Discourse markers and (dis)fluency in English and French

Variation and combination in the DisFrEn corpus

Ludivine Crible
Université catholique de Louvain

While discourse markers (DMs) and (dis)fluency have been extensively studied in the past as separate phenomena, corpus-based research combining large-scale yet fine-grained annotations of both categories has, however, never been carried out before. Integrating these two levels of analysis, while methodologically challenging, is not only innovative but also highly relevant to the investigation of spoken discourse in general and form-meaning patterns in particular. The aim of this paper is to provide corpus-based evidence of the register-sensitivity of DMs and other disfluencies (e.g. pauses, repetitions) and of their tendency to combine in recurrent clusters. These claims are supported by quantitative findings on the variation and combination of DMs with other (dis)fluency devices in DisFrEn, a richly annotated and comparable English-French corpus representative of eight different interaction settings. The analysis uncovers the prominent place of DMs within (dis)fluency and meaningful association patterns between forms and functions, in a usage-based approach to meaning-in-context.

Keywords: discourse markers, disfluency, corpus annotation, usage-based, speech

1. Introduction

Spoken language is characterized by online processes of production and comprehension happening over time. A natural consequence of this temporal nature is the presence of disfluencies or ‘fluencemes’ (Götz 2013) that can generally be defined as signals of on-going mechanisms of processing and monitoring. They include pauses, discourse markers, truncations, false starts, repetitions and substitutions, and are pervasive especially in impromptu speech, although not excluded from more prepared and

monologic interactions. Fluencemes can contribute to both fluency and disfluency, respectively marked by smoothness vs. discontinuity (e.g. Ejzenberg 2000), and as such do not always involve an error or a hesitation. For instance, Candéa (2000) distinguishes between 'structuring' and 'non-structuring' pauses (cf. also Pawley & Syder 2000 on 'planned' vs. 'unplanned' cognitive pauses), which she associates to (fluent) segmentation and (disfluent) interruption, respectively. Discourse markers (henceforth DMs) such as *so* or *well* constitute another highly ambivalent fluenceme: on the one hand, linguists (especially in second language acquisition) acknowledge the positive effects of DMs on naturalness or automaticity (e.g. Hasselgren 2002, Götz 2013); on the other hand, outside academia, these expressions are often stigmatized as "cringing verbal tics", as in a 2008 article from the LanguageLog website reporting on the US Senator Caroline Kennedy and her "inexperience as a speaker" reflected by the use of more than 200 *you know*s and many *um*s in a 30-minute interview (http://languagelog.ldc.upenn.edu/nll/?p=964). This ambivalence of fluencemes requires a careful, qualitative approach to fluency and disfluency as two sides of the same coin – hence the use of '(dis)fluency' in the remainder of this article – in order to uncover the conditions under which a particular device is used rather fluently or disfluently.

(Dis)fluency and the ambivalent nature of fluencemes attract a growing interest in a broad range of methodological and theoretical frameworks. As far as corpus linguistics is concerned, many contributions to the field consist in case studies focusing on particular devices or data types. This situation therefore calls for more comprehensive studies of all types of (dis)fluent phenomena in a variety of registers. The present study addresses such a gap by presenting the data, method and results of a bottom-up, intensive-extensive approach to fluencemes covering a large number of linguistic categories for which detailed formal and functional annotations are provided, thus adding to "the small class of corpora featuring discourse and pragmatic annotation" (Rühlemann & O'Donnell 2012: 315). The corpus under scrutiny is DisFrEn, a comparable dataset of native French and English balanced across eight spoken registers.

Within the typology of fluencemes, this research pays particular attention to discourse markers, drawing an exhaustive portrait of their many forms (conjunctions like *and*, *although*, adverbs like *well*, *anyway*, verb phrases like *I mean*, *you know*, etc.) and functions, from connective meanings such as causal or contrastive relations, to

interactional meanings such as turn-taking or monitoring (Crible 2017). DMs are presently considered as one type of fluenceme, following previous proposals where DMs are included in the typology (e.g. Shriberg 1994, Strassel 2003, Götz 2013). However, in most of these works, DMs are only selected on the basis of very restricted closed lists and/or not analyzed beyond mere frequency (i.e. not functionally annotated). In DisFrEn, a very large (bottom-up, onomasiological) category of DMs is fully described in terms of their syntactic (positional) and pragmatic (functional) behavior. In addition, the surface structure of other fluencemes (e.g. pauses, repetitions, etc.) co-occurring with DMs is annotated as a complementary level of analysis, thus uncovering patterns of forms and functions of DMs and fluencemes. In sum, this paper addresses the lack of large-scale, functional, corpus-based approaches integrating DMs and (dis)fluency, as opposed to the bulk of case studies, and provides strong insights on the crosslinguistic and register variation of DMs and their combination with other fluencemes.

After a contextualisation of the present work in corpus-based studies on discourse markers and (dis)fluency (Section 2), this article describes the data and annotation method of DisFrEn (Section 3). Section 4 situates DMs within the typology of fluencemes and answers the following research questions: with what type(s) of fluencemes do DMs tend to co-occur and how do these clusters vary across languages and registers (Section 5.1)? What formal-functional patterns or schemata can be identified from the mapping of DM functions and their co-occurring fluencemes (Section 5.2)?

2. Discourse markers and (dis)fluency in corpus linguistics

Discourse markers and (dis)fluency are predominantly treated as two independent objects of study in the literature. They both have become major trends in different disciplines of linguistics such as language acquisition or psycholinguistics. This section focuses on corpus-based works investigating DMs and (dis)fluency (especially in English and French), discussing in particular those which were relevant to the design of the present approach.

2.1. Corpora annotated with DM information

The following sections discuss how previous corpus-based research addressed definition, crosslinguistic selection and functional annotation of DMs.

2.1.1 *Definition of DMs*

The study of discourse markers arose in the 1970s with the increasing interest for discourse, pragmatics and expressions functioning beyond sentence-level. The present paper adopts Schiffrin's (1987) seminal and influential definition of DMs as "sequentially dependent elements which bracket units of talk" (Schiffrin 1987: 31). In her view, DMs contribute to coherence by punctuating segments and giving instructions on how to interpret the current utterance with respect to (i) previous or upcoming material, (ii) the speaker-hearer relationship and (iii) so-called "planes of talk", that is, generic functional domains such as content (ideational structure) or attitude (participation framework). Schiffrin (1987) brings to the forefront of her definition the connectivity of DMs, which is a recurrent feature of many DM definitions, as Schourup (1999: 230) points out in his exhaustive and structured state-of-the-art, although the connection can be more or less strictly defined and can apply to different types of units (verbally expressed or not, single or multiple, contiguous or distant; see Hansen 2006).

Schourup (1999) further mentions two consensual core features of DMs, namely syntactic optionality and non-truth-conditionality. Optionality refers to the possibility to virtually remove a DM from its host unit without altering the grammaticality of the utterance. The non-truth-conditionality of DMs, in turn, implies that the utterance remains true whether or not the DM is present. This feature sets apart DMs from "content words, including manner adverbial uses of words like *sadly*, and from disjunctive forms which do affect truth-conditions, such as evidential and hearsay sentence adverbials" (Schourup 1999: 232). Another semantic aspect often mentioned in the definition of DMs is their multifunctionality, which applies at three levels: (i) the category includes items that perform a wide range of discourse functions, such as causal

relation, reformulation or topic-shift; (ii) a single DM can be polysemous and perform different functions in different contexts; (iii) a particular instance of DM can express several simultaneous meanings in one context (Crible 2017).

Very few features of DMs are unanimous while some are rather scalar or optional. Another source of disagreement between authors is terminological, opposing, for instance, the term 'discourse marker' to the more generic 'pragmatic marker', as in Aijmer (2013): the latter include very heterogeneous elements such as "connectives, modal particles, pragmatic uses of modal adverbs, interjections, routines (*how are you*), feedback signals, vocatives, disjuncts (*frankly, fortunately*), pragmatic uses of conjunctions (*and, but*), approximators (hedges), reformulation markers" (Aijmer & Simon-Vandenbergen 2011: 227). The main issue with this broad category concerns the heterogeneity of its members, since they have little or nothing in common be it on the functional or formal levels. As a consequence, such a broad class of expressions would hardly be analyzable or interpretable coherently, which motivates the present choice of 'discourse markers' for reasons of coherence in the category and feasability of the analysis.

2.1.2 *Crosslinguistic research on DMs: From case studies to full categories*

DMs are very rarely studied in onomasiological (i.e. bottom-up, not based on closed lists) approaches, especially not in multilingual spoken data, as opposed to the bulk of (contrastive) case studies providing in-depth accounts of a very limited number of expressions, such as Willems & Demol (2006) for the English-French cognates *really*/*vraiment* or Beeching (2013) on the English translations of French *quand même*. In addition to contrastive case studies, some authors have tackled larger groups of DMs which are usually semantically coherent. These works very often focus on one semantic type of connectives which they investigate in multilingual written data, such as Zufferey & Cartoni (2012) on the translation of causal connectives in English and French, or Dupont (2015) who provides a Systemic Functional account of the position of adverbs of contrast in these two languages. The main results of both these studies converge in identifying significant differences in the use and functions of connectives in English and French: for instance, Zufferey & Cartoni (2012) show that *since* and French *puisque*, although often taken as translation equivalents, present some differences related to

information structure, namely the French connective is more frequently used to relate given information, whereas the English form tends to show the reverse preference for discourse-new information.

Few studies aim at exhaustivity over the whole DM category across several languages. One of them is reported in Zufferey & Degand (in press), who carry out a multilingual annotation experiment in English, French, German, Dutch and Italian, disambiguating the discourse relations expressed by connectives as varied as *after*, *and*, *despite*, *meanwhile* or *thus* and their translations. Lopes et al. (2014) use translation spotting techniques to build a multilingual lexicon of DMs from the Europarl corpus of parliamentary debates (i.e. cleaned transcriptions of written-to-be-spoken data) in English, Portuguese, French, German and Italian. While the transcriptions in the Europarl corpus are closer to a written style than to natural speech, other crosslinguistic works have been working with spoken data as well: Kunz & Lapshinova-Koltunski (2015) compare connectives (along with co-reference and substitution) in English and German written and spoken registers and find that connectives are more affected by crosslinguistic than register variation, for instance with a higher variety of cohesive devices in German than in English. The present annotation of spoken DMs aims at an even larger coverage of expressions including typical connectives (*so*, *however*) and more speech-specific expressions (*I mean*, *well*) as will be detailed in Section 3.2.1.

2.1.3 *Annotating the functions of discourse markers*

DM research abounds with DM-specific functional classifications (e.g. Bolly & Degand 2009 on French *donc*, "so"), which are usually very fine-grained but often not generalizable to other DMs and sometimes not replicable or operational enough to be systematically applied to authentic (spoken) data. However, several frameworks have targeted a large coverage of DMs and functions, especially in writing, thus restricting the functional spectrum of DMs to relational uses as connectives of cause or contrast (e.g. the Penn Discourse Treebank 2.0, henceforth PDTB, Prasad et al. 2008) and excluding non-relational and other speech-specific functions such as face-saving or turn-taking. In the PDTB, discourse relations are classified in four main semantic groups: "temporal", "contingency", "comparison" and "expansion". These top levels include different types and subtypes of relations, such as "cause" (further divided in

"reason" and "result") or "alternative" (three subtypes). This hierarchical taxonomy was applied to the Wall Street Journal corpus of English newspaper articles.

The PDTB has been adapted to many languages, including spoken languages (e.g. Tonelli et al. 2010 in Italian), which vouches for its robustness and inter-operability. However, it was not designed to cover the additional functions that DMs express in speech. The taxonomy in González (2005), however, specifically targets oral narratives in English and Catalan and distinguishes four different top-level meanings: "ideational" (logico-semantic relations), "rhetorical" (illocutionary intentions), "sequential" (discourse structure) and "inferential" (inference facilitators). Functions such as "emphasis", "playing for time to think" or "face-threat mitigation" are exclusive to spoken language, while also grouped with typical discourse relations ("conclusion", "addition"), which also exist in writing. González (2005) finds that very few DMs have a direct formal and functional equivalent in English and Catalan (except for *well*/*bueno* and *then/llavors*), whereas functional pairing often occurs between formally dissimilar markers. Although innovative and rather exhaustive, this proposal leaves some room for improvement regarding the operational definition of the functions and their classification in coherent categories.

All in all, the field is mainly divided between authors targeting the annotation of *discourse relations*, often in written data, and those targeting the annotation of *discourse markers*, often in spoken data, each approach involving different decisions, challenges and benefits. The present approach strives to reconcile the two options by covering both discourse-relational and speech-specific DMs which are annotated following an integrated taxonomy inspired by both the PDTB 2.0 and González (2005), as will be detailed in Section 3.2.1.

2.2. Corpora annotated with (dis)fluency information

The next sections present a selection of frameworks in the corpus-based annotation of (dis)fluency devices and discuss how past projects tackled the integration of DMs and (dis)fluency.

2.2.1 *Brief overview of (dis)fluency annotation frameworks*

The major and seminal reference in the field of (dis)fluency annotation is Shriberg (1994): she is the first to propose a comprehensive typology covering any "stretch of linguistic material [which] must be deleted to arrive at the sequence the speaker 'intended'" (Shriberg 1994: 1), thus targeting rather disruptive features of spoken (English) language. She finds that disfluencies are more frequent in the initial position of sentences, a result which can be explained by the higher planning efforts at the beginning of units and has been corroborated many times, for instance by Bortfeld et al. (2001) who show that the rate of disfluencies (5.97 per 100 words on average in their corpus) increases with heavier planning demands such as unfamiliar topics or longer speech turns.

Other typologies have been proposed for different languages, such as Pallaud et al. (2013) in French, for instance. It is not sure whether these annotation systems could be applied to languages other than their original, in keeping with the great majority of frameworks: one crosslinguistic exception is Grosjean & Deschamps (1975) who focus on temporal variables (e.g. speech rate, length pauses) in English and French where they find (i) no difference in terms of speed, (ii) preferences in the syntactic position of pauses and (iii) a stronger tendency to combine filled and silent pauses in English.[1]

At a more theoretical level, most corpus-based typologies diverge on the number and types of elements they include: for instance, Besser & Alexandersson (2007: 182) use a restricted definition of disfluency as "the interruption of the syntactic or grammatical fluency of an utterance", hence limited to disruptive elements as in Shriberg (1994), while Götz (2013) considers more pervasive and multimodal aspects of language such as lexical diversity and body language. Apart from these comprehensive typologies, corpus-based research on (dis)fluency also benefits from a flourishing number of case studies, focusing on one type of fluenceme and/or data (e.g. Bouraoui & Vigouroux 2006 on truncations in task-oriented French dialogs). This very short state of the art calls for the need to capture the different uses and contexts of single fluencemes and the inter-relations between different types of fluencemes, through corpus-based annotation systems applicable across different languages and registers.

2.2.2 *Discourse markers in previous approaches to (dis)fluency*

Discourse markers are very irregularly mentioned in fluency research, either selectively (on rather arbitrary closed lists) or not at all, and authors tend to motivate this exclusion by various reasons. One recurrent justification is found in approaches to disfluencies as removable errors, where DMs are discarded on the grounds that not all uses of DMs are disfluent: some are disruptive and removable, while others are more clearly fluent and useful. This view is represented notably by Shriberg's (1994) framework where DMs are only annotated when they occur within another disfluency (cf. a similar restriction in Pallaud et al. 2013). Her typology was adapted to Swedish by Eklund (2004), who acknowledges that some DMs could be considered as disfluencies, yet he excludes them with no further justification. Overall, these authors motivate the exclusion of DMs by acknowledging their multifunctionality: the ambivalence between fluent and disfluent uses of DMs is incompatible with the rather negative view of disfluencies in these works. Another more practical reason for the exclusion of DMs is the complexity of the category, especially in the perspective of systematic corpus annotation. The challenge of DM identification is explicitly mentioned in Meteer et al. (1995) and Strassel (2003). A different perspective is taken by studies on second-language fluency such as Müller (2005), Denke (2009) or Götz (2013) who tend to focus on a small number of DMs (usually *you know*, *I mean*, *well*) selected either for their high frequency or their relevance for learners.

Two studies, both in French, include DMs in their approach to (dis)fluency, although in a purely frequentist (i.e. not functional) approach: firstly, Beliao & Lacheret (2013) find a correlation between the frequency of DMs and that of hesitations such as lengthening or crushing voice; their results also show an effect of register variation, with lower rates of both DMs and disfluencies in planned public speeches. Secondly, Boula de Mareüil et al. (2013) study a corpus of French political interviews and found that some types of disfluencies such as repetitions and discourse markers are often involved when speakers are competing for the floor, in overlapping or turn-taking contexts. These are, to my knowledge, the only attempts at integrating DMs in (dis)fluency: while their corpus-based results suggest interesting effects of register variation, the quantitative focus of both studies leaves some room for a more qualitative approach as presently undertaken.

2.3 A usage-based approach to the integration of DMs and (dis)fluency

Many functions of DMs are directly connected to (dis)fluency (e.g. reformulation, planning) and all of them both contribute to and are windows on the cognitive mechanisms of processing and monitoring. In the present approach, the category of DMs is fully considered as one type of fluenceme, since it shares with the other members of the typology (e.g. pauses, repetitions) the characteristics of functional ambivalence (with both fluent and disfluent uses) and sensitivity to register variation and planning demands. In other words, DMs as varied as *but*, *although*, *well*, or *I mean* are all assumed to contribute either to fluency, by making the structure of discourse explicit and enhancing the fluidity and naturalness of the speech, or to disfluency, by signaling error-corrections or stalling for planning purposes. DMs are not *a priori* distinguished as either fluent or disfluent: the aim is rather to extract from the annotations any meaningful pattern of association between certain functions of DMs and co-occurrence with other fluencemes. In doing so, the role of context is put to the forefront, following the requirements of a usage-based approach to meaning-in-context (Kemmer & Barlow 2000). It is believed that such an interface is necessary since DMs do not have a uniform behavior and pragmatic role. For instance, it could be expected that DMs signaling a reformulation (e.g. *I mean*) more often co-occur with truncations or substitutions than DMs signaling causal or contrastive relations. This corpus-based study will thus contribute to both DM research and fluency research by showing attractions between functional behaviors of DMs and relative degrees of fluency, which was never carried out before.

Based on previous findings in the literature discussed earlier in this paper, DMs and fluencemes are expected to be strongly affected by crosslinguistic and register variation in terms of rate and combination patterns. Such a claim will be tested at length in this study thanks to the corpus-based methodology presented in the next section. Quantitative exploration of the data will also reveal the place of DMs in the typology of fluencemes, whether their ranking depends on language and register, and what types of fluencemes DMs tend to co-occur with. This integration of DMs and fluencemes will

also be refined by taking into account specific functional configurations of DMs, in order to uncover form-meaning patterns or schemata, which will inform us of their associated degree of (dis)fluency.

3. Methodology

The present analysis of the register-sensitivity and co-occurrence of DMs and fluencemes is carried out on the annotated dataset DisFrEn. The data and annotation procedure are described in the following sections.

3.1 Dataset construction

DisFrEn is a comparable French-English corpus containing whole interactions extracted from several source corpora according to a variationist design balanced across eight interaction settings. The main source for the English data is the British component of the International Corpus of English (ICE-GB, Nelson et al. 2002), along with the BACKBONE project (Kohn 2012). The French data come from: VALIBEL (Dister et al. 2009), CLAPI (Palisse 1997), LOCAS-F (Degand et al. 2014), C-PhonoGenre (Goldman et al. 2014), Rhapsodie (Lacheret et al. 2014) and C-Humour (Grosman 2016). The main principle behind this corpus design is context variation and the balance between languages (rather than between registers). Despite the variety of sources, the final format of DisFrEn is homogeneous after a uniform technical treatment: sound-alignment with *eLite* (Roekhaut et al. 2014) and *Train&Align* (Brognaux et al. 2012), automatic part-of-speech-tagging (Schmid 1997) and generation of the annotation layers or 'tiers'. In the end, DisFrEn comprises a total of 15 hours of speech and 161,700 words. Compared to other spoken corpora featuring pragmatic annotations (e.g. the LUNA corpus, Tonelli et al. 2010; the Spoken Turkish Corpus, Demirşahin & Zeyrek 2014), DisFrEn is larger (albeit small by corpus linguistics standards). The distribution of words per register and language is presented in Table 1.

Table 1. Words per register per language in DisFrEn

Subcorpus	**English**	**French**	**Total**
conversations	17479	17432	34911
face-to-face interviews	17055	18043	35098
radio interviews	8773	8416	17189
phone calls	9747	6783	16530
classroom lessons	9425	3723	13148
political speech	8650	7824	16474
news broadcast	7046	6788	13834
sports commentaries	8237	6279	14516
Total	**86412**	**75288**	**161700**

All files from DisFrEn were manually annotated under *EXMARaLDA* (Schmidt & Wörner 2009), a highly inter-operable open-source tool for multi-layered annotation of transcribed speech (with access to the aligned soundtrack, as seen in Figure 1), which comes with its own query and concordancer tools.

[Figure 1 here]

Figure 1. Screenshot of the *EXMARaLDA* annotation interface

3.2 Annotation schemes: A bottom-up approach to corpus annotation

The annotation procedure described in the following sections is entirely manual so that all potential tokens are selected in context (i.e. without the use of a pre-defined closed list), provided they meet the criteria defined in the protocol. The full annotation guidelines are available as technical reports in Crible (2014) and Crible et al. (2016).

3.2.1 *Discourse markers: Multi-layered annotation*

The annotation of DMs starts from the bottom-up, onomasiological selection of items according to a formal-functional definition (Crible 2017) and the following prescriptive criteria: strict syntactic optionality (thus excluding prepositions such as *because of*); discourse-level scope (i.e. presence of a finite predicate); high degree of

grammaticalization (excluding idiosyncrasies such as *and all that kind of jazz*); procedural meaning expressing either a discourse relation (e.g. cause, contrast), meta-comments, a structuring function (turn or topic structure) or managing the speaker-hearer relationship. This definition includes, in effect, expressions as varied as coordinating and subordinating conjunctions, adverbs, verb phrases, etc., provided they meet the above-mentioned criteria. It excludes, however, borderline items which are presently considered as distinct pragmatic categories, namely fillers (*uh*), interjections (*ah*), response signals (*yes*), epistemic parentheticals (*I think*), tag questions (*isn't it*) and explicit editing terms (*what is it?*) – see Crible (2014, 2017) for a full discussion of these inclusions and exclusions.

Once identified, DMs are annotated across several syntactic and pragmatic variables. The complete list of tiers and tags can be found in Appendix 1. Syntactic variables are borrowed from the Model for Discourse Marker Annotation (MDMA) project (Bolly et al. 2017) and include part-of-speech tagging, a three-fold positioning system and whether or not the item co-occurs with another DM (and where). The three types of position refer to units of different natures and sizes: the turn, the dependency structure (or macro-position) and the (sub)clause (or micro-position). They are illustrated in Examples (1) to (3).

(1) I was just trying to get (1.530) the thing going *because* we've got uh
(EN-conv-06: Turn-Medial, RIGHT of the main verb, INITIAL in the clause)

(2) <turn-break> *well* she always struck me with the exception
(EN-conv-08: Turn-Initial, PRE-field, initial periphery, INITIAL in the clause)

(3) but if you count (0.347) *like* accommodation and food on top of that
(EN-phon-02: Turn-Medial, LEFT of the main verb, MEDIAL in the clause)

The functional-pragmatic variables are also three-fold and distinguish the type (relational or non-relational), the domain (ideational, rhetorical, sequential, interpersonal) and the function of the DM from a list of thirty senses reported in Table 2. The ideational domain is linked to states of affairs in the world, semantic relations

between external events (Example (4)). The rhetorical domain is linked to the speaker's meta-comments on the on-going speech and also includes relations between epistemic or speech-act events (Example (5)). The sequential domain is linked to the structuring of local and global discourse segments such as topics and turns (Example (6)). The interpersonal domain is linked to the interactive management of the exchange and the speaker-hearer relationship (Example (7)). Definitions, criteria and examples for all thirty functions are provided in Crible (2014).

(4) we don't play any more *because* it gets dark too early (EN-conv-02)

(5) they're going to uhm (1.293) they've bought a house *well* he has (0.530) and they're going to run a bookshop (EN-conv-02)

(6) here's a napkin oops (0.280) *by the way* did I mention my dustbin's been blown over (EN-conv-04)

(7) I thought that one *you know* the Brie de Meaux is quite good (EN-conv-07)

Table 2. List of functions grouped by domain

Ideational	**Rhetorical**	**Sequential**	**Interpersonal**
cause	motivation	punctuation	monitoring
consequence	conclusion	opening boundary	face-saving
concession	opposition	closing boundary	disagreeing
contrast	specification	topic-resuming	agreeing
alternative	reformulation	topic-shifting	elliptical
condition	relevance	quoting	
temporal	emphasis	addition	
exception	comment	enumeration	
	approximation		

This taxonomy is mostly inspired by the PDTB 2.0 (Prasad et al. 2008) for the structure of the protocol and the definition of relational functions, and by González (2005) for speech-specific functions like turn-opening or face-saving. A single DM can be

assigned up to two types, domains and functions simultaneously, as in Example (8) where *meanwhile* expresses both a temporal relation and a topic-shift.

(8) in advanced stages of preparing for such an operation (0.560) *meanwhile* allied war planes continue to pound Iraqi targets (EN-news-02)

The functional disambiguation of DMs following the taxonomy in Table 2 constitutes the most challenging variable to annotate in the DisFrEn corpus, given the multifunctionality of the category and of particular instances in context. Crible & Degand (in press) conduct an annotation experiment where two expert annotators apply this taxonomy on samples of conversational French and English (55 DMs in each language) and report the following scores for inter-rater agreement: Fleiss' kappa κ = 0.563 on domains (70.9% relative agreement) and κ = 0.406 on functions (44.5%). Intra-rater reliability results are higher: one expert (the author) annotated the whole corpus and re-annotated a second time a random sample of 15% of the whole corpus (1,194 DMs) after several months, and found a substantial agreement on domains (κ = 0.779, 84%) and on functions (κ = 0.74, 75.8%). These scores indicate that sense disambiguation on (spoken) DMs is quite challenging yet reliable enough to be used, after heavy training and bearing the necessary limitations in mind.

The three-fold positioning system and the extensive coverage of functions (beyond the restrictions to discourse relations in writing-based works) stand out as innovative features of this DM-level annotation protocol. Possible uses of this scheme are for instance to check for the systematic variation of functions for a single DM lexeme depending on its position, or to look for recurrent patterns of functions for all DMs in a particular position.

3.2.2 *(Dis)fluency: Word-level tagging*

The syntagmatic annotation of fluencemes follows the procedure and coding scheme designed by Crible et al. (2016) and detailed in this section. The protocol was tested on multilingual (French, English, LSFB[2]), multimodal (spoken and signed languages), native and learner data. Labels and examples can be found in Appendix 2, and fluenceme types are listed here:

i. Simple fluencemes (one-part structure): unfilled pauses (in milliseconds), filled pauses (e.g. *uhm*), DMs (e.g. *well*), editing terms (e.g. *sorry*), truncated words, false starts;
ii. Compound fluencemes (at least two-part structure): identical repetitions, modified repetitions, morphological substitutions, propositional substitutions;
iii. Diacritics: insertion of a simple fluenceme within a compound one, change of word order, misarticulation;
iv. Other: deletions, lexical and parenthetical insertions, which are used to formally distinguish between different types of sequences, for instance a lexical insertion in a modified repetition (Example (9)) or a parenthetical insertion (Example (10)):

(9) the you know the these signifier these *natural* signifiers (EN-clas-05)

(10) there was one guy very very uhm (0.750) *hate to say it* like working class (EN-conv-03)

Concretely, this annotation is entirely manual and occupies three tiers. Firstly, all the (simple and compound) fluencemes are tagged at word-level. Some words can be assigned several fluenceme tags (e.g. a DM which is part of a repetition), and some words can be left untagged in a sequence if they do not correspond to any fluenceme. Secondly, diacritics qualify fluencemes with complementary information in a separate tier. Finally, a synthetic tag summarizes the content of the whole sequence. This synthesis takes up the order of appearance of each fluenceme, once only. The articulation of these three tiers is illustrated in Table 3, with an example of simple fluencemes, an identical repetition with an embedded DM.

Table 3. Example of three-layered annotation (EN-clas-05)

transcription	(0.200)	uhm	so	she	and	she
fluencemes	<UP>	<FP>	<DM>	<RI0	<DM>	RI1>

diacritics		<WI>
synthesis	[UP+FP+DM+RI]	

Innovative features of this annotation scheme are the bracketing and numbering systems which offer the possibility to integrate all fluencemes in the same tier without losing sight of the beginning and end of intertwined structures. A more complex sequence is provided in Table 4 to show the intricate relations between tags, brackets and numbers (in vertical format for better visualization).

Table 4. Example of embedded fluencemes (EN-intr-02)

transcription	**fluencemes**	**diacritics**	**synthesis**
(0.440)	<UP>		[UP+RM+SP+RI+FP]
it	<RM0		
’s	RM0		
so	RM0		
unbelievable	<SP0		
it	RM1		
’s	RM1		
so	RM1><RI0		
(0.600)	<UP>	<WI>	
so	RI1>		
pleasing	SP1>		
(0.507)	<UP>		
uh	<FP>		
(0.440) it’s so unbelievable it’s so (0.600) so pleasing (0.507) uh			

In this example, we can see that the second *so* is both part of the modified repetition *it’s so* (RM1), which ends with the propositional substitution of *unbelievable* by *pleasing*, and the first element of an identical repetition *so so* (<RI0). Simple fluencemes are both peripheral (no diacritics) and embedded (<WI>) in this sequence.

In DisFrEn, this complete typology of fluencemes has been used to annotate about one third of the corpus, namely the subcorpora of radio and face-to-face interviews. Given the time cost of such a procedure and the particular attention paid to DMs in this research, in the rest of the corpus other fluencemes are only annotated when they occur in the direct co-text of a DM. The whole DM category is however extensively covered throughout all subcorpora. Isolated occurrences of pauses, false starts, etc. were thus excluded from the annotation (except in radio and face-to-face

interviews), as in Example (11), but such phenomena were annotated if they were part of a larger sequence of fluencemes containing a DM, as in Example (12).

(11) I went to see a new play at the *uhm (0.300)* royal court (EN-conv-02)

(12) *so who uhm (0.200) who* finds all these places (EN-conv-02)

Additional numeric information was automatically encoded to provide information on the content of the sequence: number of fluenceme types and tokens, number of DMs in the sequence, number of annotated graphic words (i.e. length of the sequence).

In sum, DisFrEn comprises the bottom-up identification of DMs and their surrounding fluencemes in the whole corpus, and of all other fluencemes in one third of the data, thus providing a flexible resource at either level of analysis. DisFrEn offers a very rich dataset featuring the combination of multiple numeric and categorical variables at different levels of analysis.

4. Results

The research questions on register sensitivity and clustering of DMs and fluencemes are answered in the following sections, combining purely quantitative findings on distribution and ranking with more qualitative interpretations of form-function patterns and their degree of (dis)fluency. The potential of the present corpus and annotations is illustrated by – although not restricted to – these analyses, in addition to addressing the gap on crosslinguistic and register variation of DMs and (dis)fluency.

4.1. Quantitative analysis: Variation of DMs and fluencemes

The large coverage and bottom-up approach to DMs and fluencemes adopted in this study provides a comprehensive view of these categories and the relative distribution of their members. Starting with the top-five most frequent DMs in each language (all

registers combined), the data reveal a striking similarity at each rank between the English and French forms in terms of their most frequently assigned values for function and position:

i. *And* (N = 1153) / *et* (N = 870): while these two conjunctions are both tagged as "addition" (generic coordination) in 57% of all their occurrences, they are also highly multifunctional with 18 and 17 different labels, respectively;
ii. *But* (N = 513) / *mais* (N = 577): the two typically contrastive conjunctions are used in the rhetorical domain as "opposition" (epistemic or speech-act contrast) in 43% of the cases, followed by the ideational equivalent of this relation, viz. "concession" (counter-expectation), in 30% and 20%, respectively;
iii. *So* (N = 460) / *donc* (N = 315): again, the basic meaning of these conjunctions is observed in 46% and 50% of their occurrences as "conclusion" (epistemic or speech-act consequence), followed by its ideational equivalent "consequence" (logical effect between facts) in 29% and 19%, respectively, while they also perform other functions such as "topic-resuming" and "specification" (about 10%);
iv. *Well* (N = 324) / *alors* (N = 285): highest proportion of turn-initial position in the top 5 DMs of each language (54% and 17%, respectively), albeit with a greater polysemy for the French form;
v. *You know* (N = 207) / *hein* (N = 260): highest proportion of final (micro-)position in the top 5 DMs (30% and 86%, respectively), almost exclusively tagged as "monitoring" (common-ground and attention check).

Turning to fluencemes, Table 5 shows that, in radio and face-to-face ("ftf") interviews (where all fluencemes have been annotated), unfilled pauses show an outstanding proportion of almost half (41.98%) of all (dis)fluent devices, followed by DMs (29.77%) and filled pauses (10.89%). This finding confirms the ambivalence of DMs and points to their highly prominent place in the typology of fluencemes.

Table 5. Relative frequency of fluencemes in interviews per thousand words

Fluenceme	**English**	**French**	**Total**	**%**

	ftf	**radio**	**ftf**	**radio**	**ftf**	**radio**	
unfilled pause	110.88	78.31	87.62	64.52	98.92	71.56	41,98
discourse marker	62.86	54.6	71.88	52.16	67.5	53.41	29,77
filled pause	30.96	13.56	22.83	19.84	26.78	17.45	10,89
identical repetition	11.84	16.98	14.96	17.94	0.57	16.64	4,24
truncation	5.34	5.02	4.77	6.06	4.87	5.53	2,56
modified repetition	4.98	3.53	5.82	6.18	5.30	4.83	2,49
false start	3.87	3.19	7.43	6.18	13.45	4.65	4,46
prop. substitution	3.05	1.94	2.94	2.26	5.58	2.09	1,89
morph. substitution	0.76	0.34	2.94	2.85	3.39	1.57	1,22
editing term	0.18	0.11	0.94	0.24	1.88	0.17	0,50
Total	234.71	177.59	222.13	178.23	228.25	177.9	100

Broadcasting (radio vs. face-to-face) has no major effect on the relative frequency of fluencemes as far as these data are concerned. Crosslinguistically, however, we see that unfilled pauses are the only fluenceme consistently more frequent in English than in French across the two interview settings, while DMs and identical repetitions are significantly more frequent in French than in English. In other words, English and French seem to favor different types of fluencemes although, overall, the total number of fluenceme tokens is not significantly different between the two languages once the two interaction settings are combined (5561 vs. 5508; log-likelihood = 3.14, $p > 0.05$).

In the whole DisFrEn corpus, the most frequent patterns of combination are various configurations of DMs with filled and unfilled pauses (Examples (13) and (14)). This type of sequence is overwhelmingly represented in the data: 3,055 out of 7,244 sequences (i.e. 42.17%) contain DMs and pauses. This substantial proportion leads to a categorization of this pattern as a particular schema “DM + pause” of which more than 10 specific configurations have been identified (e.g. filled pause + DM or unfilled pause + DM + filled pause, see Crible et al. 2017 for an in-depth contrastive study of these clusters).

(13) push out of the way *(0.827) but I mean uh* again a lot of people (EN-clas-02)

(14) I was in here quite early *(0.773) so (0.367) you know* really I only (EN-phon-05)

Sequences of DMs and pauses are overwhelmingly frequent across all registers, yet a statistical analysis of Pearson's residuals shows that they are particularly associated to settings where only one speaker tends to hold the floor as in classroom lessons, face-to-face interviews or political speeches (i.e. monologues or non-interactive settings). Combinations of DMs with other, less frequent fluencemes such as truncations or substitutions also tend to favor specific registers: for instance, false-starts and truncations are more represented in conversations, which might reflect the effect of low planning and high interactivity on speakers' interruptions (Example (15)); repetitions are particularly frequent in radio interviews, relatively to the other seven registers, which could indicate a "radio style" whereby speakers tend to repeat themselves either for rhetorical or stylistic effects (cf. Simon et al.'s (2010) notion of phonostyle).

(15) a lot of people couldn't think of the *wo- I mean I (0.540) after* I'd done this for a hundredth time I know exactly the words (EN-conv-03)

On the whole, patterns of co-variation between sequence types and registers point to a divide between formal registers on the one hand, where sequences are rare and mostly restricted to DMs and pauses, and informal or intermediary registers on the other showing a greater diversity of fluencemes including more disruptive ones, as in the previous example.

5.2 Qualitative analysis: Usage-based schemata of forms and functions

More fine-grained analyses involving functional annotations can specify the combination patterns identified above. In the data, each domain (viz. ideational, rhetorical, sequential, interpersonal) tends to prefer the co-occurrence of specific fluencemes:

i. Ideational DMs very often do not co-occur with other fluencemes (46%, Example (16)) and especially not with false-starts or truncations (only 3%);

ii. Rhetorical DMs show a relatively strong association to mixed sequences of repetitions with truncations and/or false-starts (Example (17)) compared to the other domains;
iii. Sequential DMs are very strongly associated to pauses (Example (18)) and are relatively less frequent as isolated DMs compared to the other domains;
iv. Interpersonal DMs are only distinguished from the other domains by their positive association to false-starts and truncations (Example (19)).

(16) I know exactly the words people are trying to find *but* I'm trying not to prompt them (EN-conv-03)

(17) and (1.020) that doesn't really *I mean* I never had a career you mention the word career I have to say I never had a career (0.680) I n- didn't even have a career in Ken Russell's films (EN-intr-04)

(18) those institutions have the exercise of public power even by private bodies (1.740) *now* (0.260) we've said so far that it consists of the constitutions consists of rules (EN-clas-03)

(19) I'm not aware of it but I will keep my *you know* somebody may be doing the dirty on me (0.250) behind my back (EN-conv-05)

We can interpret these associations of DM functions and their co-occurring fluencemes in terms of (dis)fluency degrees, by decreasing order of fluency: (i) sequential DMs and their attraction to pauses indicate major discourse boundaries and perform segmentation functions; (ii) ideational DMs are well integrated in the speech flow (isolated uses) and do not often occur in the context of interruptions; (iii) the attraction of rhetorical DMs to mixed sequences might indicate their lower degree of fluency; (iv) interpersonal DMs are associated to disruptions of utterances in the form of truncations and false-starts.

Zooming in from the four domains to the thirty functions, we can expect specific senses to be conceptually related to disfluency by definition. Three functions can be identified in this respect: "reformulation" covers both paraphrases for clarification and

corrective reformulations; the role of "punctuation" is similar to written commas as floor-holders for segmentation or planning purposes; "monitoring" includes common ground, calls for attention and comprehension checks. Their hypothesized low fluency can first be tested through register information (presence or absence in formal registers): in the data, reformulative, punctuating and monitoring DMs are quasi-absent from very formal broadcast settings such as political speeches and news broadcasts and are, by contrast, strongly associated to conversations (42% of all their occurrences compared to 26% for all other 27 functions combined) and, to a lesser extent, to phone calls (17% vs. 12%).

The disfluency of these three functions is also confirmed by their association with specific types of fluencemes: compared to the other functions in the taxonomy, "monitoring", "reformulation" and "punctuation" show more occurrences than expected in sequences with false-starts, truncations, substitutions and combinations thereof. Each of the three functions is related to disfluency in their own way, either by introducing a nuance or correction ("reformulation", Example (20)), calling for cooperation and help during lexical access trouble ("monitoring", Example (21)), stalling for planning and maintaining the floor during a very long pause ("punctuation", Example (22)).

(20) I mean she she *wrote the book but uh or wrote the the chapter in the book (0.600) but (0.333)* it was after (EN-conv-01)

(21) the local councillors *etcetera have have have uh (0.450) you know have* supported us all the way through (EN-intf-02)

(22) I've long been inured to Felicity and her (2.600) pantheon of (0.410) achievements *(0.220) but uhm (1.710) I wasn't I wasn't* put out when she was (0.293) you know (1.540) sitting taking I don't know ten O-levels (EN-conv-08)

In sum, the mapping of DM functions with their co-occurring fluencemes allows to interpret their relative degree of fluency, either at a generic level (domains) or in more fine-grained associations (three potentially disfluent functions), which illustrates the

analytical potential of integrating DMs and other (dis)fluent devices in a quantitative-qualitative analysis of their variation and combination.

5. Summary and discussion

This paper has presented the data and method of the DisFrEn corpus where a large, bottom-up category of discourse markers is manually annotated, following fine-grained coding schemes covering position in three unit types, meaning-in-context from a taxonomy of four domains and thirty senses, and co-occurrence with other fluencemes such as pauses or repetitions. Applications of this resource were illustrated by a quantitative-qualitative analysis of the crosslinguistic and register variation of DMs and fluencemes across eight interaction settings in English and French, as well as an investigation of the relation between DMs and other fluencemes in the typology. The main findings include the prominent place of DMs within this typology (2nd most frequent fluenceme overall), their frequent clustering with filled and unfilled pauses suggesting an abstract "DM + pause" schema, a tentative scale of fluency for the four functional domains, and the identification of three potentially disfluent functions (viz. "monitoring", "reformulation" and "punctuation").

The identification of generalized schemata or patterns of form and function situated in discourse makes a strong case for the bottom-up corpus-based study of complex pragmatic categories. Schemata can be drawn from corpus-based annotation only if the phenomena have been approached both extensively (bottom-up identification) and intensively (multi-layered annotation). Robust annotation schemes, applicability to various data types and extensive coverage of the phenomena at stake all vouch for the methodological soundness of the present study of (dis)fluency and DMs. Further research is needed on larger samples to evaluate the statistical robustness of the observed patterns (see below for more avenues).

This paper has illustrated only a handful of applications of this approach to the integration of DMs and (dis)fluency. Such a large-scale yet fine-grained investigation of the (dis)fluency of DMs is highly innovative against the bulk of case studies approaching DMs and (dis)fluency as two separate phenomena, when our results in fact

show their frequent clustering. In particular, DisFrEn provides a speech-based annotation framework for DMs covering a large number of expressions and functions, as opposed to the more restricted lens of previous studies and to the usual focus on discourse relations in written data in the literature. The method and analytical potential of DisFrEn therefore contribute to corpus-based pragmatics by addressing the challenge of discourse annotation in spoken data, applied to a particularly tricky object of study, namely DMs, a multifunctional and heterogeneous category which still attracts the interest of many linguists to this day.[3]

DMs are themselves considered as part of the category of fluencemes, those multi-faceted structures that arguably constitute windows onto the mental processes of speech. If (dis)fluent devices are so pervasive in natural language, then comprehensive theories of speech production should account for these stretches of talk when things go wrong and explain "how a natural language handles its intrinsic troubles" (Schegloff et al. 1977: 381). Many case studies have contributed to furthering our knowledge of individual phenomena or restricted populations. Nonetheless, a great deal remains to be done so that various methods and types of evidence communicate and converge in the perspective of theory-building. The present proposal hopefully strives in this direction.

In DisFrEn, speech is studied from the point of view of production, revealing context-sensitive strategies of discourse-structuring and online planning. Nevertheless, any corpus-based approach remains silent as to the salience and entrenchment of these strategies in the speakers' cognitive systems, and even more so to the perception of these strategies by the interlocutor. Regarding the former (cognitive salience), several authors have explored the link between observed frequency and prototypicality (Gilquin 2006, Schmid 2010), and they seem to agree on a rather pessimistic conclusion that these notions do not directly overlap. The latter (perception of fluency) is affected by a wealth of context-sensitive factors (e.g. topic familiarity, relationship between speakers, accent, even physical parameters), some of which are hard to predict or even quantify objectively.

This gap between language and mind thus questions the legitimacy of corpus linguistics to access human cognition. Frequency of use of a set of structures or expressions in a particular corpus cannot be interpreted as a cue to their relative fluency. The functional annotation of DMs in DisFrEn takes us closer to filling this gap by

striving to retrieve the speaker's intended meaning as objectively as possible. For example, instances of DMs with a reformulative meaning in context (typically *I mean*) can be distinguished in terms of their relative fluency (paraphrastic, clarifying use) or disfluency (corrective use) depending on the structure they appear in: a short modified repetition of a few words separated by filled pauses (e.g. *the big cat uh I mean the big uh dog*) is likely to be more disfluent than an isolated DM connecting syntactically complete segments (e.g. *John is gone I mean he's on holiday*). Of course these examples are speculative and would need to be empirically tested, but the point remains that such distinctions can be made from a qualitative-quantitative analysis of DisFrEn.

Going a step further and directly connecting production data to perception data necessarily involves a multi-method approach, combining corpus and experimental linguistics, as recently advocated by some authors (e.g. Gilquin & Gries 2009). While psycholinguistic experiments such as response times or eye-tracking tests can be used to test corpus-based claims, their scope tends to be restricted to one or two structures in very controlled settings, as opposed to the large coverage and authentic nature of corpus data, thus arguing for the necessity of corpus linguistics in theory-building.

On a more epistemological note, this paper brings up the issue of the analyst's subjective involvement in their research, here in the form of heavy pragmatic annotation. Manual paradigmatic annotation of (dis)fluency and discourse markers is much more time-consuming than automatic identification systems or closed-list selection methods (e.g. extracting all occurrences of *I mean* in the data, sometimes with manual disambiguation afterwards). It requires the elaboration of operational coding schemes, linguistic expertise and extensive training on corpus data. The final annotation itself is a long process (partly dependent on corpus size), especially given the complexity of the phenomena at stake, which necessarily restricts the size of the dataset unless a larger team of expert annotators can be gathered, as is the case for very large projects such as the Prague Dependency Treebank (Zikánová et al. 2015). In its present state, DisFrEn does not meet the current "big data" trend in corpus linguistics, but this limitation is more related to the practical conditions of its construction, rather than theoretical motivations, and therefore does not exclude in principle application to larger corpora.

The last question raised by this paper is that of objectivity in the human sciences. Linguistics is a social science, however quantitatively oriented and objective particular studies can claim to be. Its object of study is very flexible in nature and affected by many factors that escape the analyst's control, contrary to laboratory sciences. In particular, meanings and perceptions are not fixed in absolute repertoires but rather emerge from a combination of co(n)textual variables. This does not exclude some systematicity in language, but I would like to argue for a flexible, more involved approach to the data for any cognitive-functional theory. With adequate methodological precautions, manual annotation can render the speaker's experience more faithfully than a purely automatic and quantitative treatment of the data, which presents the risk of jumping to premature conclusions. In a second step, pragmatic annotation can be mapped with more formal variables and reveal significant patterns of association. In sum, DisFrEn benefits from the soundness of quantitative methods while allowing for the flexibility of more qualitative analyses.

Acknowledgments

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). The author would like to thank the editors and anonymous reviewers for their insightful comments. Any remaining mistakes are mine.

Notes

1. Another possible exception is Eklund & Shriberg (1998), although they seem to have merged two pre-existing language-specific typologies.

2. Belgian French Sign Language.

3. See, for example, the program of the ISCH COST Action IS1312 "TextLink: Structuring Discourse in Multilingual Europe", chair Pr. Liesbeth Degand, http://textlink.ii.metu.edu.tr/.

References

Aijmer, K. (2013). *Understanding Pragmatic Markers: A Variational Pragmatic Approach*. Amsterdam: John Benjamins.

Aijmer, J., & Simon-Vandenbergen, A.-M. (2011). Pragmatic markers. In J. Zienkowski, J.-O. Östman & J. Verschueren (Eds.), *Discursive Pragmatics* (pp. 223-247). Amsterdam: John Benjamins.

Beeching, K. (2013). A parallel corpus approach to investigating semantic change. In K. Aijmer & B. Altenberg (Eds.), *Advances in Corpus-based Contrastive Linguistics. Studies in honour of Stig Johansson* (pp. 103-125). Amsterdam: John Benjamins.

Beliao, J., & Lacheret, A. (2013). Disfluency and discursive markers: When prosody and syntax plan discourse. In R. Eklund (Ed.), *Proceedings of Disfluency in Spontaneous Speech (DiSS) 2013*. *TMH-QPSR, 54*(1), 5-8.

Besser, J., & Alexandersson, J. (2007). A comprehensive disfluency model for multi-party interaction. In S. Keizer, H. Bunt & T. Paek (Eds.), *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue* (pp. 182-189).

Bolly, C., & Degand, L. (2009). Quelle(s) fonction(s) pour "donc" en français oral? Du connecteur conséquentiel au marqueur de structuration du discours. *Lingvisticae Investigationes, 32*(1), 1-32.

Bolly, C., Crible, L., Degand, L., & Uygur-Distexhe, D. (2017). 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* (pp. 71-97). Amsterdam: John Benjamins.

Bortfeld, H., Leon, S., Bloom, J., Schober, M., & Brennan, S. (2001). Disfluency rates in conversation: Effects of age, relationship, topic, role and gender. *Language and Speech, 44*(2), 123-147.

Boula de Mareüil, P., Adda, G., Adda-Decker, M., Barras, C., Habert, B., & Paroubek, P. (2013). Une étude quantitative des marqueurs discursifs, disfluences et chevauchements de parole dans des interviews politiques. *TIPA Travaux Interdisciplinaires sur la Parole et le Langage, 29*.

Bouraoui, J.-L., & Vigouroux, N. (2006). Étude de dysfluences dans un corpus linguistiquement contraint. In *Proceedings of the Journée d'Etudes sur la Parole (JEP 2006)* (pp. 429-432).

Brognaux, S., Roekhaut, S., Drugman, T., & Beaufort, R. (2012). *Train&Align*: A new online tool for automatic phonetic alignment. In *Proceedings of IEEE Spoken Language Technology Workshop (SLT)* (pp. 416-421).

Candéa, M. (2000). *Contribution à l'Etude des Pauses Silencieuses et des Phénomènes Dits "d'Hésitation" en Français Oral Spontané* (Unpublished doctoral dissertation). Université Paris III, Paris.

Crible, L. (2014). *Identifying and Describing Discourse Markers in Spoken Corpora. Annotation Protocol v.8* (Technical report). Louvain-la-Neuve, Université catholique de Louvain.

Crible, L. (2017). Towards an operational category of discourse markers: A definition and its model. In A. Sansó & C. Fedriani (Eds.), *Discourse Markers, Pragmatic Markers and Modal Particles: New Perspectives* (pp. 99-124). Amsterdam: John Benjamins.

Crible, L., & Degand, L. (in press). Reliability vs. granularity in discourse annotation: What is the trade-off? *Corpus Linguistics and Linguistic Theory*.

Crible, L., Degand, L., & Gilquin, G. (2017). The clustering of discourse markers and filled pauses: A corpus-based French-English study of (dis)fluency. *Languages in Contrast 17*(1), 69-95.

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). Louvain-la-Neuve & Namur, Université catholique de Louvain & Université de Namur.

Degand, L., Martin, L., & Simon, A.-C. (2014). *LOCAS-F: Un corpus oral multigenres annoté*. Paper presented at the Congrès Mondial de Linguistique Française, Berlin, Germany.

Demirşahin, I., & Zeyrek, D. (2014). Annotating discourse connectives in spoken Turkish. In L. Levin & M. Stede (Eds.), *LAW VIII – The 8th Linguistic Annotation Workshop* (pp. 105-109).

Denke, A. (2009). *Nativelike Performance. Pragmatic Markers, Repair and Repetition in Native and Non-native English Speech*. Saarbrücken: Verlag Dr. Müller.

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), 113-129.

Dupont, M. (2015). Word order in English and French: The position of English and French adverbial connectors of contrast. *English Text Construction*, *8*(1), 88-124.

Ejzenberg, R. (2000). The juggling act of oral fluency: A psycho-sociolinguistic metaphor. In H. Riggenbach (Ed.), *Perspectives on Fluency* (pp. 288-313). Ann Arbor: The University of Michigan Press.

Eklund, R. (2004). *Disfluency in Swedish Human-human and Human-machine Travel Booking Dialogues* (Unpublished doctoral dissertation). Linköpings Universitet, Linköping.

Eklund, R., & Shriberg, E. (1998). Crosslinguistic disfluency modeling: A comparative analysis of Swedish and American English human-human and human-machine dialogs. In R. H. Mannell & J. Robert-Ribes (Eds.), *Proceedings of the 5th International Conference on Spoken Language Processing* (pp. 2627-2630). Canberra: Australian Speech Science and Technicology Association, Incorporated (ASSTA).

Gilquin, G. (2006). The place of prototypicality in corpus linguistics. Causation in the hot seat. In S. Gries & A. Stefanowitsch (Eds.), *Corpora in Cognitive Linguistics: Corpus-based Approaches to Syntax and Lexis* (pp. 159-191). Berlin: Mouton de Gruyter.

Gilquin, G., & Gries, S. (2009). Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory*, *5*(1), 1-26.

Goldman, J.-P., Prsir, T., & Auchlin, A. (2014). C-PhonoGenre: A 7-hour corpus of 7 speaking styles in French: Relations between situational features and prosodic properties. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (Eds.), *Proceedings of the 9th Language Resources and Evaluation Conference (LREC'14)* (pp. 302-305). Paris, European Language Resources Association (ELRA).

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

Götz, S. (2013). *Fluency in Native and Nonnative English Speech*. Amsterdam: John Benjamins.

Grosjean, F., & Deschamps, A. (1975). Analyse contrastive des variables temporelles de l'anglais et du français: Vitesse de parole et variables composantes, phénomènes d'hésitation. *Phonetica, 31*(3-4), 144-184.

Grosman, I. (2016). 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* (pp. 147-175). Amsterdam: John Benjamins.

Hansen, M.-B. M. (2006). A dynamic polysemy approach to the lexical semantics of discourse markers (with an exemplary analysis of French *toujours*). In K. Fischer (Ed.), *Approaches to Discourse Particles* (pp. 21-41). Amsterdam: Elsevier.

Hasselgren, A. (2002). Learner corpora and language testing: Small words as markers of learner fluency. In S. Granger, J. Hung & S. Petch-Tyson (Eds.), *Computer-Learner Corpora,*

Second Language Acquisition, and Foreign Language Teaching (pp. 143-173). Philadelphia, PA: John Benjamins.

Kemmer, S., & Barlow, M. (2000). Introduction: A usage-based conception of language. In M. Barlow & S. Kemmer (Eds.), *Usage Based Models of Language* (pp. vii-xxviii). Stanford: CSLI.

Kohn, K. (2012). Pedagogic corpora for content and language integrated learning. Insights from the BACKBONE project. *The Eurocall Review*, *20*(2), 1-22.

Kunz, K., & Lapshinova-Koltunski, E. (2015). Cross-linguistic analysis of discourse variation across registers. *Nordic Journal of English Studies*, *14*(1), 258-288.

Lacheret, A., Kahane, S., & Pietrandrea, P. (Eds.) (2014). *Rhapsodie: A Prosodic and Syntactic Treebank for Spoken French*. Amsterdam: John Benjamins.

Lopes, A., Martins de Matos, D., Cabarrão, V., Ribeiro, R., Moniz, H., Trancoso, I., & Mata, A. I. (2015). Towards using machine translation techniques to induce multilingual lexica of discourse markers. *CoRR*, 1-6.

Meteer, M. Taylor, A., MacIntyre, R., & Iver, R. (1995). *Disfluency Annotation Stylebook for the Switchboard Corpus* (Technical report). Linguistic Data Consortium. Philadelphia, PA, University of Pennsylvania.

Müller, S. (2005). *Discourse Markers in Native and Non-native English Discourse*. Amsterdam: John Benjamins.

Nelson, G., Wallis, S., & Aarts, B. (2002). *Exploring Natural Language: Working with the British Component of the International Corpus of English*. Amsterdam: John Benjamins.

Palisse, S. (1997). *"Artisans", "Assureurs", Conversations Téléphoniques en Entreprise*. Retrieved from http://clapi-univ.lyon2.fr (last accessed March 2014).

Pallaud, B., Rauzy, S., & Blâche, P. (2013). Auto-interruptions et disfluences en français parlé dans quatre corpus du CID. *TIPA Travaux Interdisciplinaires sur la Parole et le Langage*, *29*, 2-19.

Pawley, A., & Syder, F. (2000). The one-clause-at-a-time hypothesis. In H. Riggebbach (Ed.), *Perspectives on Fluency* (pp. 163-199). Ann Arbor: The University of Michigan Press.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., & Webber, B. (2008). The Penn Discourse TreeBank 2.0. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis & D. Tapias (Eds.), *Proceedings of the 6th Language Resources and Evaluation Conference (LREC'08)* (pp. 2961-2968). Paris, European Language Resources Association (ELRA).

Roekhaut, S., Brognaux, S., Beaufort, R., & Dutoit, T. (2014). *eLite-HTS: Un outil TAL pour la génération de synthèse HMM en français*. Paper presented at the Journées d'Etude de la Parole (JEP), Le Mans, France.

Rühlemann, C., & O'Donnell, M. (2012). Introducing a corpus of conversational stories. Construction and annotation of the *Narrative Corpus*. *Corpus Linguistics and Linguistic Theory*, *8*(2), 313-350.

Schegloff, E., Jefferson, G., & Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language*, *53*(2), 361-382.

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

Schmid, H. (1997). Probabilistic part-of-speech tagging using decision trees. In D. Jones & H. Somers (Eds.), *New Methods in Language Processing* (pp. 154-164). London: UCL Press.

Schmid, H.-J. (2010). Does frequency in text instantiate entrenchment in the cognitive system. In D. Glynn & K. Fischer (Eds.), *Quantitative Methods in Cognitive Semantics: Corpus-Driven Approaches* (pp. 101-133). Berlin: Mouton de Gruyter.

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

Schourup, L. (1999). Discourse markers. *Lingua*, *107*, 227-265.

Shriberg, E. (1994). *Preliminaries to a Theory of Speech Disfluencies* (Unpublished doctoral dissertation). University of California, Berkeley, CA.

Simon, A.-C., Auchlin, A., Avanzi, M., & Goldman, J.-Ph. (2010). Les phonostyles. Une description prosodique des styles de parole en français. In M. Abecassis & G. Ledegen (Eds.), *Les Voix des Français. En Parlant, en Ecrivant, vol. 2* (pp. 71-88). Bern: Peter Lang.

Strassel, S. (2003). *Simple Metadata Annotation Specification v.5* (Technical report). Linguistic Data Consortium. Philadelphia, PA, University of Pennsylvania.

Tonelli, S., Riccardi, G., Prasad, R., & Joshi, A. (2010). Annotation of discourse relations for conversational spoken dialogs. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner & D. Tapias (Eds.), *Proceedings of the 7th Language Resources and Evaluation Conference (LREC'10)* (pp. 2084-2090). Paris, European Language Resources Association (ELRA).

Willems, D., & Demol, A. (2006). *Vraiment* and *really* in contrast: When truth and reality meet. In K. Aijmer & A.-M. Simon-Vandenbergen (Eds.), *Pragmatic Markers in Contrast* (pp. 215-235). Amsterdam: Elsevier.

Zikánová, Š., Hajičová, E., Hladká, B., Jínová, P., Mírovský, J., Nedoluzhko, A., Poláková, L., Rysová, K., Rysová, M., & Václ J. (2015). *Discourse and Coherence. From the Sentence Structure to Relations in Text.* Prague: Institute of Formal and Applied Linguistics.

Zufferey, S., & Cartoni, B. (2012). English and French causal connectives in contrast. *Languages in Contrast, 12*(2), 232-250.

Zufferey, S., & Degand, L. (in press). Annotating the meaning of discourse connectives in multilingual corpora. *Corpus Linguistics and Linguistic Theory*.

Appendices

Appendix 1. Discourse markers annotation scheme

Information	Variable	Definition	Values
Syntax	POS	source grammatical class of the (head of the) DM	CC, RB, VP, SC, WP, JJ, NN, PP, UH
	Position: macro	position referring to the dependency structure	pre-field – PRE left (integrated) – LEFT middle field – MID right (integrated) – RIGHT post-field – POST independent – IND
	Position: micro	position referring to the minimal clause	initial – ini medial – med final – fin independent – ind
	Position: turn	position referring to the turn-of-speech	turn-initial – TI turn-medial – TM turn-final – TF whole turn – TT
	Co-occurrence	whether the DM co-occurs with another DM	yes-left – Yleft yes-right – Yright yes-both – Ylr no – NO
Pragmatics	Type	whether the DM expresses a relation or not	relational – RDM non-relational – NDRM both – B
	Domain	component of language structure affected by the marker	ideational – IDE rhetorical – RHE sequential – SEQ interpersonal – INT
	Function	function in context	see Table 4

Appendix 2. (Dis)fluency annotation scheme (Crible et al. 2016)

Tags	Fluencemes	Examples
UP	unfilled pause (sec.)	(0.380)
FP	filled pause	*uhm, uh, euh*
DM	discourse marker	*so, because, well, I mean*
ET	editing term	*oops, what is it?*
FS	false-start	"places are funny on (1.060) well they don't…"
TR	truncation	"tran/ uhm (0.700) transplant"
RI	identical repetition	"they go (0.630) eh they go"
RM	modified repetition	"a lot of time a lot of money"
SP	propositional substitution	"Asian speakers well no Asian people living in the UK"
SM	morphological substitution	"but there is there are"
	Diacritics	
AR	misarticulation	"to do resiv/ residential conveyancing"
WI	embedded fluenceme	"she and she"
OR	change of order	"normally would take you would normally take you"
	Related elements	
IL	lexical insertion	"I deal with disputes, so civil disputes"
IP	parenthetical insertion	"and the rainy (0.250) well touch wood the rainy"
DE	deletion	Mary didn't want to come Mary didn't come

Author's address

Ludivine Crible

Centre Valibel – Discourse & Variation

Université catholique de Louvain

1 place Blaise Pascal

Louvain-la-Neuve 1348

Belgium

ludivine.crible@uclouvain.be