1.6	Troncation (TR)	13
3.2	Fluencèmes composés	14
3.2.1	Répétitions (RI, RM, RE, RG)	14
3.2.2	Substitutions (SM, SP)	17
3.3	Insertions	18
3.3.1	Insertion lexicale (IL)	18
3.3.2	Insertion parenthétique (IP)	19
3.4	Signes diacritiques	20
3.4.1	Articulation (AR)	20
3.4.2	Allongement (LG)	20
3.4.3	Enchâssement de fluencème simple (WI)	21
3.4.4	Changement d’ordre (OR)	21
3.4.5	Complétude (syntaxique) fonctionnelle et totale (CF et CT)	22
4	Formulation de requêtes	23
5	Conclusion	23

1 Introduction

1.1 Contexte scientifique

La recherche en linguistique sur les phénomènes de fluence et de disfluence est abondante, pluridisciplinaire et adopte des approches différentes selon les objectifs scientifiques. Cette situation aboutit aujourd’hui à un panel de protocoles d’annotation disponibles dans la littérature mais rarement comparables ou généralisables à des données de nature différente. En effet, depuis l’ouvrage précurseur de Shriberg (1994), plusieurs auteurs ont adapté son système et sa terminologie, tantôt dans une perspective d’annotation automatique, tantôt pour spécifier ou modifier des catégories d’analyse. Parmi les travaux qui ont principalement inspiré notre recherche, citons les suivants : SHRIBERG (1994), METEER (1995), CANDEA (2000), EKLUND (2004), DISTER (2007) et GÖTZ (2011). D’autres projets dans des approches similaires sont actuellement en cours, notamment CHRISTODOULIDES, AVANZI et GOLDMAN (2014), MONIZ, BATISTA, MATA et TRANCOSO (2014) et PALLAUD, RAUZY et BLACHE (2013). Nous ferons référence à des travaux plus ponctuels sur l’un ou l’autre phénomène linguistique dans les sections qui leur sont respectivement dédiées ci-dessous.

L’intérêt pour une approche compositionnelle¹ de la fluence comme proposée par GÖTZ (2011) tend à se répandre. En revanche, les éléments qui contribuent à cette vision de la fluence tels que répétition, substitution, pauses etc., ainsi que les formats d’annotation diffèrent bien souvent sur la portée des phénomènes couverts. Plus précisément, les différences entre systèmes portent sur les types de phénomènes observés, les spécificités de langues et de modalités, les choix technico-pratiques en vue de l’efficacité de l’annotation (format des labels, extraction des données), etc. De manière générale, la plupart de ces protocoles présentent un certain nombre de lacunes, que ce soit sur le plan pratique de la répliquabilité de l’annotation, de sa compatibilité avec un traitement quantitatif efficace, ou sur le plan théorique de la finesse et de la validité des catégories annotées, de la robustesse des critères qui permettent de distinguer les différents phénomènes, et de la pertinence cognitive et pragmatique du modèle dans son ensemble.

Dans cette perspective, le présent protocole se propose de résoudre un certain nombre de ces problèmes, grâce à l’apport des recherches précédentes d’une part, mais aussi grâce à la richesse des cadres théoriques et des types de données concernés dans notre démarche. En effet, l’originalité de l’approche proposée ici est de dépasser les particularités liées à un cadre unique en offrant un modèle exhaustif, flexible et modulable qui s’adapte potentiellement à un grand nombre de questions de recherche. Les sections suivantes précisent le cadre théorique et méthodologique de l’approche adoptée par ce protocole, avant de passer à la description opérationnelle des phénomènes couverts par l’annotation.

1.2 Objectif du document

L’objectif de ce document est de présenter un protocole d’annotation des marqueurs de (dis)fluence en adéquation avec une vision componentielle de la (dis)fluence (*ibid.*) et apte à prendre en charge les marqueurs de (dis)fluence en français, en anglais (L1 et L2) et en langue des signes de Belgique francophone (LSFB). La création de ce protocole est le fruit d’une collaboration entre plusieurs chercheuses dont les recherches doctorales abordent des données différentes au niveau de la langue (français, anglais et LSFB), de la modalité (langue orale versus langue des signes), des types de locuteurs (natifs versus apprenants), et enfin plus généralement au niveau de l’approche analytique (semi-automatique versus manuelle, sémasiologique versus onomasiologique). Cette multiplicité d’approches n’est pas simplement la preuve de la flexibilité de ce protocole, mais constitue surtout le fondement d’un certain nombre de décisions théoriques et pratiques qui ont forgé ce travail dès son origine. Le protocole offre la possibilité de recourir à l’utilisation d’une même étiquette pour se référer à une même notion dans le but d’assurer la comparabilité entre les travaux et se veut donc applicable et adaptable à des corpus oraux² de locuteurs natifs et non natifs. Nous posons que cette modularité linguistique, modale et théorique renforce la robustesse et la pertinence des catégories annotées et du protocole dans son ensemble.

Dans la perspective d’une répliquabilité optimale, chaque phénomène couvert par l’annotation est systématiquement présenté selon la structure suivante : définition du phénomène, critères de distinction et d’application

1. cf. section 2

2. Dans la suite du document, le mot «corpus» servira de terme générique pouvant se référer à des corpus multilingues, natifs ou non natifs, oraux ou signés

sur données, étiquette préconisée pour l’annotation et autres considérations techniques éventuelles, et illustration par plusieurs exemples authentiques issus des corpus utilisés par les auteurs.

Enfin, les définitions et critères proposés sont le produit d’une confrontation des théories existantes (envisagées selon nos objectifs) à leur application sur données authentiques de corpus par les quatre auteurs. Cette méthode permet d’asseoir l’opérationnalité du protocole en s’assurant de l’applicabilité des critères et en minimisant les différences entre annotateurs. Il n’est pas exclu que certaines catégories soient plus sujettes à l’interprétation de l’analyste que d’autres, étant donné la part inévitable – mais limitée – de considérations sémantico-pragmatiques impliquée dans toute approche de la (dis)fluence. Toutefois, la grande majorité des phénomènes et des critères retenus sont liés à des paramètres formels, des indices de surface qui limitent l’intervention subjective de l’annotateur, comme le montreront les catégories de la section 3.

2 Aperçu du protocole : design et applications

2.1 Définitions et terminologie

La (dis)fluence peut être définie d’après une approche holistique comme l’utilisation fluide, rapide et sans effort du langage (CRYSTAL 1988), selon laquelle l’évaluation fluente ou disfluente est du ressort de la perception, de l’impression globale qu’un discours produit sur l’interlocuteur. Une seconde approche dite componentielle (GÖTZ 2011), dans laquelle s’inscrit le présent document, voit la fluence et la disfluence comme une combinaison de traits qui, pris isolément, peuvent être impliqués tantôt dans la fluence tantôt dans la disfluence d’un discours selon leur fréquence, leur fonction, leur position et leur combinaison. En d’autres termes, ces marqueurs peuvent être envisagés soit comme des signaux plus ou moins délibérés utilisés avec succès pour aider à la production et à la compréhension, soit comme un signal de difficulté(s) à planifier et à encoder en direct un énoncé. Les causes possibles de ces difficultés peuvent relever de l’accès lexical (phénomène du « mot sur le bout de la langue »), de la complexité sémantique et/ou syntaxique de l’énoncé, de la charge cognitive et émotionnelle du discours, du besoin de gagner du temps pour préparer la suite du discours, ou plus directement de la perception d’une « erreur » par le locuteur entraînant sa reformulation. À l’inverse, les effets positifs des marqueurs de (dis)fluence peuvent correspondre, entre autres, à des points d’ancrage qui permettent aux locuteurs d’interagir, de structurer leur discours, de faire saillir les éléments informatifs, ou de produire certains effets stylistiques. De manière générale, il s’agirait donc d’un gain cognitif en termes du ratio effort/effet (dans la Théorie de la Pertinence de SPERBER et WILSON (1995) pour les participants d’un échange.

Notre définition de la (dis)fluence est également fondamentalement situationnelle, c’est-à-dire que la fluence ne s’évalue que par rapport aux attentes vis-à-vis d’une situation d’interaction particulière. Ainsi, les mêmes combinaisons de phénomènes (par exemple la multiplication des pauses silencieuses), peuvent avoir un effet tantôt fluent dans une situation formelle comme un discours politique, et tantôt disfluent dans le contexte plus informel d’une conversation spontanée. Cette ambivalence du phénomène nous porte à utiliser, dans ce document et d’autres publications liées, la forme « (dis)fluence » qui permet de ne pas se prononcer *a priori* sur l’une ou l’autre évaluation d’un élément.

Outre cette acception générale de fluence et de disfluence, il convient également de définir brièvement d’autres termes clés qui se réfèrent aux différents types de structures observées, notamment fluencème (simple et composé), locution et zone.

À l’instar de GÖTZ (2011), nous utiliserons le terme « **fluencème** » pour désigner un marqueur de (dis)fluence, sans jugement *a priori* de son caractère plutôt fluent ou disfluent. *Au niveau de l’annotation*, les fluencèmes seront répartis en deux catégories : les fluencèmes simples constitués d’une seule partie (marqueurs de discours, différents types de pauses, faux départs) et les fluencèmes composés qui requièrent de par leur nature au moins deux parties (répétition et substitution). La troncation est intermédiaire car elle peut être abandonnée (simple) ou complétée (composée).

Nous réserverons le terme « **locution** » pour des unités lexicalisées composées de plusieurs mots graphiques ou de plusieurs signes. L’expression s’appliquera à certains marqueurs de discours (par exemple : « *tu vois* », « *in other words* »). Le terme « **zone** » quant à lui sera utilisé pour parler de toute zone présentant au moins un fluencème simple ou composé, qu’il soit isolé, juxtaposé ou enchâssé. La zone de disfluence correspond donc à une portion de discours couverte par une ou plusieurs annotation(s) de fluencèmes.

2.2 Aspects techniques de l’annotation

L’annotation des corpus est essentiellement manuelle, parfois conjuguée à une part d’annotation automatique, et est soumise à peu de restrictions. Nous avons voulu prendre en compte les diversités techniques au sein des corpus. Le texte doit être segmenté au mot/signe, aligné à la bande sonore/vidéo, et consultable au sein d’une interface d’annotation (EXMARaLDA (SCHMIDT et WÖRNER 2012), ELAN (HULSBOSCH et SOMASUNDARAM 2013), PRAAT (BOERSMA et WEENINK 2014), etc.).

Ce protocole prévoit une annotation en deux couches d’annotation (tires). L’annotation des fluencèmes simples et composés est envisagée sur une tire unique, complétée quand nécessaire par une tire « diacritiques »³

Chaque élément d’un fluencème est annoté au niveau du mot ou du signe. Tous les éléments composant un fluencème portent une étiquette composée de crochets, de deux lettres majuscules et parfois de chiffres (cf. section 2.2). Par exemple, la locution "tu vois" reçoit une étiquette sur chacun des deux éléments graphiques, même s’il ne s’agit que d’un seul marqueur de discours, (le comptage n’est pas faussé grâce au système de crochets). Toutefois, au sein d’une zone de (dis)fluence, seuls sont annotés les éléments définis dans le présent protocole, c’est-à-dire les fluencèmes et certains phénomènes adjacents particuliers bien délimités dans le protocole comme certaines parenthèses ou certains termes insérés. On exclut donc de l’annotation tout autre élément situé dans une zone de (dis)fluence qui ne serait pas explicitement défini dans ce protocole.

De plus, quand cela s’avère pertinent, c’est-à-dire lors d’une zone de (dis)fluence comprenant plusieurs fluencèmes simples et composés, l’accent est mis sur une vue d’ensemble des phénomènes présents : il s’agit bien entendu d’annoter tous les fluencèmes présents pour eux-mêmes, mais aussi de repérer le fluencème composé soutenant l’ensemble de la zone de (dis)fluence⁴, celui qui joue en tant que structure principale de cette zone pour donner à celle-ci une annotation cohérente et représentative de ce qui a lieu dans le discours. Ainsi dans l’exemple suivant, le fluencème d’ordre supérieur est la répétition partielle de la structure « it’s a long process », réitérée deux fois avec des modulations internes :

Backbone : Bb_en009⁵

it’s	a	long	process	(345)	it	’s	a	long	haul
<RM0	RM0	RM0	RM0<SP0	<UP>	RM1	RM1	RM1	RM1<SP0	SP1>
				<WI>					

and	it	’s	a	very	expensive	process			
<DM>	RM2	RM2	RM2	<IL>	SP1>	RM2>			
<WI>									

Enfin, un système de crochets est appliqué à l’annotation pour délimiter les frontières d’un fluencème, qu’il soit simple ou composé. En d’autres termes, les crochets portent sur tous les types de fluencèmes, même ceux formés d’un seul élément (p. ex. une pause ou un marqueur du discours). La position des crochets est significative : les crochets ouverts « < » ne portent que sur l’étiquette (et son éventuel chiffre) à sa droite, tandis que les crochets fermés « > » ne portent que sur l’étiquette (et son chiffre) à sa gauche. Un crochet ne porte que sur une étiquette à la fois.

3. Par « annotation de diacritiques », nous entendons l’annotation d’éléments en relation hiérarchique et interdépendante par rapport à la tire d’annotation principale. En d’autres termes, il s’agit d’un second niveau d’annotation d’éléments précédemment annotés dans la tire principale pour apporter des éléments de détails sur l’un ou l’autre fluencème repéré dans la tire principale (cf. section 3.4). Ainsi, les différents types de fluencèmes seront tous annotés sur le même niveau. Par conséquent, lorsqu’un même mot est concerné par plusieurs phénomènes simultanément, il portera une double étiquette. Plusieurs étiquettes pourront dès lors être attribuées au même terme si celui-ci est impliqué dans plus d’un fluencème. Nous pensons que la richesse et la relative rapidité d’annotation en une tire est un atout non négligeable pour l’analyse contextuelle des fluencèmes. Grâce aux interfaces, il est toujours possible d’extraire automatiquement chaque catégorie de fluencèmes, annotée sur la tire principale, sur une couche d’annotation séparée afin de répondre aux différents besoins d’analyse.

4. L’annotation se fait principalement dans une perspective linéaire, mais des « allers-retours » droite-gauche rétrospectifs peuvent également être envisagés dans un souci de contextualisation.

5. Les exemples sont authentiques et la référence correspond au nom du texte dans son corpus. Les corpus utilisés pour les exemples du présent document sont : le composant francophone de LINDSEI (GILQUIN, DE COCK et GRANGER 2010), BACKBONE (KURT 2012), LSFB (MEURANT 2013).

Étiquettes	Fluencèmes	Fluencèmes (équivalent anglais)
<i>Tire (dis)fluence</i>		
UP	Pause silencieuse	Unfilled pause
FP	Pause pleine	Filled pause
S1/S2/S3	Arrêt des mains (LSFB)	Manual stop
PU	Palm-up (LSFB)	Palm Up
DM	Marqueur de discours	Discourse marker
ET	Terme explicite d'édition	Explicit editing term
FS	Faux départ	False start
TR	Troncation	Truncation
RI	Répétition à l'identique	Identical repetition
RM	Répétition avec modulation	Modifying repetition
RE	Répétition encadrante (LSFB)	Framing repetition
RG	Répétition grammaticale (LSFB)	Grammatical repetition
SP	Substitution propositionnelle	Propositional substitution
SM	Substitution morphosyntaxique	Morphosyntactic substitution
IL	Insertion lexicale	Lexical insertion
IP	Insertion parenthétique	Parenthetical insertion
<i>Tire diacritiques</i>		
AR	Articulation	Articulation
LG	Allongement	Lengthening
WI	Enchâssement de fluencème simple	Within
OR	Changement d'ordre	Order
CF	Complétion fonctionnelle	Functional completion
CT	Complétion totale	Total completion

Tableau 1 : Récapitulatif des étiquettes communes du protocole

La numérotation de certains éléments ne concerne que les différentes parties d'un fluencème composé, ainsi que les troncations répétées. En présence de fluencèmes enchâssés, les fluencèmes simples sont placés en position initiale de telle sorte que les tags UP, FP, DM, ET et TR précèdent toujours les tags de répétition et de substitution. Les phénomènes « diacritiques » (OR, AR, LG, CT, CF, WI) sont toujours annotés dans une couche d'annotation supplémentaire. Les exemples d'annotations contiennent donc toujours deux tires.

L'annotation présentée ci-après ne se veut ni contraignante ni restrictive : liberté est laissée de ne pas annoter certains phénomènes s'ils n'entrent pas dans les objectifs d'analyse d'une recherche particulière et d'ajouter des étiquettes spécifiques à des phénomènes propres à une modalité d'expression (par exemple l'annotation des pauses (cf. 3.1.1) ou des répétitions (cf. section 3.2.1 en LSFB).

3 Catégories de fluencèmes couvertes par le protocole d’annotation

Cette section définit et explicite toutes les valeurs des étiquettes d’annotation. Elle reprend par conséquent les ressemblances et divergences avec la littérature existante sur le sujet et les relations avec des travaux antérieurs. Quand aucune référence n’est mentionnée, c’est qu’à la connaissance des auteurs, la valeur n’a pas été documentée auparavant.

3.1 Fluencèmes simples

Pour rappel, les fluencèmes simples désignent des marqueurs étant constitués d’un seul élément. Ces phénomènes peuvent s’actualiser de manière isolée, juxtaposée à d’autres fluencèmes ou enchâssée dans des fluencèmes composés.

3.1.1 Pauses (UP, FP, S1, S2, S3)

Afin de ne pas calquer a priori les concepts développés pour les langues vocales (LV) sur les langues des signes (LS) avec le risque de trahir ainsi la spécificité multimodale de ces dernières, l’annotation distinguera les pauses en LV et en LS. En effet, bien que des arrêts des mains surviennent en LSFB entre les signes par relâchement des mains ou dans les signes par figement des mains, le canal non-manuel (c’est-à-dire les mouvements des yeux, du regard, des sourcils, de la bouche et de la tête) continue à communiquer de l’information. Ces arrêts des mains sont donc remplis par le canal non-manuel (NOTARRIGO, MEURANT et SIMON 2014). De plus, contrairement aux LV où l’opposition entre son et silence est claire, dans le cas des LS, les mains ne disparaissent pas. Les articulateurs sont toujours visibles : ils occupent de l’espace et prennent une certaine forme. Il est donc prématuré, voire non pertinent, de parler de pause vide ou de pause pleine en LS. L’analyse permettra de mettre en lumière le comportement des pauses en LSFB ouvrant ainsi la possibilité de comparer a posteriori certaines pauses en LS et en LV.

Pause silencieuse (UP) Cette catégorie couvre toute interruption du signal sonore au-delà d’un seuil fixé par les conventions d’annotation existantes des corpus⁶ ou fixé relativement au débit de parole du locuteur dans un texte donné⁷. Aucune distinction n’est faite a priori entre les types de pauses en fonction de leur longueur (LITTLE, OEHMEN, DUNN, HIRD et al. 2013).

Backbone : Bb_fr013

mon	ami	Georget	(315) ⁸	monsieur	Georget	Daumas	plutôt	quoi
<SP0	SP0	<RM0	<UP>	SP1>	RM1>		<ET>	<DM>
			<WI>					

LINDSEI : FR001-F

I	don	't	think	that	an	institution	(330)	really	er	necessary	is
							<UP>		<FP>		

Le protocole ne prévoit pas de distinguer les pauses (ou tout autre fluencème simple) en position enchâssée dans un fluencème composé, de celles qui ne sont pas enchâssées dans d’autres fluencèmes. Toutefois, il est possible de spécifier cette information dans la tire « diacritiques », afin de pouvoir identifier ces cas par la suite. Nous utilisons alors le tag « WI » pour *within* (cf. section 3.4.3).

6. ICE, VALIBEL, ont tous des conventions d’annotation des pauses différentes (seuils variables de 100 à 250 ms).

7. Par application d’une fonction optionnelle du logiciel d’annotation DisMo (CHRISTODOULIDES, AVANZI et GOLDMAN 2014) ou avec un système similaire à celui présenté par LITTLE, OEHMEN, DUNN, HIRD et al. (2013).

8. Dans les exemples, la durée (en millisecondes) des silences est annotée dans la couche d’annotation orthographique. Cette décision intervient afin d’homogénéiser la présentation des exemples entre les différents corpus.

Pause pleine (FP) Cette catégorie couvre l’occupation active et codifiée du signal auditif par un élément conventionnel d’appui ou de maintien de la parole (*euh* d’appui ou d’hésitation) (DUEZ 2001 ; STRASSEL 2003 ; VASILESCU, CANDEA et ADDA-DECKER 2004).

Il s’agit de toute vocalisation qui maintient le signal dans une position neutre (par exemple : *ah, eh, er, uh, um, euh, mh*). En cas d’hésitation entre une pause pleine et un allongement, nous prenons le parti de réserver l’allongement à tout schwa post-vocalique non précédé d’une pause silencieuse.

La catégorie de pauses pleines exclut l’annotation des vocalisations de type *mh, mm, mhm* relevant du back-channel mais peuvent inclure des *mm* d’hésitation et être transcrites par plusieurs formes (*erm, er, eh, etc.*).

Backbone :Bb_fr003

football	en	France	en	fait	euh	s’appuie	sur	deux
			<DM	DM>	<FP>			

LINDSEI : FR005-F

er	it	’s	eh	it	’s	our	our	project
<FP>	<RI0	RI0	<FP>	RI1	RI1>	<RI0	RI1>	
			<WI>					

Arrêt des mains en LS (S1/S2/S3) Les trois catégories utilisées pour qualifier les pauses en LSFB couvrent : l’arrêt des mains durant un signe (S1), l’arrêt des mains entre les signes (S2), les mouvements d’hésitation ou de recherche lexicale (S3). Le premier groupe inclut le figement des mains dans la position de départ (<S1:ST>) ou de fin (<S1:EN>) d’un signe. Le deuxième groupe reprend les moments où le signeur ne signe pas (S2). La configuration de ses mains n’est pas significative : mains croisées (<S2:CR>), mains le long du corps (<S2:BO>) et mains relâchées dans l’espace neutre devant le signeur (<S2:NE>). Le dernier groupe (<S3>) comprend le signe conventionnel équivalent au *euh* français en LSFB ainsi que d’autres réalisations plus idiosyncratiques. En raison des chevauchements, les S1 sont annotés sur une tire séparée des autres fluencèmes et des autres types d’arrêts (cf. Notarrigo 2014 pour plus de précisions).

*LSFB*⁹ : CLSFBE CAM1002

« oui c’est mieux de changer par exemple [320] il y a trop de signes avec des épellations par exemple le signe pour USB c’est mieux d’utiliser le signe pour clef »

OUI	AVENIR	MEILLEUR	CHANGER	PAR-EXEMPLE	(320)	TROP
				<RM0	<S2:CR>	<SP0
					<WI>	

EPELLATION	PAR-EXEMPLE	USB	MIEUX	CLEF	CLEF	
<S1:EN>		<S1:EN>			<S1:EN>	
SP0	RM1>	SP1	SP1	SP1><RI0	RI1>	

9. Pour tous les exemples concernant la LSFB, la ligne au dessus du tableau contient la traduction en français, sans ponctuation. La première du tableau ligne contient la glose de la source vidéo, c’est-à-dire qu’à chaque signe apparaissant dans le discours en LSFB est attribuée une étiquette correspondant à un lemme.

LSFB : CLSFBE JMS22076

«oui ça a eu lieu à Paris Paris euh il y a trois quatre ans à Paris»

OUI	LA-BAS	PARIS	PARIS	EUH	TROIS	QUATRE	ANS	PASSE	PARIS
			<S1 :EN>						
		<RI0	RI1><RE0	<S3 :EU>	<SP0	SP1>			RE1>

3.1.2 Palm-up en LS (PU)

Cette catégorie couvre la réalisation manuelle suivante : une ou deux mains tendues, les cinq doigts détachés les uns des autres, effectue(nt) un mouvement de rotation du poignet qui amène la paume des mains orientée(s) vers le ciel. Le palm up a lieu dans l'espace neutre devant le signeur.

Le palm-up peut recevoir plusieurs fonctions (LOON 2012) : commencer ou finir un tour de parole, montrer son attitude par rapport à ses propres propos, inviter l'interlocuteur à réagir, établir un lien entre deux unités du discours, remplir un arrêt en montrant que l'on veut conserver le tour de parole. De cette liste ressort la proximité entre les palm-ups et les marqueurs de discours (cf. NOTARRIGO, MEURANT et SIMON (2014) pour plus de précisions).

LSFB : CLSFBE JMS22076

« voilà je signe maintenant oui euh selon moi et bien quel est le but de la journée mondiale des sourds »

/	PT :PRO1	SIGNER	/	OUI	(220)	EUH	PT :PRO1	/	JMS
<PU>			<PU>		<S2 :CR>	<S3 :EU>		<PU>	

VRAI	SOURD	MONDE	JOUR	POUR	QUOI				

LSFB : CLSFBE JMS22076

« les personnes sourdes oh elles ne sont vraiment pas assez impliquées dans le monde de la surdité »

SOURD	/	PAS	ASSEZ	DANS	PLUS	MONDE	SOURD
	<PU>						

3.1.3 Marqueur de discours (DM)

Cette catégorie couvre des éléments grammaticalement hétérogènes et multifonctionnels qui sont syntaxiquement optionnels et orientés vers les processus d'interprétation. Ils fonctionnent au niveau métadiscursif en tant qu'indices pour situer leur unité-hôte dans une représentation du discours dialogiquement co-construite (BRINTON 1996; HANSEN 2006). Ils peuvent soit signaler une relation de discours entre l'unité-hôte et son contexte, soit expliciter l'organisation structurelle des segments discursifs, soit exprimer les méta-commentaires du locuteur sur sa formulation, ou soit contribuer à la collaboration interpersonnelle.

Les marqueurs de discours couvrent donc des éléments très divers comme des conjonctions de coordination et de subordination (*donc, parce que*), des adverbes (*enfin, en quelque sorte*), des particules (*ben*), etc.

Pour conserver une annotation au niveau du mot/signe, chaque élément d'une locution de marqueur de discours (par exemple *je veux dire*) portera l'étiquette de la catégorie. Cela ne signifie pas pour autant que nous accordons le statut de marqueur de discours à ces éléments pris séparément. Grâce au système de crochets (ouvert et fermé uniquement sur le dernier terme d'une locution), il est envisageable de fusionner les étiquettes attribuées à chaque terme afin d'avoir une unité unique en post-traitement (par exemple : la locution « je veux dire » sera annotée de la manière suivante <DM DM DM>).

La limite entre marqueur de discours et termes d'édition peut être parfois subtile. En conséquence, les critères suivants seront utilisés pour les distinguer pour exclure certaines locutions de la catégorie des marqueurs de discours :

- référence explicite à des troubles de l'accès lexical (ce qui exclut des expressions du type *comment, comment dire, comment dirais-je, si je peux dire*)
- faible degré de grammaticalisation (juxtaposition libre et transparence sémantique)
- présence de contenu propositionnel

Les cas limites sont, par exemple, *si vous voulez, si tu veux, if you will, je dirais, on va dire, I don't know*, qui présentent un haut degré de figement mais qui sont une référence explicite à l'acte de parole ou de pensée. Selon le contexte, ils peuvent toutefois être considérés comme marqueurs de discours. Il reste que la sélection des marqueurs de discours ne peut jamais se passer de considérations intra-linguistiques et qu'elle reste en partie intuitive quand il s'agit d'éléments aux limites de la catégorie.

Backbone : Bb_en019

wapping	which	is	/	sort	of	/	I	suppose	would	be	classified as	the	East
		<SM0	<UP>	<DM	DM>	<UP>	<DM	DM>	SM1	SM1>			
			<WI>	<WI	WI>	<WI>	<WI	WI>					

LINDSEI : FR014-F

I	don	't	know	but	in	fact	I	went	to	er	the	London	University
				<DM>	<DM	DM>				<FP>			

Backbone : Bb_fr021

loc/	euh	enfin	locales	euh	enfin	du	coin	enfin	c'est	pas	euh
<TR	<FP>	<DM>	TR>	<SP0	<FP>	<DM>	SP1	SP1>	<DM>		<FP>
	<WI>	<WI>		<WI>	<WI>						

Backbone : Bb_fr013

mon	ami	Georget	/	monsieur	Georget	Daumas	plutôt	quoi
<SP0	SP0	<RM0	<UP>	SP1>	RM1>		<ET>	<DM>
			<WI>					

LSFB : CLSFBI1905

« par exemple le premier l'ASL c'est des signes avec de l'iconicité c'est visuel »

EXEMPLE	PT :UN	FS :ASL(ASL)	LS	ICONICITE	VISUELLE	LS	PT :UN
	<DM>	<RE0	<RE0			RE1>	<DM>RE1>

LSFB : CLSFBI1905

« si tu n'as pas compris je signe »

SI	PT:PRO2	COMPRENDRE-NEG	PT:PRO1	LS
<DM>				

3.1.4 Terme explicite d'édition (ET)

Cette catégorie couvre tout élément qui indique explicitement la reconnaissance d'une disfluente par le locuteur (STRASSEL 2003 : 8). Elle exclut les marqueurs du discours et les pauses pleines qui ont leur propre étiquette. Si le terme explicite d'édition est une locution, seuls le premier et dernier mot/signe qui la composent reçoivent un crochet ouvrant et fermant. Une zone de disfluences peut inclure plus d'un terme explicite d'édition. Ces derniers ne sont annotés que lorsqu'ils se situent dans le contexte gauche ou droit d'un autre fluencème dont ils sont la reconnaissance explicite ou quand ils sont encadrés par d'autres fluencèmes. Les termes explicites d'édition peuvent être verbaux ou non-verbaux en LS.

Backbone : Bb_fr004

beaucoup	de	euh	comment	(242)	d'	exploitants
	<SM0	<FP>	<ET>	<UP>	SM1>	
		<WI>	<WI>	<WI>		

Backbone : Bb_en0014

it's	still	just	over	a	year	(186)	what	is	it ?	(311)	fifteen	months
			<SP0	SP0	SP0	<UP>	<ET	ET	ET>	<UP>	SP1	SP1>
						<WI>	<WI	WI	WI>	<WI>		

LINDSEI : FR015-F

there	is	(380)	less	no	it	's	er	less	noisier	than	during	the	night
<SP0	<RM0	<UP>	RM0	<ET>	SP1>	RM1	<FP>	RM1>					
		<WI>		<WI>			<WI>						

LSFB : CLSFBE JMS20060

« ensuite il y a des conférences qui nous informent pardon c'est confus je suis fatiguée euh »

ENSUITE	CONFERENCE	INFORMATION	/	PARDON	CONFUSION	MOI	FATIGUER	EUH
<DM>			<PU>	<ET>	<ET>	<ET	ET>	<S3>

3.1.5 Faux départ (FS)

Cette catégorie couvre les moments d'auto-interruption laissant un segment discursif syntaxiquement ou sémantiquement inachevé et abandonné et qui ne fait l'objet d'aucune reprise (Pallaud, Rauzy & Blache 2014). Aucun élément du faux départ ne doit se retrouver dans le segment suivant (pas de reprise au niveau du lemme) (BIBER, JOHANSSON, LEECH, CONRAD et al. 2000) ; dans le cas contraire, on parlera de cas de substitution morphosyntaxique (SM) et/ou propositionnelle (SP) (cf. section 3.2.2). L'étiquette s'applique au dernier mot de la séquence fonctionnelle minimale abandonnée.

Backbone : Bb_fr008

et	nous	sommes	devenus	donc	vraiment	<i>mais</i>	<i>la</i>	<i>Provence</i>
<DM>				<DM>	<FS>	<DM>		

Backbone : Bb_en014

in	the	UK	for	women	possibly	to	have	you	know	(231)	they	're	getting	married
							<FS>	<DM	DM>	<UP>				

LSFB : CLSFBE-JMS20064

« ils apprennent des nouv/ c'est comme s'ils découvrent »

OUI	APPRENDRE	OUI	NOUVEAU	COMME	DECOUVRE
			<TR><FS>		

3.1.6 Troncation (TR)

Cette catégorie couvre tout fragment de mot, qu'il soit complété (avec ou sans délai) ou abandonné. Ce phénomène révèle l'incomplétude formelle d'un morphème ou d'un mot pouvant rester incomplet ou être repris et modifié (PALLAUD et HENRY 2004). Similairement, en LSFB, une troncation est l'ébauche d'un signe interrompu. Le signe doit être reconnaissable de par la configuration des mains et la localisation initiée (même si incomplète). Le signe ainsi tronqué peut être achevé directement avec reprise du fragment, ou un peu plus loin après l'insertion d'un ou quelques signes ou gestes, ou être abandonné (HENRY et PALLAUD 2003 : 78). Dans ce dernier cas uniquement, il s'agit d'un fluencème simple, sinon dans tous les autres cas, il est composé.

Nous annotons le fragment et, le cas échéant, sa forme complétée. Un fragment abandonné porte des crochets ouvert et fermé. Sans preuve du contraire, nous partons du principe que la reprise est la complétion du fragment de mot, même si seul le premier phonème est commun au fragment et à sa reprise. Dans le cas de combinaison d'une troncation de mot avec un abandon de structure, nous faisons précéder l'élément simple avant l'élément composé, comme dans l'exemple suivant.

LSFB : CLSFBI3406

« parfois une personne est énervée alors elle me signe rapidement »

DEPEND	LS	DEPEND	PERSONNE	NERVEUX	LS
<TR	<TR	TR>			TR>

LSFB : CLSFBE-JMS22-068

« Les cours privés, tu peux y aller si tu en as besoin pour rencontrer les membres de l'ASBL et recevoir un enseignement, des explications »

COURS	PRIVE	BESOIN	ALLER	ASBL	RENCONTRER	ENSEIGNER	ENSEIGNER	EXPLIQUER
						<TR	TR>	
						<AR>		

Conformément à la définition de la répétition à l'identique (définition selon laquelle une répétition ne porte que sur un segment lexicalisé) (cf. section 3.2.1), une troncation ne peut pas être affectée par un fluencème de répétition. La succession d'un même fragment sera annoté par une numérotation, comme dans l'exemple suivant :

(Exemple construit)

la	m/	m/	maman
	<TR0	TR1	TR>

Lorsqu'un fragment est complété, cela peut avoir lieu directement, ou après une insertion lexicale ou un autre fluencème.

9. Le signeur commence le signe standard « ENSEIGNER ». Le signe démarre du corps du signeur vers l'interlocuteur. Ensuite, le signeur s'interrompt pour remplacer le signe standard « ENSEIGNER » (dans le sens de « j'enseigne ») par la version dérivée du verbe (« je reçois l'enseignement »). Les mains du signeur pour cela changent de direction et le signe est accompli de l'interlocuteur vers le signeur.

Backbone : Bb_fr018

ils	ét /	euh	ils	étaient	tout	gênés
<RM0	<TR	<FP>	RM1>	TR>		
		<WI>				

Les deux cas suivants sont des exemples où le recours à l'enregistrement sonore est nécessaire pour déterminer le statut du mot tronqué, i.e. s'il s'agit d'une erreur d'articulation ou s'il s'agit d'une substitution propositionnelle.

Backbone : Bb_fr003

intégration	de	la	poli/	de	la	population	d'origine	étrangère
	<RI0	RI0	<TR	RI1	RI1>	TR>		
			<AR>					

Backbone : Bb_en009

and	po/	after	that	it'	s	partnership
<DM>	<TR	<IL	IL	IL	IL>	TR>

3.2 Fluencèmes composés

Comme mentionné plus haut (cf. section 2.1), un fluencème est dit « composé » lorsque son fonctionnement structurel requiert au moins deux parties. Il n'est pas exclu que ces phénomènes s'appliquent à des fluencèmes annotés par ailleurs, notamment les marqueurs de discours, qui peuvent être répétés ou substitués.

Pour l'annotation des fluencèmes composés, un système de numérotation est utilisé : les numéros identiques correspondent aux mots d'une même zone répétée ou substituée ; les chiffres croissants représentent le nombre de fois que la zone a été répétée ou substituée.

3.2.1 Répétitions (RI, RM, RE, RG)

En accord avec le positionnement théorique selon lequel on ne peut pas se prononcer *a priori* sur le caractère fluent ou disfluent d'un fluencème, l'annotation porte sur tout type de répétition, même les répétitions « rhétoriques » et les répétitions causées par une intervention extérieure dans la situation d'interaction, pourvu que ce soit dans le même tour de parole. Le protocole prévoit l'annotation d'une répétition où les éléments du ré-

pété changent d'ordre par rapport à leur apparition dans le répétable. Sur la tire diacritique, l'étiquette <OR> indique un changement d'emplacement. De plus, toujours sur la tire des diacritiques, les étiquettes <CF> et <CT> signalent le degré de complétude syntaxique du segment répété ou substitué (cf. section 3.2.1 et 3.2.2).

Répétition à l'identique (RI) Cette catégorie couvre un mot ou une séquence de mots (quasi-)contigus répétés formellement à l'identique, c'est-à-dire sans plus-value sémantique (CANDEA 2000). La quasi-contiguïté se rapporte à la possibilité d'insérer un élément à contenu propositionnel faible ou nul entre le répétable et le répété, soit les possibilités suivantes : UP, DM, ET et IP (étant donné que le contenu de la parenthèse n'affecte pas le contenu des éléments répétés). Comme précisé plus haut, la répétition ne peut porter que sur des éléments propositionnels complets, en d'autres termes ni sur une troncation de mot, ni sur une pause pleine (à titre de vocalisation non lexicale). *Backbone* : Bb_fr004

suite	à	euh	suite	à	quelques	euh	comment	dire,	mauvais	résultats
<RI0	RI0	<FP>	RI1	RI1>		<FP>	<ET	ET>		
		<WI>								

LINDSEI : FR002-F

they	er	they	go	(630)	eh	they	go	to	bed
<RI0	<FP>	RI1><RI0	RI0	<UP>	<FP>	RI1	RI1>		
	<WI>			<WI>	<WI>				

LSFB : CLSFB2406

« ou ou il y a beaucoup de périodes où je rencontre trop souvent des oralistes¹⁰ »

OU	OU	/	BEAUCOUP	/	MOMENT	/
<S1 :EN>					<S1 :EN>	
<DM><RI0	<DM>RI1>	<S2 :NE>	<RE0			
				<S2 :NE>		<S2 :NE>

BEAUCOUP	/	ORAL	RENCONTRER	RENCONTRER	BEAUCOUP	
		<S1 :ST>				
RE1>	<S2 :NE>		<RG0	RG1>		

Répétition avec modulation (RM) Cette catégorie couvre un mot ou une séquence de mots (quasi-)contigüe entièrement ou partiellement répétés mais dont le contenu est modifié. Ce phénomène repose sur une définition moins restrictive que la répétition à l'identique car il reconnaît la possibilité d'une modulation syntactico-sémantique qui peut prendre la forme d'une insertion lexicale ou d'une substitution. Une RM se distingue d'une RI par la modification de son contenu ou par un ou plusieurs éléments propositionnels. Cette catégorie ne peut pas être appliquée aux marqueurs de discours et aux termes explicites d'édition. *Backbone* : Bb_en021

asian	speakers	well	no	asian	people	living	in	the	UK
<RM0	<SP0	<DM>	<ET>	RM1>	SP1>				

Backbone : Bb_en009

a	lot	of	time	a	lot	of	money
<RM0	RM0	RM0	<SP0	RM1	RM1	RM1>	SP1>
		<CF	CF>				

LINDSEI : FR002-F

they	are	(250)	they	form	circle
<RM0	<SP0	<UP>	RM1>	SP1>	
		<WI>			

Backbone : Bb_fr008

c'	était	défendu	à	l'	époque	c'	était	défendu	de	parler
<RM0	RM0	RM0	<IL	IL	IL>	RM1	RM1	RM1>		
		<CF>								

10. Sourds qui communiquent à l'aide d'une LV]

LSFB : CLSFBI1905

« ici en Europee plus dans la région de Belgique et de France on est contre la dactylogologie on n'aime pas la dactylogologie »

EUROPE	CA-VEUT-DIRE	NS :Belgique	PLUS.P	NS :Belgique	NS :FRANCE(F)
					<S1 :ST>
		<RM0	<IL>	RM1>	
ENDROIT	(320)	DACTYLOGOLOGIE	CONTRE	DACTYLOGOLOGIE-neg	
			<S1:EN>		
	<S2:NE>	<RM0	<IL>	RM1>	

LSFB : CLSFBI306

« et en plus s'il y a une petite erreur on m'accuse de cette erreur et en plus l'erreur elle est liée à un problème de communication »

PLUS	APPARAÎTRE	PETITE	ERREUR	ACCUSER	ERREUR	ACCUSER
<DM>			<RM0 ¹¹	RM0	RM1	RM1<RG0
ACCUSER	PLUS	ERREUR	COMMUNICATION	COMMUNICATION – PAS	ERREUR	COMMUNICATION – PAS
<S1 :EN>						
RM1>RG1>	<DM>	<RM0	RM0<RG0	RG1>	RM1	RM1>

Répétition encadrante en LS (RE) Cette catégorie couvre les répétitions appelées dans la littérature des LS « doubling » (KIMMELMAN 2013) ou « reduplication » (VERMEERBERGEN et VRIENDT 1994). Selon les auteurs, ce genre de répétitions appartient à des structures syntaxiques possibles en LS (par exemple, il pourrait y avoir deux places de sujet pour un seul verbe (SVS)). Elles se présentent sous la forme X Y X, selon le modèle établi par KIMMELMAN (2013). Ce type de répétitions peut toucher toutes les sortes de classes de signes (noms, verbes, termes interrogateurs, adjectifs, pronoms, adverbes, etc.) et porter sur un signe à l'intérieur d'une unité phrastique ou sur l'unité phrastique elle-même qui vient dès lors encadrer une autre unité. Nous avons donc un répétable et un répété qui encadrent un autre élément propositionnel. (Pour plus de précisions, cf. NOTARRIGO, MEURANT et SIMON (2014)).

Pour des raisons pratiques, nous ne notons pas les insertions (parenthétiques ou lexicales) entre le répétable et le répété car la répétition encadrante peut couvrir de larges empan. Dès lors tout notre texte recevrait une étiquette <IL>, ce qui rendrait l'annotation contreproductive.

LSFB : CLSFBI306

« on jouait à se battre »

RIRE	JOUER	BOXE	RIRE
<RE0			RE1>

11. Dans cet exemple, la présence d'une répétition avec modulation est motivée par la présence d'une répétition grammaticale dans uniquement l'une des parties de la répétition. Cette absence de répétition grammaticale dans l'autre partie empêche la présence d'une répétition à l'identique car la répétition grammaticale entraîne une nuance de pluriel ou d'aspect.

LSFB : CLSFBI1905

« le deuxième c'est comme le français signé ça suit la grammaire "la, signe, à, de, à " »

PT :DEUX	FRANCAIS.LS	SUIVRE	MOT	LA	LS (SI- GNER)	FS : ADEA	PT : DET	SUIVRE	PT :DEUX
<DM><RE0		<RE0						RE1>	<DM>RE1>

Répétition grammaticale en LS (RG)¹² Cette catégorie couvre les répétitions dites « morphologiques » par Filipczak et Mostowski (2013) et appelées « morphosyntactic reduplication » par WILBUR (2005). Elles prennent en charge des fonctions grammaticales : marque du pluriel, de la portée d'un verbe ou d'un aspect. Par exemple, pour exprimer le fait qu'il y a plusieurs maisons, le signeur produira n fois le signe MAISON d'affilée soit au même endroit soit en les disposant dans l'espace devant lui. Un autre cas est celui d'une action réitérée avec parfois une notion d'ennui. Par exemple, le signeur peut produire n fois d'affilée le signe TRAVAILLER pour exprimer l'idée qu'il ne cesse de travailler.

Nous acceptons une certaine variabilité entre les deux parties de la répétition : entre le répétable et le répété. En effet, "repeated signs are not necessarily identical; one occurrence can present some modulation of the citation form or be accompanied by non-manual features that are absent in the other occurrence." (VERMEERBERGEN et VRIENDT 1994). KIMMELMAN (2013) aussi constate des variations phonologiques, phonétiques et grammaticales entre les deux parties d'une répétition. Par exemple, un répété peut être articulé de manière plus relâchée et présenter un mouvement plus court et faible que le répétable. Il souligne aussi parfois la différence de la localisation du signe lors de la seconde occurrence ou du mouvement pour ajouter une information aspectuelle, modale, etc. (Pour plus de précisions, cf. NOTARRIGO, MEURANT et SIMON (2014))

LSFB : CLSFBI1905

« mais la structure est la même que le français « De, la, à » ils sont enlevés La structure des signes suit le français »

MAIS	(830)	STRUCTURE	AUSSI	FRANÇAIS	FS : ADELA	RIEN	ENLEVER
				<S1 :EN>			
<DM>	<SE :NE>	<RE0	RE0	RE0			<RG0

ENLEVER	ENLEVER	MOT	STRUCTURE	AUSSI	LS	FRANCAIS	
						<S1 :EN>	
RG1	RG2>		RE1	RE1	<IL>	RE1>	

3.2.2 Substitutions (SM, SP)

Substitution morphosyntaxique (SM) Cette catégorie couvre toute modification morphologique. La substitution morphosyntaxique s'observe au niveau du lemme complet (excluant donc les troncations). Toute altération du lemme sera notée, c'est-à-dire ajout ou changement de morphème sur base d'un lemme commun. Ces substitutions incluent également les cas d'élision. Les cas de substitution incluent souvent des cas de répétitions partielles.

Backbone : Bb_fr004

qui	présentent	les	meilleurs	euh	la	meilleure	volaille
		<SM0	SM0	<FP>	SM1	SM1>	
				<WI>			

12. Savoir s'il existe une équivalence en langue orale est un travail en cours d'élaboration.

Backbone : Bb_en022

you	can	make	have	anything	made	in	Birmingham
		<SM0	SM1	<IL>	SM1 ¹³ >		

Substitution propositionnelle (SP) Cette catégorie couvre toute reprise d'un segment discursif par un autre introduisant une nuance ou modification sémantique. La substitution peut inclure des cas de répétitions avec modulation (DUEZ 2001). En d'autres termes, cette catégorie concerne tout élément faisant l'objet d'un remplacement, y compris dans une séquence interrompue. Comme pour les catégories de répétitions, nous distinguons les substitutions portant sur des segments complets (voir tire des diacritiques) de substitutions portant sur un segment incomplet. Toutefois, cette incomplétude n'est pas notée : seule la complétude fait l'objet d'une étiquette. Le cas d'un abandon total sans reprise formelle ni sémantique est analysé comme un faux départ et non une substitution. Cette distinction permet d'évaluer différemment une phrase comme « on va euh nous allons au marché » (substitution d'une structure incomplète par une complète) de « on va au marché euh nous allons au marché black » black (substitution de structures complètes). Pour déterminer où s'arrête la substitution

(frontière droite), s'il est possible de retrouver une équivalence grammaticale mot à mot, elle sert d'empan d'annotation. Sinon l'appréciation de la délimitation est laissée à l'annotateur sur base de critères sémantiques, avec un biais minimaliste (ce qui forme un contenu informationnel complet, cohérent, au minimum). Ce choix est pris pour une plus grande homogénéité de l'annotation, et ne pas surcharger l'annotation.

Backbone : Bb_fr008

j'habitais	douze	boulevard	Jean	Jaurès	/	non	douze	boulevard	du	vieux	pont
	<RM0	RM0	<SP0	SP0	<UP>	<ET>	RM1	RM1>	SP1	SP1	SP1>
					<WI>	<WI>					

Backbone : Bb_en021

asian	speakers	well	no	asian	people	living	in	the	UK
<RM0	<SP0	<DM>	<ET>	RM1>	SP1>				
		<WI>	<WI>						

LSFB : CLSFB1306

« mais mes amis et mes parents savent que pour que je travaille bien il faut bien m'encadrer »

MAIS	AMI	PARENT	SAVOIR	BIEN	TRAVAILLER	BIEN	ENCADRER	ENCADRER
<DM>				<RM0	<SP0	RM1>	SP1><RG0	RG1>

Pour déterminer la frontière droite d'une SP, on utilise une équivalence du nombre d'unités lexicales si possible afin de limiter l'empan de la répétition (cf. exemple bb_en021).

3.3 Insertions

3.3.1 Insertion lexicale (IL)

Cette catégorie couvre tout élément lexical propositionnel inséré à l'intérieur d'un fluencème composé de répétition modulée ou de troncation. Nous ne considérons pas cet élément en tant que fluencème à part entière, mais son annotation sert à la contextualisation des fluencèmes de répétition et de troncation. Il s'agit donc d'une insertion qui vient modifier le co-texte par l'ajout d'une information propositionnelle. Une insertion peut être envisagée en contexte comme la motivation pour une répétition ou une troncation, et participe donc à la fluence

13. L'alternance « make » et « have made » pourrait être considérée comme un changement de structure plutôt que de lemme. En cas d'hésitation, la SM est privilégiée.

ou disfluency du discours. Par souci d'économie, nous regroupons entre un seul crochet ouvert et fermant tous les éléments co-occurent ensemble.

L'annotation est limitée à la frontière gauche d'un fluencème composé. L'insertion lexicale apparaissant après un crochet fermant, mettant fin au fluencème composé (répétition ou troncation), ne sera pas annotée.

L'annotation des insertions lexicales se limite à celles présentes dans le répété ou entre le répétable et le répété ou entre un fragment et sa complétion. Si un terme est présent dans le répétable mais n'apparaît plus dans le répété, ce qui correspond à ce que Shriberg (1997) appelle « deletion », il ne recevra pas d'annotation. Nous excluons ces cas de notre protocole. L'insertion, en tant que phénomène lexical apportant une modification

sémantique à un contenu préexistant, se distingue des fluencèmes simples se trouvant en position enchâssée dans une zone de fluencèmes composés. Dans ce second cas, le diacritique « WI » est utilisé (cf. section 3.4.3).

Backbone : Bb_en009

I	deal	with	disputes,	so	civil	disputes,	predominantly
			<RM0	<DM>	<IL>	RM1>	
				<WI>			

Backbone : Bb_fr008

c'	était	défendu	à	l'	époque	c'	était	défendu	de	parler
<RM0	RM0	RM0	<IL	IL	IL>	RM1	RM1	RM1>		
		<CF>								

LSFB : CLSFB1306

« tu vois comme si, par exemple comme si on écrivait en français »

VOIR	TITRE	EXEMPLE	TITRE	ECRIRE	FRANÇAIS
					<S1 :EN>
	<RM0	<IL>	RM1>		

LSFB : CLSFB1306

« c'est normal même les plus vieux se concentrent pour chercher moi je réfléchis aux réponses puis je prépare avant de les écrire à la machine »

NORMAL	PLUS	AGE	CONCENTRER	CHERCHER	PT :PRO1	CHERCHER
			<S1 :EN>			
				<RM0		<TR

PENSER	CHERCHER	REPONDRE	AUSSI	PREPARER	CLAVIER	
<IL>	TR> RM1>		<DM>			

3.3.2 Insertion parenthétique (IP)

Cette catégorie couvre tout énoncé propositionnel fonctionnant comme une parenthèse (« parenthetical aside » (SHRIBERG (1994 : 61) et STRASSEL (2003 : 42)). Les parenthèses ne sont pas considérées comme des fluencèmes en elles-mêmes. Elles ne sont annotées que si elles sont comprises dans une zone de (dis)fluence (co-texte droit ou gauche d'un fluencème ou enchâssement dans un fluencème composé). Le contenu de l'énoncé, bien que relié contextuellement aux segments adjacents, n'affecte qu'indirectement ces segments par un ajout d'information. Il reste cependant indépendant, non intégré syntaxiquement. Comme pour l'insertion lexicale, le crochet s'ouvre uniquement sur le premier terme de la parenthèse et se referme sur le dernier terme de celle-ci.

Backbone : Bb_en009

normally	would	take	you	(257)	before	you	're	a	fully
<RI0	RI0	RI0	RI0	<UP>	<IP	IP	IP	IP	IP
<OR	OR			<WI>					

qualified	solicitor	(222)	would	normally	take	you	a	minimum	of
IP	IP>	<UP>	RI1	RI1	RI1	RI1>			
		<WI>	OR	OR>					

Backbone : Bb_fr019

puisque	nous	avons	je	l'	ai	pas	dit	dans	la	présentation
<DM>			<IP	IP		IP	IP	IP	IP	IP

au	début	mais	euh	un	euh	un	sous-sol			
IP	IP	<DM>IP>	<FP>	<RI0	<FP>	RI1>				

LSFB : CLSFBI1905

« avant je moi ma vraie voie d'expression c'est la *LSFB* je trouve ça bien agréable pourquoi parce qu'avant je parlais »

AVANT	PT:PRO1	PT:PRO1	VRAI	EXPRESSION	LSFB	BIEN
<RI0	RI0	<IP	IP	IP	IP	IP
<OR	OR					

AGREABLE	POURQUOI	PARCE	QUE	PT:PRO1	AVANT	PARLER
IP	IP	<DMIP	DM>IP>	RI1	RI1>	
				OR	OR>	

3.4 Signes diacritiques

Certains phénomènes locaux ne sont pas traités comme des fluencèmes à proprement parler, mais viennent ajouter une information sur un autre élément de la chaîne linguistique. Tous les diacritiques seront annotés sur une seule tire, différente de la tire des autres étiquettes de fluencèmes.

3.4.1 Articulation (AR)

Cette catégorie couvre tout élément étant identifié comme déviant d'une prononciation « standard », « correcte » selon le locuteur. Cette erreur de prononciation doit être explicitée par le locuteur par le biais d'un marqueur (ET, DM, etc.), sans quoi elle ne sera pas annotée afin d'éviter toute référence à une norme. Les erreurs d'articulation excluent de fait une simple référence à une « norme » de prononciation absolue.

Backbone : Bb_en009

to	do	resiv/	residential	conveyancing
		<TR0	TR1>	
		<AR>		

3.4.2 Allongement (LG)

Cette catégorie couvre tout maintien d'un phonème en début, fin ou intérieur de mot, relativement à sa valeur attendue. Cette annotation est perceptive et, par conséquent, subjective. L'étiquette s'applique au mot affecté par l'allongement sans égard à la position de l'allongement dans le mot.

LINDSEI : FR002-F

it	was	in	relation	(310)	to	(480)	er	(720)	a	disease
				<UP>		<UP>	<FP>	<UP>		
					<LG>					

3.4.3 Enchâssement de fluencème simple (WI)

Cette catégorie couvre tout fluencème simple encadré par un autre fluencème composé. Par exemple, une pause au milieu d'une répétition (entre le répétable et le répété ou entre deux répétés d'un même répétable) porte le tag <WI> (pour *within*) sur la tire réservée aux diacritiques. L'information ainsi conservée permet d'extraire automatiquement les deux types d'environnements (isolé et enchâssé) pour une éventuelle analyse contextuelle des marqueurs concernés. Les insertions lexicales et parenthétique ne reçoivent jamais le trait <WI> étant donné qu'il est définitoire de cette catégorie.

Backbone : Bb_fr004

beaucoup	de	euh	comment	(231)	d'	exploitants
	<SM0	<FP>	<ET>	<UP>	SM1>	
		<WI>	<WI>	<WI>		

Backbone : Bb_en019

wapping	which	is	/	sort	of	/	I	suppose
		<SM0	<UP>	<DM	DM>	<UP>	<DM	DM>
				<WI	WI>	<WI>	<WI	WI>

would	be	classified as	the	East				
SM1	SM1>							

LSFB : CLSFB1306

« j'étais triste et frustré je me sentais frustré. mais moi c'est comme si j'avais ressenti cela avec beaucoup de retard [par rapport aux autres membres de sa famille] »

FRUSTRER	TRISTE	FRUSTRER	/	SENTIR	FRUSTRER
<RE0		RE1><RM0	<SE :NE>	<IL>	RM1>
			<WI>		

/	MAIS	PT : PRO1	TITRE	RETARD	RETARD
		<S1 :EN>			
<PU>	<DM>			<RG0	RG1>

3.4.4 Changement d'ordre (OR)

Cette catégorie couvre tout élément répété à l'identique tant au niveau formel que sémantique, mais dans un nouvel ordre syntagmatique. Puisque les éléments répétés sont identiques, hormis le changement d'ordre, ces cas font toujours partie du premier type de répétition « à l'identique » (RI). Le tag s'applique à tous les éléments affectés par le changement d'ordre, et uniquement à ceux-ci, et ce dans les deux parties du fluencème (RI0 et RI1). Les crochets s'ouvrent sur le premier des termes dans le répétable et se ferment sur le dernier des termes ayant subi un changement d'ordre dans le répété.

Backbone : Bb_en009

normally	would	take	you	before	you're	a	fully	qualified	solicitor
<RI0	RI0	RI0	RI0	<IP	IP	IP	IP	IP	IP>
<OR	OR								

would	normally	take	you	a	minimum	of			
RI1	RI1	RI1	RI1>						
OR	OR>								

LSFB : CLSFB1306

« à la mort de mon oncle ma maman mon frère et moi étions assis côte à côte maman et moi étions assis côte à côte »

PT : PRO1	MAMAN	FRERE	A-COTE	MAMAN	A-COTE	PT : PRO1	PT : PRO3	MORT
<RI0	RI0		RI0	RI1	RI1 <S1 :EN>	RI1>		<SI :ST><S1 :EN>
<OR	OR		OR	OR	OR	OR>		

3.4.5 Complétude (syntaxique) fonctionnelle et totale (CF et CT)

Cette catégorie couvre deux niveaux de complétude : une complétude fonctionnelle, minimale, qui correspond à la réalisation totale d'une séquence fonctionnelle complète (par exemple : « a lot of money a lot of money » à distinguer d'une répétition incomplète « a lot of a lot of money »), et une complétude totale, qui correspond à une unité entière, une structure de dépendance avec un prédicat et ses éléments valentiels, typiquement une proposition (voir exemple ci-dessous « I win a lot of money. I win a lot of money »). Ces deux étiquettes sont appliquées aux éléments répétés ou substitués et ne portent que sur le dernier mot de la première partie du fluencème composé. L'incomplétude (cas où répétition et substitution portent sur des segments incomplets) n'est pas annotée afin de ne pas trop charger l'annotation.

Backbone : Bb_fr021

alors	un	canton	euh	un	canton	c'est
	<RI0	RI0	<FP>	RI1	RI1>	
		<CF>	<WI>			

Backbone : Bb_en021

and	that	's	I	use	that	I	use	that
<DM>		<FS>	<RI0	RI0	RI0	RI1	RI1	RI1>
					<CT>			

Dans le cas particulier d'une répétition avec modulation accompagnée par une substitution propositionnelle (voir exemple ci-dessous), la zone de (dis)fluence recevra deux étiquettes de complétude : l'une pour la répétition du ou des mots/signes et l'une pour la répétition de la structure grâce à la substitution du dernier terme. Pour garder l'information que ces deux annotations sont liées à la même zone, un crochet sera ouvert sur la première et un autre sera fermé sur la seconde.

Backbone : Bb_en009

a	lot	of	time	a	lot	of	money
<RM0	RM0	RM0	<SP0	RM1	RM1	RM1>	SP1>
		<CF	CF>				

4 Formulation de requêtes

Le système d'annotation présenté ci-dessus permet de prévoir la formulation de requêtes par le chercheur, avec pour principe général que lors d'une requête sur un marqueur, tous les items concernés par le marqueur sont comptés, sauf les éléments internes à un item composé (p. ex. « parce que », « in other words » compté qu'une fois, et non 2 et 3 respectivement). Ce système inclut donc les phénomènes répétés : un fragment répété 3 fois sera compté comme 4 fragments. Comme toutes les étiquettes sont composées de deux lettres, une recherche par étiquette ne devrait jamais recouvrir des occurrences non souhaitées.

La combinaison des étiquettes, crochets et numéros permet d'extraire automatiquement des ensembles et sous-ensembles de phénomènes sans passer par des requêtes trop complexes. Ainsi, pour repérer tous les fluencèmes et autres éléments annotés d'un même type, il suffira de chercher la formule : [**<TAG**]. Par exemple, la requête [**<RM**] donnera le nombre de répétitions avec modulation, et localisera le premier élément de ce type de fluencème.

On peut également envisager des requêtes sur les deux tires ou sur les doubles étiquettes dans la même tire pour spécifier certains contextes. Par exemple, toutes les pauses silencieuses en position « within », c'est-à-dire à l'intérieur d'un fluencème composé, pourront être extraites en filtrant les tags **<WI>** qui s'appliquent à des **<UP>**. De même, pour extraire tous les marqueurs de discours répétés, on pourra entrer [**<DM<R**]. L'assistance de scripts automatiques plus complexes pourra régler au cas par cas des requêtes spécifiques non prévues par le protocole.

5 Conclusion

Le présent protocole a pour but d'offrir un cadre méthodologique pour l'annotation de plus d'une vingtaine de phénomènes de (dis)fluence testée sur des corpus oraux et multimodaux alignés temporellement en anglais, français et langue des signes de Belgique francophone.

Son objectif d'exhaustivité va de pair avec son caractère flexible : il a pour but de s'adapter au plus grand nombre de questions de recherche tout en fournissant une définition opérationnalisable des phénomènes étudiés. Son objectif d'impartialité théorique vis-à-vis de l'interprétation de marque comme porteuse de fluence ou disfluence élargit considérablement les types de phénomènes annotés par les protocoles antérieurs sans ajout de difficulté. Nous espérons avoir mis en avant l'existence de critères d'analyse formellement répliquables mais non restreints à une théorie syntaxique, prosodique ou sémantique prédéfinie et limitant les considérations sémantico-pragmatiques subjectives impliquées dans l'approche de la (dis)fluence. Le protocole continue d'être testé sur corpus et l'analyse de l'accord inter-annotateur est en cours d'élaboration.

Références

- BIBER, D., S. JOHANSSON, G. LEECH, S. CONRAD et E. FINEGAN. 2000. *Longman Grammar of Spoken and Written English*. Harlow : Pearson Education.
- BOERSMA, P. et D. WEENINK. 2014. *Praat : doing phonetics by computer [Computer program]. Version 5.3.80*.
- BRINTON, L. J. 1996. *Pragmatics markers in English. Grammaticalization and discourse functions*. New York : Mouton de Gruyter.
- CANDEA, 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 de Doctorat Nouveau Régime. Université Paris III Sorbonne Nouvelle. Thèse de doctorat non publiée, 1–475.
- CHRISTODOULIDES, G., M. AVANZI et J.-p. GOLDMAN. 2014. « DisMo : A Morphosyntactic , Disfluency and Multi-Word Unit Annotator An Evaluation on a Corpus of French Spontaneous and Read Speech Presentation of DisMo ».
- CRYSTAL, D. 1988. « Another look at, well, you know... » *English Today* 4.1, 47–49.
- DISTER, A. 2007. « De la transcription à l'étiquetage morphosyntaxique. Le cas de la banque de données textuelles orales VALIBEL ». Thèse de doct. Université de Louvain.
- DUEZ, D. 2001. « Signification des hésitations dans la parole spontanée ». *Revue Parole* 17-18-19.33, 113–138.
- EKLUND, R. 2004. « Disfluency in Swedish human – human and human – machine travel booking dialogues ». Thèse de doct.

- GILQUIN, G., S. DE COCK et S. GRANGER. 2010. *LINDSEI Louvain International Database of Spoken English Interchange. Handbook and CD-ROM*. Louvain-la-Neuve, Belgique : Presses Universitaires de Louvain.
- GÖTZ, S. 2011. « Fluency in Native and Nonnative English Speech : Theory , description , implications ». Thèse de doct.
- HANSEN, M.-B. M. 2006. « A dynamic polysemy approach to the lexical semantics of discourse markers (with an exemplary analysis of French toujours) ». *Approaches to discourse particles*. Sous la dir. de K. FISCHER. Amsterdam : Elsevier, 21–41.
- HENRY, S. et B. PALLAUD. 2003. « Word fragments and repeats in spontaneous spoken French ». *Gothenburg Papers in Theoretical Linguistics 90*. Sous la dir. de R. EKLUND, 77–80.
- HULSBOSCH, M. et A. SOMASUNDARAM. 2013. « ELAN - Linguistic Annotator ».
- KIMMELMAN, V. 2013. « Doubling in RSL and NGT : a Pragmatic Account ». *Interdisciplinary Studies on Information Structure 17*, 99–118.
- KURT, K. 2012. « Pedagogic corpora for content and language integrated learning. Insight from the BACKBONE Project ». *The Eurocall Review 20.2*, 3–22.
- LITTLE, D. R., R. OEHMEN, J. DUNN, K. HIRD et K. KIRSNER. Mar. 2013. « Fluency Profiling System : An automated system for analyzing the temporal properties of speech ». *Behavioral Research Methods 45.1*, 191–202.
- LOON, E. van. 2012. *What's in the palm of your hands ? Discourse functions of Palm-up in Sign Language of the Netherlands*. Thèse de Master.
- METEER, M. 1995. « Dysfluency annotation stylebook for the switchboard corpus. Linguistic Data Consortium. »
- MEURANT, L. 2013. *Corpus LSFB. Un corpus informatisé en libre accès de vidéos et d'annotations de la langue des signes de Belgique francophones. Laboratoire de langue des signes de Belgique francophone (LSFB-Lab)*.
- MONIZ, H., F. BATISTA, A. I. MATA et I. TRANCOSO. Juin 2014. « Speaking style effects in the production of disfluencies ». *Speech Communication 65*. May, 20–35.
- NOTARRIGO, I., L. MEURANT et A. C. SIMON. 2014. « Les marqueurs de (dis) fluence en Langue des Signes de Belgique Francophone (LSFB) Rapport préliminaire à l ' épreuve de confirmation ».
- PALLAUD, B. et S. HENRY. 2004. « Troncations de mots, reprises et interruption syntaxique en français parlé spontané ». *Le poids des mots*. Sous la dir. de G. PRUNELLE, C. FAIRON et A. DISTER. Louvain-la-Neuve 10-12 mars 2004 / March 10-12, 2004 Gérard : Presses Universitaire de Louvain, 707–715.
- PALLAUD, B., S. RAUZY et P. BLACHE. 2013. « Auto-interruption et disfluences en français parlé dans quatre corpus du CID ». *TIPA. Travaux interdisciplinaires sur la parole et le langage 29*.
- SCHMIDT, T. et K. WÖRNER. 2012. « EXMARaLDA ». *Handbook on Corpus Phonology*. Sous la dir. de J. DURAND, G. ULRIKE et G. KRISTOFFERSEN. Oxford : Oxford University Press, 402–419.
- SHRIBERG, E. E. 1994. « Preliminaries to a Theory of Speech Disfluencies. » Thèse de doct. University of California at Berkeley, 225.
- SPERBER, D. et D. WILSON. 1995. *Relevance. Communication and Cognition [2nd Edition]*. Oxford : Blackwell.
- STRASSEL, S. 2003. « Simple Metadata Annotation Specification », 1–43.
- VASILESCU, I., M. CANDEA et M. ADDA-DECKER. 2004. « Hésitations autonomes dans 8 langues : une étude acoustique et perceptive ». *Colloque MIDL 2004, Modélisations pour l'identification des langues et des variétés dialectales, 29-30 novembre*, 29–30.
- VERMEERBERGEN, M. et S. de VRIENDT. 1994. « The repetition of signs in Flemish Sign Language ». *Perspectives on sign language usage. Proceedings of The Fifth International Symposium on Sign Language Research*. Sous la dir. d'I AHLGREN, B. BERGMAN et M BRENNAN. England : University of Durham, 201–214.
- WILBUR, R. B. 2005. « A Reanalysis of Reduplication in American Sign Language ». *Studies on Reduplication. Empirical Approaches to Language Typology*. Berlin : Mouton de Gruyter, 595–623.