Discourse Markers and Disfluencies
Integrating Functional and Formal Annotations

Ludivine Crible
Université catholique de Louvain
1 Place Blaise Pascal, 1348 Louvain-la-Neuve, Belgium
E-mail: ludivine.crible@uclouvain.be

Abstract

While discourse markers (DMs) and (dis)fluency have been extensively studied in the past as independent phenomena, combining DM-level and disfluency-level annotations however has never been carried out before at a fine-grained level of informativeness. It is argued that the integration of formal and functional annotations, while facing a number of methodological and theoretical challenges, is not only possible and innovative (addressing the lack of consensus in the field) but also highly relevant to the investigation of form-meaning patterns. This paper reports the methodological aspects of an annotation protocol which integrates formal identification of (dis)fluency markers and a multi-layered description of discourse markers featuring, among others, semantic-pragmatic variables such as their domain and function in context. The challenges and merits of this integration are illustrated by a comparison of clustering tendencies between different functions of DMs in *DisFrEn*, a French-English comparable dataset. Quantitative results allow us to generate tentative interpretations of the relative fluency of some DM functions based on co-occurrence patterns, in line with a cognitive-functional approach to spoken language.

Keywords: discourse markers, disfluency, annotation integration

1. Introduction

Spoken language in its most natural, spontaneous forms is characterized by the frequent – yet mostly unnoticed – occurrence of so-called disfluencies, which are generally considered to be cues of ongoing processes of language production and comprehension (e.g. Alter & Oppenheimer, 2009). Disfluencies generally include filled and silent pauses, repetitions, truncations, false starts and reformulations, taking up the seminal typology by Shriberg (1994). The formal and functional diversity of these elements is the direct consequence of the multifaceted nature of the abstract constructs of fluency and disfluency, which are in fact two sides of the same coin, hence the terminological choice of "(dis)fluency markers" in the remainder of this paper, to take this ambivalence into account. Concretely, the same device (e.g. a repetition) can either be perceived as strategic or disruptive depending on a wide range of linguistic (e.g. position, co-occurrence patterns) and extralinguistic (settings, speaker profile) factors. As a result, to date, corpora annotated with (dis)fluency markers are rarely comparable since they do not always include the same types of elements, and even within one type do not always follow the same definitions. Crible (in press) has shown that the same problem applies to the functional category of discourse markers (henceforth DMs) where it has been resolved by means of a corpus-based operational definition and annotation procedure (see Crible & Zufferey, 2015). DMs can be broadly defined as syntactically optional, metadiscursive cues constraining the interpretation of discourse by signaling coherence relations, topic structure and/or interpersonal strategies (Schiffrin, 1987; Brinton, 1996). DMs are here considered to be a type of (dis)fluency markers, although their inclusion is not consensual (e.g. Eklund, 2004; Beliao & Lacheret, 2013) nor always consistent with the exhaustive definition of the category stated above (cf. the use of closed-lists of DMs in Strassel, 2003 or Pallaud, Rauzy & Blâche, 2013).

This paper reports the methodological aspects of an annotation protocol which integrates formal identification of (dis)fluency markers and a multi-layered description of DMs featuring, among others, semantic-pragmatic variables such as their domain and function in context. The challenges and merits of this integration will be illustrated by a comparison of clustering tendencies between different functions of DMs in a French-English comparable dataset, thus uncovering form-meaning patterns. In the following sections, background and key notions will be briefly outlined (Section 2); the data and annotation procedure will be presented (Section 3); results of the quantitative study will be discussed in Section 4, before concluding on some methodological perspectives (Section 5).

2. (Dis)fluency and Discourse Markers in Corpus Linguistics

2.1 Fluency and Disfluency in Native Speech

This study follows a componential approach to (dis)fluency according to which different features or markers contribute to the relative fluency of discourse depending on their frequency, combination patterns and contextual distribution. It is argued that a fluent/disfluent interpretation is always context-bound, and that all markers in the typology are potentially fluent, thus refraining from any premature bias or interpretation during the first stages of the analysis (cf. Section 3.2). Although applicable to different populations, the present definition targets native speakers, thus effectively excluding learner and pathological (dis)fluency from the scope of this paper. In the absence of a reference standard for native speakers (as opposed to the native-like target for learners), language use

in native speech can only be assessed relative to social and contextual expectations.

Most contributions to the study of L1 (dis)fluency are either corpus-based or experimental, usually focusing on specific markers in one register (e.g. Rendle Short, 2004 on filled pauses in academic discourse; Fung, 2007 on repetitions in business discourse). However, a number of exhaustive annotation campaigns in recent years have provided substantial contributions to the field, such as Shriberg (1994), Meteer et al. (1995), Besser & Alexandersson (2007), Dister (2007) or Götz (2013) among others.

2.2 Defining and Annotating DMs

DMs have been extensively studied in the past thirty years in a variety of frameworks, methods and languages, which results in a lack of consensus both at the theoretical level for defining the boundaries of the category, and at the operational level for annotating various features of their behavior in corpus data (see Crible, in press for a full review). The major disagreement probably lies in the divide between i) relational DMs or "connectives" such as semantic uses of *so* or *because*, which are sometimes excluded from the category as in Lewis (2006), and ii) non-relational DMs, most of which are speech-specific such as non-propositional uses of *you know* or *sort of*, thus absent from most written-based accounts (e.g. Sanders, Spooren & Noordman, 1992; Fraser, 1999).

Current research is mostly focused on designing cognitively valid and operationally robust categories for the annotation of DM functions, in both speech and writing[1]. This endeavour faces numerous issues, namely particularism (language- or medium-related specificities), varying granularity, poor replicability and the intrinsic under-specification and multifunctionality of language. Major corpus-based frameworks include the Penn Discourse TreeBank (Prasad et al., 2008), Rhetorical Structure Theory (Mann & Thompson, 1988; Taboada & Mann, 2006), Segmented Discourse Representation Theory (Asher & Lascarides, 2003) and other functional taxonomies for speech such as González (2005) or Cuenca (2013).

Apart from their function(s), authors have been concerned with other features of DMs behavior especially their syntactic integration. In particular, the MDMA project (Bolly et al., 2015; in press) has designed a coding scheme based on French corpus data covering the following variables: part-of-speech category, position in the dependency unit, syntactic mobility, basic semantic value, procedural vs. conceptual meaning, presence of a contiguous pause and position in the speech turn. Another ongoing endeavour is that of the Val.Es.Co group (Briz & Pons Bordería, 2010) which combines information about the position, type of host-unit and function of the DM in a corpus of Spanish conversations.

2.3 DMs as (Dis)fluency Markers

DMs are often studied for their role in cognitive processing and overall (dis)fluency, although not always in these terms (e.g. facilitating interpretation, enhancing cooperation between participants, sounding natural and convincing, etc.). In fact, experts in the study of DMs in spoken language rarely deal with this cognitive aspect of their use and functions, let alone in a crosslinguistic perspective (see Fox Tree & Schrock, 1999 for an exception on English DM oh). However, their important role in online production and their high contextual variation advocate for their treatment as (dis)fluency markers similar to filled pauses or editing expressions.

While DMs and (dis)fluency have been extensively studied in the past as independent phenomena, combining DM-level and disfluency-level annotations however has, to the author's knowledge, never been carried out before at such a fine-grained level of informativeness as what is proposed in the present model, especially regarding the annotation of syntactic and pragmatic features of DMs. It is argued that the integration of formal and functional annotations, while facing a number of methodological and theoretical challenges, is not only possible and innovative but also highly relevant to the investigation of form-meaning patterns. Concretely, this integration combines a syntagmatic or "horizontal" level (identification of (dis)fluency markers based on formal features only) with a multi-layered "vertical" level focusing on several features of DMs.

Both annotation levels were designed following the general principles of flexibility (to different languages, registers, modes, even technical formats and theoretical frameworks) and exhaustivity in the selection of observed phenomena. While exhaustivity can hardly be evaluated on open-class categories such as DMs or (dis)fluency markers, several applications of the present model to different corpora can vouch for its flexibility: the (dis)fluency typology has been tested on spoken French (Grosman, in press; Crible, in press) and English native and nonnative corpora (Dumont, 2014) as well as French Belgian Sign Language (Notarrigo, 2016); the functional taxonomy for DMs has been used by Dobrovoljc (2016) in Slovene and Gabarró-López (forthc.) in French Belgian Sign Language.

3. Data and Procedure

3.1 The *DisFrEn* Dataset

For this study, a French-English comparable dataset has been sampled from several existing corpora[2] in order to cover eight contextual settings in similar proportions in the two languages: conversations, private phone calls, face-to-face interviews, radio interviews, classroom lessons, sports commentaries, political speeches and news broadcasts. *DisFrEn* comprises 15 hours of speech and 163,620 words

[1] See the program of the ISCH COST Action IS1312 "TextLink : Structuring Discourse in Multilingual Europe" (chair L. Degand) at http://textlink.ii.metu.edu.tr/. See also the ISO 24617-8 standard for discourse relations (Prasad & Bunt, 2015).

[2] For reasons of limited space, only the main corpus in each language will be mentioned here: ICE-GB (Nelson, Wallis & Aarts, 2002) for English and VALIBEL (Dister et al., 2009) for French.

in total. It is segmented at word-level, sound-aligned and all transcriptions are provided with their audio file. All files have been formatted to be manually annotated under EXMARaLDA (Schmidt & Wörner, 2009), an open source software for multi-layered annotation, as can be seen in Figure 1.

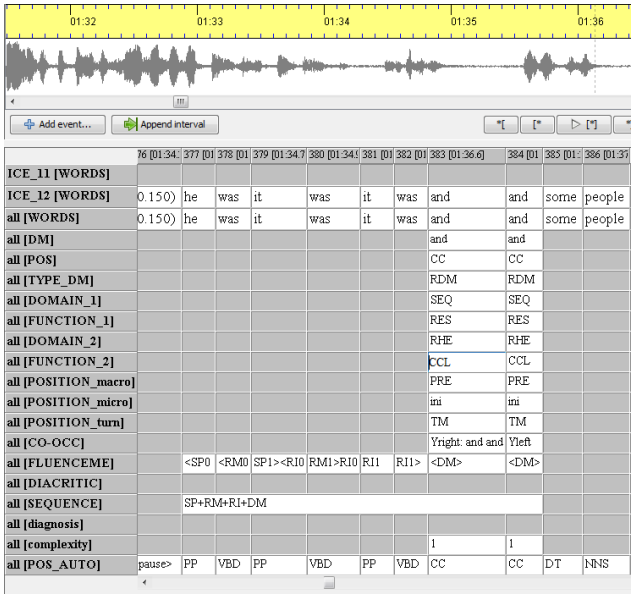

Figure 1: The EXMARaLDA interface

3.2 Formal Annotation of (Dis)fluency Markers

The "horizontal" level follows the typology and procedure detailed in Crible et al. (2016). It distinguishes between ten types of (dis)fluency markers for spoken language, as well as other secondary phenomena such as deletion or misarticulation. Their identification is entirely based on formal and structural criteria alone, making no *a priori* distinctions between potentially fluent or disfluent markers: all types of repetitions, pauses, DMs, etc. are selected. However, this annotation refrains from any judgment of grammaticality (unless explicitly noticed by the speakers themselves), in order to avoid any reference to a linguistic norm, as opposed to the approach taken by Besser & Alexandersson (2007) who use grammatical correctness as selective criterion for their "uncorrected" category. The complete list of markers can be found in Table 1.

Innovatives features include a bracketing and numbering system that enables the annotation of complex nested structures with high precision. Example 1 shows a sequence of filled and unfilled pauses, two DMs and an identical repetition. The last row summarizes the marker types by order of appearance, and it is manually assigned during the annotation.

(1)

(0.200)	uhm	so	she	and	she
<UP>	<FP>	<DM>	<RI0	<DM>	RI1>
				<WI>	
[UP+FP+DM+RI]					

The distinction between DMs, filled pauses (FPs) and editing terms (ET) is particularly relevant to the issue of integrating different levels of annotation. While some authors consider FPs such as *uhm* to be DMs (e.g. Tottie, 2015), in the present approach they had to be distinguished in order to be compatible with the literature on (dis)fluency. Similarly, some editing expressions such as *I don't know* or *oops* share some characteristics with DMs (i.e. grammatically optional, metadiscursive meaning), thus calling for more prescriptive criteria to keep the two types of markers apart.

Tags	Markers
UP	unfilled pause (seconds)
FP	filled pause
DM	discourse marker
ET	editing term
FS	false-start
TR	truncation
RI	identical repetition
RM	modified repetition
SP	propositional substitution
SM	morphological substitution
	Diacritics
AR	misarticulation
WI	embedded fluenceme
OR	change of order
	Related elements
IL	lexical insertion
IP	parenthetical insertion
DE	deletion

Table 1: Typology of (dis)fluency markers based on Crible et al. (2016).

A number of post-treatment operations were necessary to synthesize all possible combinations of markers in the data into quantitatively manageable categories. Several macro-labels with different levels of granularity, as well as numeric variables (e.g. length of the sequence, number of marker types etc.) were semi-automatically assigned to the whole dataset. The most coarse-grained categorization of (dis)fluent sequences amounts to six hierarchical levels based on the content of the sequence: DMs alone; pauses (UP/FP + DMs); interruptions (TR/FS + pauses and DMs); repetitions (RI/RM + pauses and DMs); substitutions (SM/SP + repetitions, pauses and DMs); and mixed sequences (interruptions/repetitions/substitutions + all others). These groupings made no use of DM-based annotations, which are processed separately.

3.3 Multi-layered annotation of discourse markers

The "vertical" level follows the corpus-based annotation scheme by Crible (in press). Identification and annotation of DMs are entirely manual, bottom-up and aim at exhaustivity (as opposed to closed-list selections). The annotation procedure is quite unrestrictive and allows the analyst to listen to the audio as often as necessary. Each

variable is annotated on a separate layer (cf. Figure 1). The scheme contains five formal variables and three functional ones which only apply to the DMs previously identified as part of the "horizontal" (dis)fluency annotation (i.e. any item that meets the DM definition, and nothing else). The variables are:

- formal: part-of-speech; position in the dependency structure; position in the clause; position in the speech turn; co-occurrence with another DM;
- functional: (non-)relational type; domain; function.

The most informative (at a semantic-pragmatic level) and challenging annotation is the functional interpretation based on a list of thirty values grouped in four domains (or macro-functions) which can be found in Table 2. Because of space constraints, only the four domains will be briefly defined here: ideational functions are semantic relations between real-world events; rhetorical functions express the speaker's subjectivity towards their discourse; sequential functions cover turn exchange and topic structure; interpersonal functions concern the speaker-hearer relationship. This taxonomy has been extensively tested and validated with satisfying intra-rater reliability (K= 0.74) (see Crible & Zufferey, 2015 for a discussion of inter-annotater agreement).

Ideational	Rhetorical	Sequential	Interpersonal
cause	motivation	punctuation	monitoring
consequ.	conclusion	opening	face-saving
concession	opposition	closing	disagreeing
contrast	specification	topic-resume	agreeing
alternative	reformul.	topic-shift	elliptical
condition	relevance	quoting	
temporal	emphasis	addition	
exception	comment	enumeration	
	approx.		

Table 2: Functional Taxonomy for DMs in speech based on Crible (in press).

All annotations are extracted variable by variable, using EXMARaLDA's built-in concordancer EXAKT. In order to retrieve (dis)fluency markers that are clustered in a single sequence (i.e. contiguously co-occurring markers), horizontal annotations are extracted at sequence-level, so that each line in the concordancer corresponds to one sequence, possibly containing more than one (dis)fluency markers. Through the use of unique identification codes, each sequence is semi-automatically matched with the DM(s) it contains, and vice versa, so that both annotation levels can be cross-tabulated. For instance, the sequence "so (0.480) and once" (ID code: ClaEN4t7) contains three DMs, all of which receive a corresponding ID code (ClaEN4t7DM9, ClaEN4t7DM10 and ClaEN4t7DM11). This system, while time-consuming and not required for general corpus results, is necessary for more qualitative analyses and offers a structured representation of the two annotation layers.

4. Clusters of (Dis)fluency and Discourse Markers

The annotations carried out in DisFrEn provide the material for a wide range of research questions and make it possible to draw detailed DM-specific, (dis)fluency-specific and/or function-specific profiles, by querying the dataset for any item of interest, in several registers of French and English. The integration of formal and functional annotations is particularly useful to uncover patterns of objective features that corroborate semantic-pragmatic categorization by showing contrasts and preferences. The analyses in this section will mainly make use of the six hierarchical macro-labels presented in Section 3.2 which summarize the types of (dis)fluency markers contained in a sequence, and cross-tabulate them with DM-based annotations.

Starting with the generic level of functional domains (viz. ideational, rhetorical, sequential and interpersonal), some clustering tendencies can be observed. Table 3 shows that clusters of pauses and DMs (or "P-sequences") are the most frequent type overall, followed by DMs alone ("D-sequences"), leaving the other types of more complex (dis)fluency markers to very low frequencies.

	Ideat.	Rhet.	Sequ.	Interp.	Total
Pauses	42%	42%	59%	47%	47%
DMs	45%	39%	24%	31%	36%
Repet.	7%	9%	8%	9%	8%
Interrupt.	3%	5%	5%	8%	5%
Mixed	1%	3%	2%	3%	2%
Substit.	2%	2%	2%	2%	2%

Table 3: Distribution of functional domains by (dis)fluent sequences.

The ideational domain is the only one where D-sequences are the most frequent type (44.71%). For the interpersonal, rhetorical and sequential domains, clusters of pauses and DMs are preferred. This first result could indicate that the objectivity of ideational functions seems to be paired with a high integration in the speech flow (i.e. no pause boundaries before or after the DM). On the other hand, sequential functions show the biggest proportion of P-sequences (almost 60%), which can be related to the structuring and punctuating role of these functions (e.g. topic-shift, enumerating). This clustering tendency of pauses with sequential DMs could be expected from previous findings on the high frequency of (dis)fluency markers at discourse boundaries (e.g. Roberts & Kirsner, 2000). It should be noted that these differences are observed on a subset of the data which only includes turn-medial occurrences, since turn-initial and turn-final positions necessarily exclude the possibility to co-occur with a pause at their left and right periphery, respectively.

Another potentially interesting difference is the slight reversal between D- and P-sequences across ideational and rhetorical DMs. These two domains contain several functions which only differ in their degree of subjectivity, such as cause (objective/semantic) vs. motivation (subjective/pragmatic), consequence vs conclusion,

condition vs. relevance, etc. The preference for pause clusters in rhetorical functions could be connected to well-established hypotheses on the higher cognitive complexity of subjective relations (e.g. Canestrelli, Mak & Sanders, 2013) and/or to the larger scope of these relations, usually applying to larger or more distant segments. However, any stronger conclusion would require to investigate these differences at a more fine-grained level, taking more variables into account.

In this perspective, the following results focus on specific functions, thus acknowledging the variation within a single domain. These analyses report the distribution of the top seven functions, which are the only ones with more than 200 occurrences in each language: addition (continuation in the same topic without any other value), monitoring (checking for understanding and attention), specification (elaboration with more detail or an example), opposition (pragmatic contrast or concession), temporal (chronological relation), consequence (logical effect of a previous situation) and conclusion (pragmatic result, includes summaries). Table 4 shows that only the temporal function (as well as specification, to a lesser extent) occurs more frequently as DMs alone than clustered with pauses, which could again be interpreted in terms of greater prosodic integration. On the other hand, the monitoring function shows the larger proportions of R- (repetitions) and F-sequences (interruptions), which could in turn be a symptom of disfluency, since these types of (dis)fluency markers are generally more intrusive and less directly functional than pauses.

A more fine-grained view of the content of the sequences shows that, for all seven functions, the top ten sequences are clusters of DMs, unfilled and filled pauses in various configurations (see Crible, Degand & Gilquin, in press for a detailed study of these clusters), in considerably higher proportions than all the other types of (dis)fluency markers. Only additive DMs are more frequent clustered with another marker (DM or other) than alone, while monitoring DMs stand out as rarely preceded by an unfilled pause, a configuration which is very frequent for all other functions (second most frequent type). This last result on the monitoring function can be explained by the typically final position of these markers (as in example 2), which in turn results in a higher frequency of pauses at their right periphery, compared to the other functions.

(2) I only put the alarm back twice **you know (0.567)** it's really good (PhoEN5t66)

Going a step further, this type of comparison between functions can even be refined DM by DM. For instance, looking at the occurrences of *so* signaling either a relation of consequence (objective) or of conclusion (subjective), the tendency identified above regarding the contrast between ideational and rhetorical functions seems to be confirmed: subjective *so* is much more frequently clustered with pauses (69.64%) than its objective uses (39.34%). However, the data doesn't support the general conclusion that subjective relations always trigger more clustering with (dis)fluency markers such as pauses, since it is for instance not the case for *because* which always prefers D-sequences regardless of its objective vs. subjective meaning.

5. Conclusion

This paper presented the methodological and technical aspects of the integration of two annotation layers: a horizontal annotation of (dis)fluency markers based on formal criteria, and a vertical annotation of syntactic and functional characteristics of DMs. By keeping these two levels of analysis independent during the annotation procedure, this model avoids circularity (e.g. only selecting DMs that are perceived by the analyst as disfluent), strives towards exhaustivity, and therefore provides a rich resource for the investigation of (dis)fluency and discourse markers in several registers of French and English.

The quantitative study which illustrates the analytical potential of *DisFrEn* has generated tentative interpretations of the relative fluency of some functions: ideational DMs appear to be more prosodically integrated than the other domains (i.e. fewer co-occurrences with pauses); monitoring DMs are potentially more disfluent than other functions (i.e. they co-occur more frequently with intrusive (dis)fluency markers); the objective-subjective distinction is not systematically associated with different clustering tendencies. Ongoing research is working towards the validation of these observations across language and context variation, as well as their extension to other DM functions and their refinement by positional variables.

Function	**Pauses**	**DMs**	**Repet.**	**Interrupt.**	**Mixed**	**Substit.**	**Total**
Addition	60,54%	24,06%	7,61%	3,59%	2,01%	2,19%	100,00%
Monitoring	45,42%	31,54%	9,64%	7,84%	3,43%	2,12%	100,00%
Specification	42,11%	43,42%	7,07%	4,44%	1,48%	1,48%	100,00%
Opposition	49,80%	36,00%	7,00%	3,40%	2,20%	1,60%	100,00%
Temporal	39,63%	47,43%	7,39%	3,90%	0,62%	1,03%	100,00%
Consequence	47,51%	41,21%	6,07%	1,52%	1,30%	2,39%	100,00%
Conclusion	55,20%	28,64%	8,31%	2,31%	4,39%	1,15%	100,00%
Total	**50,09%**	**34,33%**	**7,63%**	**3,98%**	**2,17%**	**1,79%**	**100,00%**

Table 4: Distribution of the Seven Most Frequent Functions by (Dis)fluent Sequences.

In addition to these methodological and empirical results, the present work offers a first step towards a more general purpose, which is to show how theoretical and methodological decisions (such as functional categorization and annotation of DMs) can be motivated by empirically-sound clusters of structural characteristics. This endeavour is in line with cognitive corpus linguistics (Arppe et al., 2010) and usage-based cognitive semantics (Glynn, 2010) whereby theory and data feed each other by validating certain abstract groupings through their actual use in native language.

6. Acknowledgements

This research benefits from the support of the ARC-project "A Multi-Modal Approach to Fluency and Disfluency Markers" granted by the Fédération Wallonie-Bruxelles (grant nr. 12/17-044).

7. Main References

Alter, A. & Oppenheimer, D. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review*, 13:219–235.

Arppe, A.; Gilquin, G.; Glynn, D.; Hilpert, M. & Zeschel, A. (2010). Cognitive Corpus Linguistics: five points of debate on current theory and methodology. *Corpora*, 5(1), pp. 1–27.

Asher, N. & Lascarides, A. (2003). *Logics of Conversation*. Cambridge: Cambridge University Press.

Beliao, J. & Lacheret, A. (2013). Disfluency and discursive markers: When prosody and syntax plan discourse. In R. Eklund (Ed.), *Proceedings of the 6th Workshop on Disfluency in Spontaneous Speech (DiSS) 2013*, 54(1), pp. 5–8.

Bolly, C.; Crible, L.; Degand, L. & Uygur-Distexhe, D. (2015). MDMA. Identification et annotation des marqueurs discursifs "potentiels" en contexte. *Discours* 16, pp. 3–32.

Bolly, C.; Crible, L.; Degand, L. & Uygur-Distexhe, D. (in press). Towards a Model for Discourse Marker Annotation. From potential to feature-based discourse markers. In C. Fedriani & A. Sansó (Eds.), *Discourse Markers, Pragmatic Markers and Modal Particles: New Perspectives*. Amsterdam: John Benjamins.

Brinton, L. (1996). *Pragmatic markers in English. Grammaticalization and discourse functions*. New York: Mouton de Gruyter.

Briz, A. & Pons Borderia, S. (2010). Unidades, marcadores discursivos y posicion. In O. Loureda & E. Acin (Eds.), *Los Estudios sobre Marcadores del Discurso*. Madrid: Acro/Libros, pp. 523–557.

Canestrelli, A.; Mak, W. & Sanders, T. (2013). Causal connectives in discourse processing: How differences in subjectivity are reflected in eye movements. *Language and Cognitive Processes* 28(9), pp. 1394–1413.

Crible, L. (in press). Towards an operational category of discourse markers: A definition and its model. In C. Fedriani & A. Sansó (Eds.), *Discourse Markers, Pragmatic Markers and Modal Particles: New Perspectives*. Amsterdam: John Benjamins.

Crible, L. & Zufferey, S. (2015). Using a unified taxonomy to annotate discourse markers in speech and writing. In H. Bunt (Ed.), *Proceedings of the 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (isa-11), April 14th, London, UK*, pp. 14-22.

Crible, L.; Dumont, A.; Grosman, I. & Notarrigo, I. (2016). Annotation manual of fluency and disfluency markers in multilingual, multimodal, native and learner corpora. Version 2.0. *Technical report*. Université catholique de Louvain and Université de Namur.

Crible, L.; Degand, L. & Gilquin, G. (in press). The clustering of discourse markers and filled pauses: a corpus-based French-English study of (dis)fluency. *Languages in Contrast*.

Cuenca, M.J. (2013). The fuzzy boundaries between discourse marking and modal marking. In L. Degand, B. Cornillie & P. Pietrandrea (Eds.), *Discourse markers and modal particles. Categorization and description*. Amsterdam: John Benjamins, pp. 191–216.

Dister, A. (2007). De la transcription à l'étiquetage morphosyntaxique – Le cas de la banque de données textuelles orales VALIBEL. PhD thesis. Université catholique de Louvain.

Dobrovoljc, K. (2016). Annotation of multi-word discourse markers in spoken Slovene. Poster presented at *Discourse Relational Devices (LPTS 2016)*, January 24-26, Valencia, Spain.

Dumont, A. (2014). Annotation of fluency and disfluency markers in nonnative spoken corpora. Paper presented at the *Interlanguage Annotation Workshop (Societas Linguistica Europaea - 47th Annual Meeting)*, September 11-14, Poznań, Poland.

Eklund, R. (2004). Disfluency in Swedish human-human and human-machine travel booking dialogues. PhD thesis. Linköping Studies in Science and Technology.

Fox Tree, J. & Schrock, J. (1999). Discourse markers in spontaneous speech: Oh what a difference an Oh makes. *Journal of Memory and Language* 40, pp. 280–295.

Fraser, B. (1999). What are discourse markers? *Journal of Pragmatics* 31, pp. 931–952.

Fung, L. (2007). The communicative role of self-repetition in a specialised corpus of business discourse. *Language Awareness* 16(3), pp. 224–238.

Gabarró-López, S. (forthcoming). The hotchpotch of buoys: A corpus study on their use across genres and discourse functions in French Belgian Sign Language (LSFB).

Glynn, D. (2010). Testing the hypothesis. Objectivity and verification in usage-based Cognitive Semantics. In D. Glynn & K. Fischer (Eds.), *Quantitative methods in cognitive semantics: corpus-driven approaches*. Berlin: De Gruyter Mouton, pp. 239–269.

González, M. (2005). Pragmatic markers and discourse coherence relations in English and Catalan oral narrative. *Discourse Studies* 77(1), pp. 53–86.

Götz, S. (2013). *Fluency in native and nonnative English speech*. Amsterdam: John Benjamins.

Grosman, I. (in press). How do French humorists manage their persona across situations? A corpus study on their prosodic variation. In L. Ruiz-Gurillo (Ed.),

Metapragmatics of Humor: Current Research Trends. Amsterdam: John Benjamins.

Lewis, D. (2006). Discourse markers in English: A discourse-pragmatic view. In K. Fischer (Ed.), *Approaches to Discourse Particles*. Amsterdam: Elsevier, pp. 43–59.

Mann, W. & Thompson, S. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text* 8(3), pp. 243–281.

Meteer, M.; Taylor, A.; MacIntyre, R. & Iver, R. (1995). Disfluency annotation stylebook for the Switchboard corpus. *Technical report*. Linguistic Data Consortium.

Notarrigo, I. (2016). Les marqueurs de (dis)fluence en Langue des Signes de Belgique Francophone (LSFB). PhD thesis. Université de Namur.

Pallaud, B.; Rauzy, S. & Blâche, P. (2013). Identification et annotation des auto-interruptions et des disfluences dans les corpus du CID. *Unpublished technical report*. Laboratoire Parole et Langage.

Prasad, R. & Bunt, H. (2015). Semantic relations in discourse: The current state of ISO 24617-8. In H. Bunt (Ed.), *Proceedings of the 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (isa-11), April 14th, London, UK*, pp. 80–92.

Prasad, R.; Dinesh, N.; Lee, A.; Miltsakaki, E.; Robaldo, L.; Joshi, A. & Webber, B. (2008). The Penn Discourse TreeBank 2.0. In *Proceedings of LREC, June 2008, Marrackech, Morroco*, pp. 2961–2968.

Rendle-Short, J. (2004). Showing structure : using um in the academic seminar. *Pragmatics* 14(4), pp. 479–498.

Roberts, B. & Kirsner, K. (2000). Temporal cycles in speech production. *Language and Cognitive Processes* 15(2), pp. 129–157.

Sanders, T.; Spooren, W. & Noordman, L. (1992). Toward a taxonomy of coherence relations. *Discourse Processes* 15, pp. 1–35.

Schiffrin, D. (1987). *Discourse markers*. Cambridge: Cambridge University Press.

Schmidt, T. & Wörner, K. (2009). EXMARaLDA – Creating, analysing and sharing spoken language corpora for pragmatic research. *Pragmatics* 19(4), pp. 565–582.

Shriberg, E. 1994. Preliminaries to a theory of speech disfluencies. PhD thesis. University of California at Berkeley.

Strassel, S. (2003). Simple metadata annotation specification v.5. *Technical report*. Linguistic Data Consortium.

Taboada, M. & Mann, W. (2006). Rhetorical Structure Theory: Looking back and moving ahead. *Discourse Studies* 8(3), pp. 423–459.

Tottie, G. (2015). *Uh* and *um* in British and American English: Are they words? Evidence from co-occurrence with pauses. In N. Dion, A. Lapierre & R. Torres Cacoullos (Eds.), *Linguistic variation: Confronting Fact and Theory*. New York: Routledge, pp. 38–54.

8. Language Resource References

Dister, A.; Francard, M.; Hambye, P. & Simon, A.-C. (2009). Du corpus à la banque de données. Du son, des textes et des métadonnées. L'évolution de la banque de données textuelles orales VALIBEL (1989-2009). *Cahiers de Linguistique* 33(2), pp. 113–129.

Nelson, G.; Wallis, S. & Aarts, B. (2002). *Exploring natural language: Working with the British component of the International Corpus of English*. Amsterdam: John Benjamins.