The clustering of discourse markers and filled pauses: a corpus-based French-English study of (dis)fluency[1]

Ludivine Crible, Liesbeth Degand, Gaëtanelle Gilquin

Université catholique de Louvain

This article presents a corpus-based contrastive study of (dis)fluency in French and English, focusing on the clustering of discourse markers (DMs) and filled pauses (FPs) across various spoken registers. Starting from the hypothesis that markers of (dis)fluency, or 'fluencemes', occur more frequently in sequences than in isolation, and that their contribution to the relative fluency of discourse can only be assessed by taking into account the contextual distribution of these sequences, this study uncovers the specific contextual conditions that trigger the clustering of fluencemes in the two languages. First, the contexts of appearance of DMs and FPs are described separately, both in English and French, focusing on their distribution, position and co-occurrence patterns. Then, the combination of DMs and FPs in sequences and their different configurations (DM+FP, FP+DM, etc.) are investigated. Overall, it appears that FPs function differently depending on whether they are clustered with DMs or not, and this difference consists in either maintaining or erasing inter- and intra-linguistic contrasts.

Keywords: fluency; filled pauses; discourse markers; comparable corpus; English/French

1. Introduction

Componential approaches to fluency (Hieke, 1985) have shown that different features can contribute to the fluency (or disfluency) of discourse, among which speech rate, (filled and unfilled) pauses or discourse markers (henceforth DMs). These 'fluencemes' (the term is taken from Götz, 2013) seem to be present in all languages. However, the relatively few studies that have performed a contrastive analysis of fluencemes reveal crosslinguistic differences. Thus, it is well known that filled pauses (henceforth FPs) are language-specific, from English *uh* or *uhm* and French *euh* to Spanish *pues* or Japanese *eeto*, for instance (see Clark and Fox Tree, 2002: 92). Their use and selection may also differ significantly from one language to the other, as demonstrated by Zhao and Jurafsky (2005) for English and Mandarin. Similarly, discourse markers appear to have several equivalents crosslinguistically, which points towards language-specific functions (e.g. Aijmer and Simon-Vandenbergen, 2006).

[1] This research benefits from the support of the ARC-project "Fluency and disfluency markers. A multimodal contrastive perspective" granted by the Fédération Wallonie-Bruxelles (grant nr.12/17-044).

Discourse markers are generally viewed as contributing to speakers' fluency (Hasselgren, 2002; Müller, 2005; Götz, 2013), although sometimes they are stylistically stigmatised as "a sign of dysfluency and carelessness" (Brinton, 1996: 33) by authors who associate them with "unskilful speakers" or "powerlessness" (O'Donnell and Todd, 1980: 67; Ragan, 1983: 166). Similarly, filled pauses have been shown to positively help speech production and processing (O'Connell and Kowal, 2005), while also seen as encoding hesitations and difficulties (e.g. Mahl, 1987; Levelt, 1989). Studying combinations of DMs and FPs is intended to shed light on the relative contribution of these fluencemes to (dis)fluency, both clustered and isolated, following the claim that (dis)fluency is a multifaceted phenomenon which should be investigated by means of an integrated approach (Götz, 2013).

In this article, we propose a French-English contrastive study of the clustering of discourse markers and filled pauses in naturally-occurring speech to test the hypothesis that fluencemes usually do not occur in isolation, but tend to combine (cf. Stenström, 1990: 222; Aijmer, 1997: 27). Background and key notions are presented in Section 2. Using comparable corpora representing a variety of spoken situations (e.g. conversation, news broadcast) in French and British English (Section 3), we start by profiling the linguistic environment of DMs and FPs, i.e. describing their position and distribution in the various subcorpora as well as their co-occurrence with unfilled pauses (Section 4). We then examine the specific patterns characterizing the combinations of DMs and FPs to uncover any quantitative and/or qualitative differences as compared to their individual contexts of appearance, and to check whether their behavior differs crosslinguistically (Section 5). Tentative interpretations of their role as either fluency or disfluency markers will be drawn from the synthesis of our corpus-based observations.

2. An integrated approach to native (dis)fluency

2.1. A multifaceted phenomenon

Unlike our daily language-user experience of fluency as a global, holistic impression of efficiency and/or naturalness, most corpus-based studies examine individual fluencemes (or combinations thereof), thus adopting a componential approach to the phenomenon. The research presented here is in line with this latter approach. The rationale for decomposing a holistic impression into a typology of temporal and linguistic measures is the following: fluency and disfluency are two sides of the same coin, and the "diagnosis" can only be drawn relatively to the combination and distribution of different types of discursive devices with regard to situational expectations and prototypicality. This usage-based framework vouches for a definition of (dis)fluency as (i) sequential (fluency is the result of specific patterns of combination or "sequences"), (ii) situational (these patterns are confronted with social and contextual norms) and (iii) ambivalent (a particular pattern can be either fluent or disfluent depending on its distribution in the micro and macro-context). This ambivalence led to diverging accounts in the literature on one particular fluenceme, namely filled pauses (Maclay and Osgood, 1959; Levelt, 1989). Clark and Fox Tree (2002) summarized two types of interpretations, either "filler-as-symptom" (involuntary, meaningless effect of a

problem) or "filler-as-signal" (functional, interpretable cue for the listener). We believe that this duality should instead be considered on a scale of (dis)fluency, and extended to all fluencemes, since certain ambivalent functions (e.g. hesitating vs. holding the floor) seem to apply to other fluencemes too (FPs, but also DMs, repetitions, word-final lengthening, etc.). Genre variation (Broen and Siegel, 1972; Merlo and Mansur, 2004), speaker profiles (Bortfeld *et al.*, 2001; Tottie 2011) and linguistic surface features (position in and mean length of utterance, e.g. Oviatt, 1995; Auer, 2005) are often found to be relevant factors in the distribution of fluencemes, which in turn provides insight into the cognitive-functional preference for certain forms and structures as either fluent or disfluent.

2.2. Contrastive fluency: discourse markers and filled pauses

Numerous authors have examined the lexical, syntactic and functional contrasts of DMs in many languages (e.g. Cuenca, 2003; Aijmer and Simon-Vandenbergen, 2006; Fagard and Degand, 2010; Hasselgård, 2014). These corpus-based studies reveal formal, functional and distributional differences within and between languages. For instance, Bazzanella *et al.* (2007) have shown that the Italian-French cognates *allora* / *alors* share the same meanings but that these meanings present a different degree of prototypicality, i.e. more or less central vs. peripheral in their category, with the temporal value being more prominent for the Italian DM. A more general finding regarding the distribution of the whole DM category can be found in González (2005), who observed three times more DMs in Catalan narrative speech than in English. The present article aims at following this categorical, paradigmatic line of research in French and English, where it is still missing to date, as opposed to the bulk of contrastive case studies (e.g. Willems and Demol, 2006; Defour *et al.*, 2010).

On the other hand, corpus-based crosslinguistic investigations of filled pauses are not as common. Eklund and Shriberg (1998) show that English and Swedish FPs are mostly sentence-initial, but that compounding languages like Swedish allow word-internal FPs as well. Zhao and Jurasfky (2005) work on Mandarin and English and compare different types of FPs (*uh/uhm* vs. demonstratives) in various syntactic contexts. Vasilescu *et al.* (2007) investigate the link between vocalic hesitation and the phonemic systems of English, French and Spanish. As far as we are aware, these references are the only ones available. In general, it appears that these lexicalized vocalizations are not universal in form and function either, although generally built around central vowels and usually followed by a nasal consonant. Clark and Fox Tree (2002) provide a review of these expressions in languages where they have been studied; we can see in Table 1 that Spanish and Japanese differ the most from other languages with their use of demonstratives as FPs (*este, ano, sonoo*). Functionally, most studies focus on intra-linguistic differences such as the preference for *uhm* in macro-planning contexts as opposed to *uh* in local planning uses (Shriberg, 1994).

The present article builds on a previous paper by Degand and Gilquin (2013) which is, to the best of our knowledge, the first and only contrastive study of the DM+FP pattern in French and English – or any other pair of languages, for that matter. Our analysis seeks to confirm their results in a more varied corpus and to supplement the profiles of DMs and FPs with a systematic comparison of the fluencemes in individual vs. combined contexts.

Table 1. Intra- and inter-linguistic variation of FPs (adapted from Clark and Fox Tree, 2002: 92).

Language	**Fillers**
German	äh, ähm
Dutch	uh, um
Swedish	eh, a, mm
Norwegian	e, e=, eh, m, hm, aj
French	euh, em, n
Hebrew	eh, e-h, em, ah
Spanish	eh, em, este, pues
Japanese	eeto, ano, konoo, sonoo

2.3. Where are we (dis)fluent?

Hypothesizing that fluencemes occur more frequently in sequences than individually, we claim that their relative contribution to the (dis)fluency of discourse comes from their combination. Crosslinguistically, Grosjean and Deschamps (1975) show that FPs are often combined with silent pauses (68.29% of occurrences in English vs. 47.26% in French). From a computational point of view, Brennan and Schober (2001) report that combinations of cues are better signals of an interruption of fluency than one single fluenceme. Working on the English modal particle *I think*, Aijmer (1997) identifies two privileged slots where "modal and interpersonal elements cooccur with hesitation noises, word-search, repetition and self-corrections", namely beginning and end of the speaker turn, respectively as signals of planning or "to assure the hearer that the relationship will continue" (1997: 27).

Studies on FPs dealing with their surrounding fluencemes often consider the presence of unfilled pauses as a relevant factor (Swerts, 1998; Clark and Fox Tree, 2002), while Degand and Gilquin (2013) additionally note the presence of co-occurring DMs in the immediate periphery of the FP. They found that, though most frequent in the right periphery in both languages, FPs tend to occur at the left periphery of the DM in higher proportions in English than in French. We will therefore focus on the different configurations of DM+FP clusters (with and without unfilled pauses) in our search for contrastive differences within and between languages. Special attention will be given to the position of the FP with respect to the DM in the sequence, as well as the position of the whole sequence in the utterance.

Given the "stalling" effect that FPs can have in online spoken production, we expect them to frequently occur at syntactic boundaries, where the cognitive load of speech planning is high. This hypothesis draws on experimental (e.g. Roberts and Kirsner, 2000) and corpus-based (e.g. Schneider, 2014) evidence that fluencemes tend to cluster at major boundaries, or after the first element in the sentence. Clark and Fox Tree (2002) and Shriberg (1994) both claim that English *uhm* occurs more often at sentence boundaries than within utterances, a position which is more typical of *uh*. This finding will be tested in our English corpus, where it will also be compared with the position of DM+FP clusters and with the behavior of the French FP *euh*. We will thus attempt to identify intra- and inter-linguistic patterns of DMs and FPs and the factors relevant to their variation.

3. *DisFrEn*: an enriched comparable dataset

3.1. Data

The corpus study made use of *DisFrEn*, a comparable French-English dataset assembling spoken texts from various existing resources, primarily ICE-GB, the British component of the International Corpus of English (Nelson *et al.*, 2002) for the English data, and VALIBEL (Dister *et al.*, 2009) for French. These corpora have in common that they gather transcripts from a variety of interactional settings, although not in the same quantity. ICE-GB and VALIBEL are thus not directly comparable as such, but the dataset resulting from their sampling was built with this purpose in mind.

DisFrEn comprises six contextual settings (phone calls, conversations, classroom lessons, sports commentaries, news broadcasts, political speeches) balanced across the two languages, for a total of 10 hours of speech and over 100,000 words. The distribution of words per situation and language is represented in Table 2.

Table 2. Word count per situation and language in *DisFrEn*.

	English	**French**	**Total**
phone calls	9747	6783	16530
conversation	17479	17432	34911
classroom lessons	9425	3723	13148
sports commentaries	8237	6279	14516
news broadcasts	7046	6788	13834
political speeches	8650	7824	16474
Total	**60584**	**48829**	**109413**

Table 2 shows that balance between languages for each situational subcorpus (except classroom lessons due to missing data) has been prioritized over balance between situations: conversations are deliberately more represented than other situations since they are the most natural and frequent type of spoken language. All texts were sound-aligned with eLite (Roekhaut *et al.*, 2014) and Train&Align (Brognaux *et al.*, 2012), thus making the audio track available throughout the annotation process.

3.2. Annotation of fluencemes

Our analysis proceeded in five steps: identification of DMs; annotation of DMs; annotation of fluencemes; post-treatment; extraction of all FPs. First, all DMs in the transcription were manually identified, using the EXMARaLDA annotation tool (Schmidt and Wörner, 2009) and following a broad, function-based definition: candidate items must be syntactically optional, have a procedural

meaning, and function either as connectors between utterances, discourse-structuring devices or tools managing the speaker-hearer relationship (see Crible and Zufferey, 2015; Crible, in press). As a result, this definition includes both so-called 'connectives' such as *and* or *although*, and non-relational, speech-specific DMs such as *well* or *I mean*, provided they meet the criteria stated above. Since the identification and annotation was entirely manual and bottom-up, propositional, non-DM uses of these expressions were disambiguated in context and excluded from the selection (e.g. an intra-clausal use of *and* as in *Cats and dogs don't get along*). These items were then described through a number of functional and syntactic variables, following an operational corpus-based protocol detailed in Crible (2014, in press) including, among others, a taxonomy of thirty functional values (e.g. cause, reformulation, monitoring) and a three-fold positioning system based on Bolly *et al*. (in press):

- position in the clause: initial, medial, final, independent;

- position in the dependency unit (with respect to the governing verb): pre-field (non-integrated, left periphery), initial field (integrated, left periphery), middle field (within the verb clause), end-field (integrated, right), post-field (non-integrated, right), independent;

- position in the turn: turn-initial (very first word after a change of speaker), turn-medial (anywhere but initial and final), turn-final (very last word before a change of speaker), whole turn.

In a third step, all fluencemes in the direct context of these DMs were identified and annotated following a multimodal and multilingual protocol (Crible *et al.*, 2016; see Dumont, 2014; Grosman, 2016 or Notarrigo *et al.*, 2016 for applications) which covers the following phenomena: unfilled and filled pauses, DMs, editing terms, false-starts, truncations, modified and identical repetitions, morphological and propositional substitutions, deletions, lexical and parenthetical insertions. Other secondary phenomena were also accounted for, such as misarticulation, nested fluencemes or re-ordering. It should be noted that final-word lengthening was not considered as a special case of filled pauses, since this information was not systematically or reliably provided in our data. Hence we included among FPs only the lexicalized vocalizations which had been identified as such by the transcribers in the original corpus.

After extracting the annotated DMs and FPs, several post-treatment categorizations were added to the dataset in order to offer complementary filters into the annotations. One that concerns the present study is the coding of all possible combinations between DMs, FPs and unfilled pauses (UPs), as shown in Table 3. Of these configurations, only those containing at least a DM and a FP were extracted for this analysis (see tags in bold).

Table 3. Different configurations for the clusters of discourse markers and pauses.

Tags	Content	Configuration	Example
N/A	No pause	DM	well
UPL	Unfilled Left	UP+DM	(0.530) well
UPR	Unfilled Right	DM+UP	well (0.238)
FPL	Filled Left	FP+DM	uh well
FPR	Filled Right	DM+FP	well uhm
UPB	Unfilled Both sides	UP+DM+UP	(0.530) well (0.238)
FPB	Filled Both sides	FP+DM+FP	uh well uhm
UFL	Unfilled, Filled Left	UP+FP+DM	(0.530) uh well
UFR	Unfilled, Filled Right	DM+UP+FP	well (0.275) uh
FUL	Filled, Unfilled Left	FP+UP+DM	uhm (0.400) well
FUR	Filled, Unfilled Right	DM+FP+UP	well uhm (0.400)
UDF	Unfilled, DM, Filled	UP+DM+FP	(0.250) well uhm
FDU	Filled, DM, Unfilled	FP+DM+UP	uh well (0.387)
MIX	Any other cluster	e.g. FP+DM+UP+FP	uh well (0.387) uhm

These clusters of DMs and FPs can occur either by themselves (1), with other fluencemes such as false-starts (2) or inside compound, two-part structures such as a repetition (3). Such distinctions will not be studied any further here.

(1) where we were staying *because uhm* it was the local one [EN-conv-07][2]

(2) so we might *for example uhm* there's a technique [EN-clas-02]

(3) I'd dearly love to *uh you know* to be spending time [EN-phon-01]

In a last step of the analysis, we automatically extracted all occurrences of FPs in the corpus, whether in the context of a DM or not. Following the literature and the transcription conventions used in our corpus, we searched for the following three forms: French *euh*, English *uh* and *uhm* (queries for *um*, *hum* and *mm* were either unsuccessful or returned other elements such as backchanneling devices). From the 1,570 automatically identified FPs, a random sample of 200 items in each language was extracted and manually coded for:

- position in the clause: either "between" (at a boundary) or "within" (inside the clause); the "between" position includes quasi-initial, initial and final positions in the clause, as well as cases of interruptions, either internal or external, since they signal an end, although not always a voluntary one;

- presence of an unfilled pause in the direct periphery of the FP, using the pauses identified in the transcriptions (150ms and longer);

- presence of a DM (and potentially other fluencemes) in the direct context, i.e. whether the FP is part of a sequence or not.

[2] ID codes in *DisFrEn* show the language, speaking task and text number the example was extracted from.

In the remainder of this article, all cross-tabulations of DM-based and FP-based annotations will be provided in both absolute and normalized frequencies (relatively to the number of words per subcorpus[3]) to remain comparable, except for the sample analysis which will make no use of situational metadata. Further information about the annotated features of DMs can be found in Crible (in press).

4. Crosslinguistic profiles of individual fluencemes

In this first results section, isolated occurrences of DMs (Section 4.1) and FPs (Section 4.2) are analysed with a systematic description of their frequency (Sections 4.1.1 and 4.2.1), position (Sections 4.1.2 and 4.2.2) and co-occurrence patterns focusing on unfilled pauses (Sections 4.1.3 and 4.2.2). Apart from frequency, more fine-grained analyses of FPs (position, co-occurrence) are only computed for the 400-item sample. An additional section mentions further distinctions between the FP forms *euh*, *uh* and *uhm* (Section 4.2.3) to be compared with the clustered contexts (Section 5.3). Finally, note that FPs are considered "isolated" when they do not co-occur with DMs, thus potentially including cases of co-occurrence with other fluencemes such as unfilled pauses, repetitions or false-starts.

4.1. Isolated DMs

4.1.1. *Frequency*

Like any contrastive research, we will first address the long-lasting claim that English and French have different distributions. Starting with DMs, this question has mostly been tackled intuitively in the literature, with no corpus data to support it, which has led to contradictory conclusions: Guillemin-Flescher (1981) argues that coordination (as opposed to juxtaposition) is more frequent in English than in French, while Vinay and Darbelnet (1995) claim the contrary.[4] In our data, normalized frequencies are significantly greater for French as we can see in Table 4 (21 DMs per 1000 words vs. 16/1000 in English, LL = 40.24, $p < 0.001$).

[3] Mean values per register are preferred over speaker-dependent frequencies since this research draws on strong variationist hypotheses, while individual or sociolinguistic tendencies are not the focus of our project. Also, the different subcorpora are not perfectly balanced (see Section 3.1).

[4] Coordinating DMs (e.g. inter-clausal conjunctions such as *and* or *but*) are, in our view, only one type of DMs. The reference to these studies merely illustrates the types of analysis that were common in the pre-corpus era.

Table 4. Absolute and normalized frequency for isolated DMs.

	English	1/1000	French	1/1000	Total	1/1000
phone call	251	26	233	34	484	29
conversation	397	23	583	33	980	28
classroom	130	14	37	10	167	13
sports	138	17	112	18	250	17
news	26	4	51	8	77	6
political	26	3	20	3	46	3
Total	**968**	**16**	**1036**	**21**	**2004**	**18**

This table also shows some situational preferences, in both languages, for spontaneous contexts (phone calls and conversations), followed by professional but unscripted situations (classroom lessons and sports commentaries), leaving very low frequencies of isolated DMs in very formal, scripted situations (news broadcasts and political speeches).

Table 5. Top five most frequent DM lexemes in English and French.

English	French
and (194)	*et* 'and' (191)
well (139)	*mais* 'but' (157)
but (127)	*quoi* 'you know' (93)
so (86)	*hein* 'eh'[5] (72)
if (44)	*ben* 'well' (60)

Zooming in on the most frequent DMs, Table 5 shows a few crosslinguistic differences: while English and French share some generic, semantically close conjunctions in their top five (*and*, *but* respectively ranked 1st and 3rd; *et*, *mais* respectively 1st and 2nd), French isolated DMs tend to be more interactive and speech-specific (*quoi*, *hein*, *ben*, respectively 3rd, 4th and 5th), whereas in English only *well* matches this definition (ranked 2nd). Although typical conjunctions like *and* are highly polysemous, including more interactive meanings such as turn-taking, it is striking that some of the most frequent DMs in French have a much narrower functional range mostly represented by one or two meanings-in-context (as opposed to generic, multifunctional conjunctions), and are strongly associated with informal conversation.

4.1.2. *Position of DMs*

Similar proportions between languages are observed for positional variables in isolated contexts. Distributions of DMs in their clause and in their turn-of-speech are reported in Table 6. A clause is defined as a minimal unit with a predicate (including subclauses). Turns are defined as continuous stretches of talk produced by the same speaker. Turn-initial and turn-final positions may include non-lexical

[5] The closest English equivalent to French *hein* is probably a tag question (e.g. *isn't it?*), however tag questions were excluded from our selection of DMs due to their high formal/syntactic variation (see Crible, in press).

items before or after the DM, e.g. a filled pause or other para-verbal behavior. Turn boundaries have been identified following the transcription conventions in the source corpora.

Table 6. Position of isolated DMs in the clause and in the turn.

	Position in the clause			**Position in the turn**		
	English	**French**	**Total**	**English**	**French**	**Total**
initial	803	716	1519	180	229	409
medial	61	70	131	716	643	1359
final	86	209	295	61	122	183
independent	14	45	59	11	42	53
Total	968	1036	2004	968	1036	2004

We see that DMs are mostly initial in their host clause (1,519 out of the total 2,004), followed by the final position mostly represented by French DMs (209 *vs*. 86 in English), then medial (131) and independent position (59). The higher frequency of final position in French (LL = 82.84, $p < 0.001$) is very likely to be the direct consequence of the previous results on typically interactive DMs such as *quoi* and *hein* which are usually clause-final, as in (4).

(4) il arrive à gagner *quoi* c'est un (0.253) pourtant c'est un [FR-conv-02]

"he manages to win *quoi* ('you know') it's a (0.253) and yet it's a"

With respect to the position in the turn-of-speech, again the distribution of isolated DMs is similar crosslinguistically. Turn-medial occurrences are unsurprisingly the most frequent with 68% of all DMs (716 in English, 643 in French). Another 20% of isolated DMs are used turn-initially (180 in English, 229 in French). Turn-final DMs (9%) are twice as frequent in French as in English (11.8% vs. 6.3%), a result which is again mainly due to the use of *hein* and *quoi* in this position.

4.1.3. *Co-occurrence with other DMs and unfilled pauses*

In *DisFrEn*, the isolated contexts described above are not the most frequent, which confirms their tendency to combine. Their co-occurrence with FPs will be discussed below (Section 5), but they cluster with other fluencemes as well. Starting with sequences of several DMs, we counted 412 DMs directly co-occurring with one another, forming a total of 197 clusters. DM-only sequences contain up to three DM tokens; longer sequences necessarily include additional fluencemes (e.g. unfilled pauses) in our data. There seems to be a significant difference between French and English sequences of DMs: 9.95% of all French DMs occur in DM-only sequences, against 5.11% of all the English ones (LL = 79.87, $p < 0.001$). Again, this result could be related to our previous finding on the frequency of *quoi*, which is often found in co-occurring patterns, as in (5) (see Crible, 2015 for a contrastive study of co-occurring DMs).

(5) c'est leur métier *quoi mais je veux dire* ils c'est une [FR-conv-05]

"it's their job *quoi* ('you know') *mais* ('but') *je veux dire* ('I mean') they"

By contrast, the proportion of DMs co-occurring with unfilled pauses (and nothing else) is slightly higher in English with 36.54% of all English DMs against 30.65% in French, although this result is not significant in terms of frequency in each corpus (LL = 1.59, $p > 0.05$). These clusters do not affect the distribution of positional variables described in the previous section (i.e. mostly initial in both languages, with a higher proportion of final position in French). This result only considers sequences strictly containing DMs and unfilled pauses, leaving out of this analysis patterns including more types of fluencemes, given the diversity of configurations in the data.

4.2. Isolated FPs

4.2.1. *Frequency*

Similarly to the results for DMs, FPs display a higher normalized frequency in French than in English, as shown in Table 7 (respectively, 12 vs. 9 items per 1,000 words, LL = 16.93, $p < 0.001$).

Table 7. Frequency of isolated[6] filled pauses in *DisFrEn*: absolute and normalized frequencies (per 1,000 words).

Situation	English	1/1000	French	1/1000	Total	1/1000
phone	191	19	107	16	298	18
conversation	211	12	225	13	436	12
classroom	116	12	63	17	179	14
sports	24	3	77	12	101	7
news	16	2	76	11	92	7
political	5	1	31	4	36	2
Total	**563**	**9**	**579**	**12**	**1142**	**10**

Register analysis also shows that FPs are scarce in very formal situations in English (news broadcasts, political speeches), whereas they do occur in French, albeit to a lesser extent than in more casual situations. This could suggest that FPs are not as stigmatized in French as in English, since French speakers in professional, scripted settings do not exclude them from their repertoires. It remains to be investigated whether filled pauses are part of similar discourse-functional strategies (e.g. planning time, punctuation, lexical salience) in different situations, but since this question is probably related to issues of professional or individual speaking styles, it would require closer, more qualitative analysis. Examples (6) and (7) respectively illustrate the punctuating and word-search uses of *euh* in news broadcasts and sports commentaries, where the normalized frequencies are very similar despite contextual differences expected from the spontaneous, "live" nature of sports commentaries.

(6) les marchés ont poursuivi leur redressement *euh* ce matin [FR-news-04]

"markets continued their recovery *euh* this morning"

[6] Note that these figures may include cases of FPs co-occurring with fluencemes other than DMs (e.g. unfilled pauses or repetitions), cf. the introduction to this section.

(7) avec un Liège toujours aussi *euh* faiblard j'ai envie de dire [FR-spor-01]

"with Liège still as *euh* weak so to say"

Additional evidence would require careful examination of the specific contexts of these FPs and how often they occur in sequences marked by interruptions or reformulations, as in example (8), in which case they would be symptoms of disfluency, all the more so as their use is not typical of these registers.

(8) que notre peuple fasse en/ (0.158) s'a/ *euh* s'/ *euh* se fasse entendre (0.505) [FR-poli-07]

"that our nation makes hea/ (0.158) mak/ *euh* ma/ *euh* makes itself heard"

In any case, Table 7 shows that in English and French, FPs are the least frequent in broadcast situations (the last three in the table), which makes private settings the most natural environment for these fluencemes.

4.2.2. *Sample analysis: position and co-occurrence*

Given the large number of occurrences of FPs in *DisFrEn*, it was not possible to describe each item in terms of syntactic position and patterns of co-occurrence. However, in the sample data (i.e. 200 occurrences in each language), such information was manually encoded. As opposed to the positional variables for DMs which distinguished four potential slots, namely initial, medial, final and independent, for FPs we reduced the possibilities to two values: "between" clauses or "within" clauses, since it is not semantically relevant to determine whether a FP "opens" or "closes" an utterance. This dual variable is nevertheless quite informative as to the relative (dis)fluency of the items, with "within" FPs (9) being potentially more disruptive than "between" FPs, which strengthen the marking of syntactic boundaries (10).

(9) now what about what about the *uh* (1.370) the theory that [EN-clas-04]

(10) it's now definite (0.410) *uhm* and I I definitely want to [EN-phon-02]

In the sample, 71% of FPs (284 out of the total 400) occur in isolation (i.e. they do not co-occur with a DM but may combine with other fluencemes). This proportion is very similar between the languages, with 140 isolated FPs in English and 144 in French. For these isolated items, we observe a significant preference for the potentially disruptive "within" position, with 172 cases against 112 at clause boundary ($LL = 12.77$, $p < 0.001$). This difference is reproduced in similar proportions across the two languages: 86 vs. 54 in English, 86 vs. 58 in French. English and French thus display a very similar behavior, with isolated FPs occurring more frequently within their host unit than between clauses, which argues for a negative, disfluent interpretation of FPs in isolation.

As for their co-occurrence patterns, 105 FPs appear jointly with an unfilled pause, out of the 284 items extracted in the sample which were not part of a cluster with a DM, leaving 179 cases of FPs that do not co-occur with a DM nor a pause. Again, this difference is observed across languages in similar proportions, as can be seen in Table 8: the only crosslinguistic contrast is the larger discrepancy between the two types of context (i.e. with or without an unfilled pause) in English than in French, where the ratio is about 1.4 as opposed to 2 in English.

Table 8. Cross-tabulation of co-occurring pauses and position of FPs.

		between	within	Total
English	without pause	30	64	**94**
	with pause	24	22	**46**
	Total	54	86	**140**
French	without pause	34	51	**85**
	with pause	24	35	**59**
	Total	58	86	**144**
Total	without pause	64	115	**179**
	with pause	48	57	**105**
	Total	**112**	**172**	**284**

Table 8 also shows the impact of position on the distribution of co-occurring pauses. We can see that in English the presence of an unfilled pause evens out the difference between the "within" (22) and the "between" (24) position, whereas the former is twice as frequent as the latter when the FP does not co-occur with an unfilled pause (LL = 12.58, $p < 0.001$). In French, both types of context prefer the "within" position in similar proportions. This is quite divergent from Swerts' (1998) finding on Dutch FPs which prefer a "between" position when they are clustered with an unfilled pause.

Overall, this result suggests a closer investigation of the local contexts of FPs to verify whether FPs in "within" position are actually disruptive and hesitant, or rather used for lexical salience: if clause-internal FPs occur more frequently without an unfilled pause, then the interruption in the flow of speech is limited to the duration of the FP, hence restricting their disfluent effect. Browsing through the sample, it appears that many isolated FPs directly precede a discourse-new concept (as in (11)), whereas co-occurrence with unfilled pauses can also be found in the context of a hesitation or mistake, as in example (12).

(11) back at him that was *uh* Kevin Gallagher chasing back [EN-spor-05]

(12) Thursday's oh sorry Wednesday *uh (0.860)* I'm meeting [EN-conv-04]

One tentative interpretation of this tendency would be that FPs are more commonly used as "symptoms" than as "signals" when they combine with unfilled pauses. A qualitative or perceptual analysis could determine whether the combination of filled and unfilled pauses is more disfluent than isolated FPs, regardless of their position.

4.2.3. *Euh, uh and uhm in isolated contexts*

Looking into the specific forms of the FPs, it turns out that some of the differences observed above are in fact explained by language-internal specializations, namely the different behaviors of *uh* vs. *uhm* in English. Our sample analysis partially confirms previous findings in the literature regarding the positions of *uhm* and *uh*: as stated by Clark and Fox Tree (2002) and Shriberg (1994), *uh* is significantly

associated with the "within" position (75% of isolated *uh*s, LL = 16.74, p < 0.001), as in example (13).

(13) since then the *uh* local government has been reformed [EN-poli-07]

However, our data does not support the claim that *uhm* is more frequent at syntactic boundaries, at least not in isolated contexts (without a clustered DM) where we found a perfect equality across "between" and "within" positions, regardless of the presence of an unfilled pause.

Since this study only considers the FP form *euh* for French, such internal comparisons are not possible. In fact, *euh* is quite neutral regarding the variables of position and co-occurrence, without any significant difference between the different configurations, apart from the general results already mentioned above (more isolated FPs without unfilled pauses and more "within" position).

All these results on DMs and FPs taken individually either settle formerly unresolved issues or confirm previous works in contrastive fluency research. We saw particularly interesting differences between more or less formal situations, as well as crosslinguistic patterns of preference between position and co-occurrence.

5. The impact of clustering on the distribution of DMs and FPs

We now turn to the clustering of DMs and FPs to see whether the distinctions established above are confirmed or even refined when DMs and FPs cluster together, and whether additional contrastive differences emerge from their combination. In this section, we will refer to any cluster of DMs and FPs as [DM+FP], regardless of the internal structure of the cluster (whether the DM precedes or follows the FP, number of items in the sequence, etc.).

5.1. Distribution of the clusters

5.1.1. *Frequency*

In *DisFrEn*, after extraction of all DMs forming a cluster with a FP, 474 occurrences were found across the different configurations detailed in Section 3.2. Counting by occurrences of FPs, 428 items were annotated as part of a [DM+FP] configuration; in other words, 27% of all FPs in *DisFrEn* are clustered with a DM. Together, DMs and FPs form 313 clusters, some of them containing several DMs and/or FPs, possibly with other fluencemes as well.

Table 9 shows the number of [DM+FP] clusters in the different subcorpora. We can see that in normalized frequencies, French and English show the same normalized number of clusters, viz. 3 per 1,000 words (LL = 0.58, p > 0.05). The same situational variation noted for the isolated contexts can be observed for the clusters, with a preference for spontaneous, unscripted settings. As for formal, broadcast situations, while isolated DMs and FPs were rare in the news and political subcorpora, clusters are completely absent from these situations in English and insignificant in French, confirming our previous observations.

Table 9. [DM+FP] clusters in absolute and normalized frequencies (per 1,000 words).

Situation	English	1/1000	French	1/1000	Total	1/1000
phone	77	8	39	6	116	10
conversation	55	3	74	4	129	6
classroom	43	5	12	3	55	5
sports	5	1	5	1	10	1
news	0	0	3	0.4	3	0.3
political	0	0	0	0	0	0
Total	**180**	**3**	**133**	**3**	**313**	**4**

Zooming in on the DMs in these clusters, the top three are translation equivalents in the two languages (although with different ranks): *and* (71), *but* (43), *so* (34); *et* 'and' (46), *donc* 'so' (40), *mais* 'but' (32), together accounting for 56% of all clustered DMs. We notice the absence of speech-specific expressions in this ranking, as opposed to isolated DMs which included more typically interactive DMs such as *well* or *quoi* (see Section 4.1.1). In clusters, DMs are mostly conjunctions, which leads to the conclusion that the presence of a FP tends to constrain the semantic nature of the neighboring DM, restricting their functional range to very generic connecting devices.

5.1.2. *Position of the clusters*

Turning to the syntactic position of the clusters, we used the annotation of the DMs as a reference for the whole cluster, given that non-lexical elements do not affect the syntactic annotation. A FP clustered with an initial or final DM will systematically be at a boundary ("between"), while the medial position corresponds to the "within" slot for FPs.

In *DisFrEn*, clause-initial strikes as the most frequent slot (81.2% of the clusters) across all languages and situations, which can be related to the prominence of this position for DMs in general (Schiffrin 1987) and in our corpus as isolated occurrences (see Section 4.1.2). Other positional slots are ranked in the same order as for isolated contexts of DMs, in the same proportions (more than twice as many final clusters in French as in English). Medial position, which was the most frequent for isolated FPs, only covers 6.5% of all clusters, which means that DMs attract FPs, and not the opposite.

5.1.3. *Co-occurrence with unfilled pauses*

Among the 313 [DM+FP] clusters, two thirds (199) contain an unfilled pause, which points to the special attraction between DMs and the two types of pause (filled and unfilled). As we can see in Table 10, this is especially true in English, where the two types of cluster (with or without an unfilled pause) show significantly different frequencies ($LL = 31.34$, $p < 0.001$), as opposed to French where they are almost equally distributed. This is in keeping with the co-occurrence patterns of isolated DMs, where unfilled pauses are more frequent in the vicinity of English DMs, albeit to a lesser extent than in clusters, and also corroborates Grosjean and Deschamps' (1975) findings.

Table 10. Presence of an unfilled pause in the [DM+FP] clusters.

Type of cluster	English	%	French	%	Total	%
With pause	127	70.6%	72	54.1%	199	63.6%
Without pause	53	29.4%	61	45.9%	114	36.4%
Total	180	100%	133	100%	313	100%

This result contributes to the debate regarding the status of FPs as words or not (Clark and Fox Tree, 2002; Corley and Stewart, 2008; Tottie, 2015a): in our view, their frequent co-occurrence with unfilled pauses is part of the evidence that these two phenomena are functionally and cognitively close. We would then argue for grouping them into a broader category of pauses rather than words, although linguistic categories tend to be fuzzy and language change might blur the picture, as Tottie (2015b) has recently shown with occurrences of FPs in written data.

5.2. Clustering configurations

Taking up the configurations of [DM+FP] clusters presented above (see Section 3.2), interesting tendencies emerge from the analysis of the position of FPs in the clusters. We report in Table 11 the frequency of each configuration per language. Only two of these configurations (see lines in bold) show statistically significant differences between the two languages: "DM+FP" ($LL = 13.99$, $p < 0.001$) and "UP+FP+DM" ($LL = 4.50$, $p < 0.01$).

Table 11. Configurations of [DM+FP] clusters in *DisFrEn*.

Configuration	Tag	English	French	Total
FP+DM+UP	FDU	1	1	2
FP+DM+FP	FPB	2	1	3
FP+DM	FPL	28	16	44
DM+FP	**FPR**	**34**	**60**	**94**
DM+FP+UP	FUR	28	17	45
UP+DM+FP	UDF	14	6	20
UP+FP+DM	**UFL**	**36**	**14**	**50**
Other	MIX	37	18	55
Total		**180**	**133**	**313**

The major difference between the two languages lies in the "FPR" configuration (e.g. *well uh*), which shows the largest standardized residuals and is much more frequent in French than the other French configurations, and than its English counterpart. This pattern thus comes out as the most representative French example, as in example (14). In English, however, the distribution is more homogeneous and does not return a clear prototype.

(14) il déteste l'atelier *et euh* c'est loin quoi [FR-conv-05]

"he hates the workshop *et euh* ('and uh') it's far you know"

Looking at the position of these configurations, it appears that the final position, which we previously identified as more frequent in French, is mainly represented by occurrences of *tu vois* 'you see' and *donc* 'so' either in "DM+FP+UP" or

"DM+FP" configurations. These contexts always correspond to the right boundary of a syntactic structure, and even in some cases to the end of the turn, as in example (15). This finding confirms Degand's (2014) observation on the emergence of turn-final *donc euh* as a "typically (turn) final pattern which appears to express a specific meaning, namely that of a conclusive relation that the addressee is invited to infer" (2014: 169).

(15) <spk1> ils s/ enfin ils se mariaient carrément quoi *donc euh*
<spk2> et tous les autres ils sont déjà parents [*DisFrEn* FR-conv-05]
<spk1> "they g/ well they even got married you know *donc euh* ('so uh')
<spk2> and all the others they are already parents"

This type of cluster illustrates the punctuating and interactional function of DMs and FPs, particularly in French where they can be used to signal transition and turn-yielding, thus no longer being the "symptom" of production trouble but a hearer-oriented discourse-functional device.

5.3. *Euh, uh and uhm in clusters*

When comparing the FP forms, some contrasts emerge, with a higher frequency of *uh* in "other" and "UP+DM+FP" patterns, as in example (16), and a higher frequency of *uhm* in "DM+FP+UP" and "UP+FP+DM", as in example (17). French *euh* does not resemble one English FP much more than the other, its specificity rather lying in the "DM+FP" configuration detailed above (Section 5.2).

(16) I did I did the *(0.200) sort of uh* acousticky bit [EN-conv-06]

(17) come through *(0.580) uhm but I mean* why did he get [EN-clas-02]

However, in sharp contrast with our previous results regarding the position of isolated FPs, the distribution of *uh* and *uhm* can no longer be distinguished, which leads to the conclusion that the presence of a DM in the vicinity of a FP "erases" the positional specificities of the two English FPs.

5.4. Can we predict the clustering of DMs and FPs?

So far, our results have shown a significant impact of the clustering of DMs and FPs, either by specifying contrastive differences (e.g. FP at the right of a DM ("FPR") as the French prototype; high frequency of the final position in French; more frequent co-occurrence with unfilled pauses in English) or cancelling the effect of previous distinctions (e.g. association of isolated *uh* and *uhm* with "within" or "between" positions; generic conjunctions in the top three clusters for both languages as opposed to more speech-specific lexemes), while other observations were merely confirmed in combined contexts (e.g. preference for non-broadcast situations, position of DMs). In this last results section, a tentative model of the contextual and discourse-functional clustering of DMs and FPs will be statistically computed in order to articulate all variables analysed so far. If the crosslinguistic and language-specific tendencies that we observed are robust enough, we should be able to reliably predict the features of context that will trigger the clustering of DMs and FPs.

To do so, we ran a stepwise binomial logistic regression on the sample data, which returned the variables *presence of an unfilled pause* and *position=within* as highly significant to predict the environment of a FP (i.e. either clustered with a DM or not). Since the overall predictive power of this model is just below acceptability ($C = 0.762$, $p < 0.01$), multiple correspondence analysis (MCA) was also computed to visualize the interplay of these factors (see Figure 1). Good variance coverage in the model (51.42%) allows us to draw a number of conclusions.

FIGURE 1 SHOULD BE INSERTED NEAR HERE

Figure 1 shows that, apart from *euh* which is neutral (i.e. occurs indiscriminately in all types of local contexts), the other variables, viz. form of the FP ("uh" and "uhm"), presence of an unfilled pause ("no_pause", "yes_pause"), syntactic position ("within", "between") and environment (in a "cluster" or "alone") tend to co-vary together in two distinct patterns, on each side of the graph, respectively illustrated in examples (18) and (19).[7]

(18) society had outgrown *uh* police institutions (0.347) and the time

[EN-poli-07: alone, no unfilled pause, clause-internal]

(19) oh right well it's now definite *(0.410) uhm and* I I definitely want

[EN-phon-02: cluster, with unfilled pause, clause-initial]

Conversely, situational variation does not produce any significant effect on the clustering of FPs and DMs, which is further proof that any interpretation of fluency solely based on individual – albeit important – factors is a simplistic view of how (dis)fluent discourse unfolds in natural speech. In accordance with our definition of (dis)fluency as a sequential, situational and ambivalent phenomenon, even clear-cut profiles such as the ones presented in this section need to be interpreted carefully by taking into account other co-textual information (e.g. part of speech of surrounding words, presence of other fluencemes) and scientific knowledge (e.g. role of FPs as discourse-new expectation triggers, cf. Barr and Seyfeddinipur, 2010; Bosker *et al.*, 2014). Based on this multivariate statistical analysis alone, jumping to the conclusion that clusters of FPs, DMs and unfilled pauses at syntactic boundaries are more fluent than clause-internal individual FPs would completely overlook the facts that (i) boundaries are typical projective slots where hesitations related to speech planning are often found (Hawkins, 1971), and (ii) internal FPs have repeatedly been shown to precede salient lexical information (e.g. Arnold *et al.*, 2003, 2007; Barr and Seyfeddinipur, 2010).

At this preliminary interpretative stage, what our corpus analysis can confirm is that the clustering of FPs and DMs appears to be discourse-functional, and that French and English clusters behave in similar patterns, strongly affected

[7] The distance between points in MCA graphs represents the statistical association or co-variation of variables. It should roughly be read as: the closer the values on the graph, the more closely associated they are.

by surface features which effectively predict the environment of FPs without any contrastive preference.

6. Conclusion

This study has shown that FPs function differently whether they are clustered with DMs or not, and this difference consists in either maintaining or erasing inter- and intra-linguistic contrasts. In particular, a French specificity emerges from the clustered contexts (viz. DM followed by a FP at the right periphery of the utterance), whereas the differences between English *uh* and *uhm* no longer hold when clustered with DMs. The presence of a DM thus explains the situational and syntactic distribution of the clusters, and any remaining variation is either due to contrastive or discourse-functional preferences. Overall, previous findings on French and English FPs have been confirmed and refined by our systematic comparison of individual vs. clustered contexts of these two fluencemes, which stand out as particularly ambivalent and multifunctional.

The multifunctionality that characterizes DMs and FPs also applies to many other fluencemes, and pleads for an integrated approach to fluency studies such as ours, covering a wide range of discourse devices and taking into account different layers of linguistic and contextual information. One particularly informative – yet challenging – source of knowledge is function or meaning-in-context, which could refine the purely formal approach adopted in this paper. By combining syntactic (position), pragmatic (function) and syntagmatic variables (co-occurrence and clusters of other types of fluencemes), as well as rich metadata (language, register, possibly speaker information), one might obtain a more comprehensive picture of the relative contribution of particular structures to the overall perception of fluent and disfluent discourse.

What quantitative analyses such as those proposed here bring to the contrastive study of (dis)fluency markers is a robust method to identify discourse-functional tendencies in different populations, observe how specific structures are used in various registers, and uncover both universal and language-specific profiles. We have shown some of the intertwined factors that influence the use of DMs and FPs, and how the same pattern can be used in very different ways (e.g. utterance-medial FPs as symptoms of repair or as signals of lexical salience). One promising avenue for corpus-based fluency research is to dig further into the relationship between corpus frequency, cognitive salience or prototypicality (see Gilquin, 2006, 2008 on these notions) and perceived fluency, testing the hypothesis that high frequency of use leads to higher cognitive entrenchment and therefore facilitates the production and perception of a particular structure. Nevertheless, we would argue that going a step further and actually interpreting the relative fluency of utterances would require the combination of various quantitative and qualitative methods, namely experimentation, corpus linguistics, conversation analysis and possibly others.

References

Aijmer, K. 1997. *I think* – An English Modal Particle. In *Modality in Germanic Languages. Historical and Comparative Perspectives*, T. Swan and O. Westvik (eds), 1-47. Berlin: Mouton de Gruyter.

Aijmer, K. and Simon-Vandenbergen, A.-M. 2006. *Pragmatic Markers in Contrast*. Amsterdam: Elsevier.

Arnold, J., Fagnano, M. and Tanenhaus, M. 2003. Disfluencies Signal theee, um, New Information. *Journal of Psycholinguistic Research* 32(1): 25-36.

Arnold, J., Hudson-Kam C. and Tanenhaus, M. 2007. If you ay thee uh you are describing something hard: The On-line Attribution of Disfluency during Reference Comprehension. *Journal of Experimental Psychology: Learning, Memory and Cognition* 33(5): 914-930.

Auer, P. 2005. Delayed Self-repairs as a Structuring Device for Complex Turns in Conversation. In *Syntax and Lexis in Conversation*, A. Hakulinen and M. Selting (eds), 75-102. Amsterdam: John Benjamins.

Barr, D. and Seyfeddinipur, M. 2010. The Role of Fillers in Listener Attributions for Speaker Disfluency. *Language and Cognitive Processes* 25(4): 441-455.

Bazzanella, C., Bosco, C., Garcea, A., Gili Fivela, B., Miecznikowsky, J. and Tini Brunozzi, F. 2007. Italian *allora*, French *alors*: Functions, Convergences and Divergences. *Catalan Journal of Linguistics* 6: 9-30.

Bolly, C., Crible, L., Degand, L. and Uygur-Distexhe, D. (in press). Towards a Model for Discourse Marker Annotation. From Potential to Feature-based Discourse Markers. In *Discourse Markers, Pragmatic Markers and Modal Particles: New Perspectives*, C. Fedriani and A. Sansó (eds). 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.

Bosker, H.R., Quené, H., Sanders, T. and de Jong, N. 2014. Native 'um's Elicit Prediction of Low-frequency Referents, but Non-native 'um's Do Not. *Journal of Memory and Language* 75: 104-116.

Brennan, S.E. and Schober, M.F. 2001. How Listeners Compensate for Disfluencies in Spontaneous Speech. *Journal of Memory and Language* 44: 274-296.

Brinton, L. 1996. *Pragmatic Markers in English. Grammaticalization and Discourse Functions*. New York: Mouton de Gruyter.

Broen, P. and Siegel, G. 1972. Variations in Normal Speech Disfluencies. *Language and Speech* 15: 219-231.

Brognaux, S., Roekhaut, S., Drugman, T. and Beaufort, R. 2014. Train&Align: Un Outil d'Alignement Phonétique Automatique Disponible en Ligne. Paper presented at the Journées d'étude de la parole (JEP), Le Mans.

Clark, H. and Fox Tree, J. 2002. Using *uh* and *um* in Spontaneous Speaking. *Cognition* 84: 73-111.

Corley, M. and Stewart, O. 2008. Hesitation Disfluencies in Spontaneous Speech: the Meaning of *um*. *Language and Linguistics Compass* 2(4): 589–602.

Crible, L. 2014. Identifying and Describing Discourse Markers in Spoken Corpora. Annotation Protocol v.8. Unpublished working draft, Université Catholique de Louvain.

Crible, L. 2015. Étude Contrastive des Marqueurs de Discours Français et Anglais: Approche Onomasiologique sur Corpus Comparable. Paper presented at the 4th International Symposium "Discourse Markers in Romance Languages: a Contrastive Approach", Heidelberg, 6-9 May 2015.

Crible, L. (in press). Towards an Operational Category of Discourse Markers: A Definition and its Model. In *Discourse Markers, Pragmatic Markers and Modal Particles: New Perspectives*, C. Fedriani and A. Sansó (eds). Amsterdam: John Benjamins.

Crible, L., Dumont, A., Grosman, I. and Notarrigo, I. 2016. Annotation Manual of Fluency and Disfluency Markers in Multilingual, Multimodal, Native and Learner Corpora, v.2.0. Technical Report, Université Catholique de Louvain and Université de Namur.

Crible, L., Zufferey, S. 2015. Using a Unified Taxonomy to Annotate Discourse Markers in Speech and Writing. In *Proceedings of the 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (isa-11), 14 April 2015, London*, H. Bunt (ed.), 14-22.

Cuenca, M.-J. 2003. Two Ways to Reformulate: A Contrastive Analysis of Reformulation Markers. *Journal of Pragmatics* 35: 1069-1093.

Defour, T., D'Hondt, U., Vandenbergen, A.-M. and Willems, D. 2010. *In fact, en fait, de fait, au fait*: a Contrastive Study of the Synchronic Correspondences and Diachronic Development of English and French Cognates. *Neuphilologische Mitteilungen* 111(4): 433-463.

Degand, L. 2014. 'So very fast, very fast then' Discourse Markers at Left and Right Periphery in Spoken French. In *The role of the Left and Right Periphery in Semantic Change: Crosslinguistic Investigations of Language and Language Change*, K. Beeching and U. Detges (eds), 151-178. Brill: Leiden.

Degand, L. and Gilquin, G. 2013. The Clustering of 'Fluencemes' in French and English. Paper presented at the *7th International Contrastive Linguistics Conference (ICLC 7) – 3rd Conference on Using Corpora in Contrastive and Translation Studies (UCCTS 3)*, Ghent, 11-13 July 2013.

Dister, A., Francard, M., Hambye, P. and Simon, A.-C. 2009. Du Corpus à la Banque de Données. Du Son, des Textes et des Métadonnées. L'Évolution de Banque de Données Textuelles Orales VALIBEL (1989-2009). *Cahiers de Linguistique* 33(2): 113-129.

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)*, Poznań, 11-14 September 2014.

Eklund, R. and Shriberg, S. 1998. Crosslinguistic Disfluency Modelling: a Comparative Analysis of Swedish and American English Human-human and Human-machine Dialogs. Paper presented at the *5th International*

Conference on Spoken Language Processing, Sydney, 30 November-4 December 1998.

Fagard, B. and Degand, L. 2010. Cause and Subjectivity, a Comparative Study of French and Italian. *Lingvisticae Investigationes: Revue Internationale de Linguistique Française et de Linguistique Générale* 33(2): 179-193.

Gilquin, G. 2006. The Place of Prototypicality in Corpus Linguistics. Causation in the Hot Seat. In *Corpora in Cognitive Linguistics: Corpus-based Approaches to Syntax and Lexis*, S. Gries and A. Stefanowitsch (eds), 159-191. Berlin: Mouton de Gruyter.

Gilquin, G. 2008. What You Think ain't what You Get: Highly Polysemous Verbs in Mind and Language. In *Du Fait Grammatical au Fait Cognitif. From Gram to Mind: Grammar as Cognition*. *Volume 2*, J.-R. Lapaire, G. Desagulier and J.-B. Guignard (eds), 235-255. Pessac: Presses Universitaires de Bordeaux.

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. and 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: 144-184.

Grosman, I. 2016. How do French Humorists Manage their Persona across Situations? A Corpus Study on their Prosodic Variation. In *Metapragmatics of Humor: Current Research Trends*, L. Ruiz-Gurillo (ed.), 147-175. Amsterdam: John Benjamins.

Guillemin-Flescher, J. 1981. *Syntaxe Comparée du Français et de l'Anglais*. Paris: Ophrys.

Hasselgård, H. 2014. Discourse-structuring Functions of Initial Adverbials in English and Norwegian News and Fiction. *Languages in Contrast* 14: 73-92.

Hasselgren, A. 2002. Learner Corpora and Language Testing: Smallwords as Markers of Learner Fluency. In *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, S. Granger, J. Hung and S. Petch-Tyson (eds), 143-173. Amsterdam: John Benjamins.

Hawkins, P.R. 1971. The Syntactic Location of Hesitation Pauses. *Language and Speech* 14: 277-288.

Hieke, A. 1985. A Componential Approach to Oral Fluency Evaluation. *The Modern Language Journal* 69(2): 135-142.

Levelt, W. 1989. *Speaking. From Intention to Articulation*. Cambridge: MIT Press.

Maclay, H. and Osgood, C. 1959. Hesitation Phenomena in Spontaneous English Speech. *Word* 15: 19-44.

Mahl, G.F. 1987. *Explorations in Nonverbal and Vocal Behavior*. Hillsdale: Erlbaum.

Merlo, S. and Mansur, L. 2004. Descriptive Discourse: Topic Familiarity and Disfluencies. *Journal of Communication Disorders* 37: 489-503.

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

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

Notarrigo, I., Meurant, L. and Simon, A.-C. 2016. Repetition of Signs according to Language Background. Paper presented at the *12th Conference on Theoretical Issues in Sign Language Research (TISLR)*, Melbourne, 4-7 January 2016.

O'Connell, D. and Kowal, S. 2005. *Uh* and *um* Revisited: are they Interjections for Signaling Delay? *Journal of Psycholinguistic Research* 34: 555-576.

O'Donnell, W. and Todd, L. 1980. *Variety in Contemporary English*. London: Allen and Unwin.

Oviatt, S. 1995. Predicting Spoken Disfluencies during Human-computer Interaction. *Computer Speech and Language* 9.

Ragan, S. 1983. Alignment and Conversational Coherence. In *Conversational Coherence: Form, Structure and Strategy*, R. Craig and K. Tracy (eds), 157-171. Beverly Hills: Sage Publications.

Roberts, B. and Kirsner, K. 2000. Temporal Cycles in Speech Production. *Language and Cognitive Processes* 15(2): 129-157.

Roekhaut, S., Brognaux, S., Beaufort, R. and 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.

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

Schmidt, T. and Wörner, K. 2009. EXMARaLDA – Creating, Analysing and Sharing Spoken Language Corpora for Pragmatic Research. *Pragmatics* 19: 565-582.

Schneider, U. 2014. *Frequency, Hesitations and Chunks. A Usage-based Study of Chunking in English*. Freiburg: NIHIN Studies.

Shriberg, E. 1994. Preliminaries to a Theory of Speech Disfluencies. PhD thesis, University of California at Berkeley.

Stenström, A.-B. 1990. Pauses in Monologue and Dialogue. In *The London-Lund Corpus of Spoken English: Description and Research*, J. Svartvik (ed.), 211-252. Lund: Lund University Press.

Swerts, M. 1998. Filled Pauses as Markers of Discourse Structure. *Journal of Pragmatics* 30: 485-496.

Tottie, G. 2011. *Uh* and *um* as Sociolinguistic Markers in British English. *International Journal of Corpus Linguistics* 16: 173-197.

Tottie, G. 2015a. *Uh* and *um* in British and American English: Are they Words? Evidence from Co-occurrence with Pauses. In *Linguistic variation:*

Confronting Fact and Theory, N. Dion, A. Lapierre and R. Torres Cacoullos (eds), 38-54. New York/ Routledge.

Tottie, G. 2015b. From Pause to Word: *Uh* and *um* in Written Language. Paper presented at *ICAME 36*, Trier, 27-31 May 2015.

Vasilescu, I., Nemoto, R. and Adda-Decker, M. 2007. Vocalic Hesitations vs Vocalic Systems: a Cross-language Comparison. In *Proceedings of the ICPhS 16th International Congress of Phonetic Science*.

Vinay, J.-P. and Darbelnet, J. 1995 [1958]. *Comparative Stylistics of French and English: A Methodology for Translation*. Translated and ed. by J. Sager and M.-J. Hamel. Amsterdam: John Benjamins.

Willems, D. and Demol, A. 2006. *Vraiment* and *Really* in Contrast: When Truth and Reality Meet. In *Pragmatic Markers in Contrast*, K. Aijmer and A.-M. Simon-Vandenbergen (eds), 215-235. Amsterdam: Elsevier.

Zhao, Y. and Jurafsky, D. 2005. A Preliminary Study of Mandarin Filled Pauses. In *Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop, September 10-12, Aix-en-Provence, France*, 179-182.

Corresponding author's address

Ludivine Crible

Institute for Language & Communication

Université catholique de Louvain

Place Blaise Pascal, 1

1348 Louvain-la-Neuve

Belgium

ludivine.crible@uclouvain.be

Liesbeth Degand

Institute for Language & Communication

Université catholique de Louvain

Place Blaise Pascal, 1

1348 Louvain-la-Neuve

Belgium

liesbeth.degand@uclouvain.be

Gaëtanelle Gilquin

Institute for Language & Communication

Université catholique de Louvain
Place Blaise Pascal, 1
1348 Louvain-la-Neuve
Belgium
gaëtanelle.gilquin@uclouvain.be