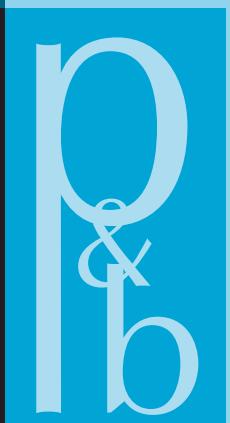
Discourse Markers and (Dis)fluency

Forms and functions across languages and registers

Ludivine Crible



John Benjamins Publishing Company

Discourse Markers and (Dis)fluency

Pragmatics & Beyond New Series (P&BNS)

ISSN 0922-842X

Pragmatics & Beyond New Series is a continuation of Pragmatics & Beyond and its Companion Series. The New Series offers a selection of high quality work covering the full richness of Pragmatics as an interdisciplinary field, within language sciences.

For an overview of all books published in this series, please see *http://benjamins.com/catalog/pbns*

Volume 286

Discourse Markers and (Dis)fluency
Forms and functions across languages and registers
by Ludivine Crible

Discourse Markers and (Dis)fluency

Forms and functions across languages and registers

Ludivine Crible
Université catholique de Louvain

John Benjamins Publishing Company
Amsterdam / Philadelphia

The paper used in this publication meets the minimum requirements of the American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984.

DOI 10.1075/pbns.286

Cataloging-in-Publication Data available from Library of Congress:
LCCN 2017059002 (PRINT) / 2018000403 (E-BOOK)

ISBN 978 90 272 0046 4 (HB)
ISBN 978 90 272 6430 5 (E-BOOK)

John Benjamins Publishing Company · https://benjamins.com

Table of contents

List of figures

List of tables

List of abbreviations and acronyms

A-repair	(generic) appropriateness repair
AA-repair	ambiguity appropriateness repair
AL-repair	level of precision appropriateness repair
AR	misarticulation
CC	coordinating conjunction (also coord. conj)
CLAS	subcorpus of classroom lessons
CONV	subcorpus of conversations
D-repair	delay repair
D-sequence	sequence containing only discourse marker(s)
DE	deletion
DM	discourse marker
E-repair	error repair
EF-repair	phonetic error repair
EL-repair	lexical error repair
EN	English
EP	editing phase
ES-repair	syntactic error repair
ET	explicit editing term
FS	false-start
FP	filled pause
FR	French
F-sequence	sequence containing false-starts and/or truncations
FTF	face-to-face
ICE-GB	British component of the International Corpus of English
IDE	ideational domain
IL	lexical insertion
INT	interpersonal domain
INTF	subcorpus of face-to-face interviews
INTR	subcorpus of radio interviews
IP	parenthetical insertion
JJ	adjective
L1	first language
L2	second language
LEFT	left-integrated macro-syntactic position
LL	log-likelihood
MCA	multiple correspondence analysis

MDMA	Model for Discourse Marker Annotation
MID	middle-field macro-syntactic position
NEWS	subcorpus of news broadcasts
NN	noun phrase
OR	change of order
PDF	Potentially Disfluent Function
PDTB	Penn Discourse TreeBank
PHON	subcorpus of phone calls
POLI	subcorpus of political speeches
POS	part-of-speech
POST	post-field macro-syntactic position
PP	preposition phrase
PRE	pre-field macro-syntactic position
P-sequence	sequence containing discourse markers and pauses
PTW	per thousand words
R-repair	resonance repair
RB	adverb
RHE	rhetorical domain
RIGHT	right-integrated macro-syntactic position
R-sequence	sequence containing repetitions
RI	identical repetition
RM	modified repetition
SC	subordinating conjunction (also subord. conj.)
SEQ	sequential domain
SM	morphosyntactic substitution
SP	propositional substitution
SPOR	subcorpus of sports commentaries
S-sequence	sequence containing substitutions
TF	turn-final position
TI	turn-initial position
TM	turn-medial position
TR	truncation
TT	whole turn position
UH	interjection
UP	unfilled pause
VP	verbal phrase
WI	within
WP	pronoun
Z-sequence	sequence combining false-starts and/or truncations with repetitions and/or substitutions

Acknowledgments

This book is a revised version of my doctoral dissertation, completed at the Université catholique de Louvain (Belgium) in February 2017. I therefore wish to thank the many wonderful people who contributed to making my four years of PhD an amazing personal and professional experience. First and foremost, I have been carefully coached and tutored by not one but two dedicated promoters, the professors Liesbeth Degand and Gaëtanelle Gilquin, who managed to never contradict each other and instead provided me with complementary input in the most fruitful and educative way a PhD student can wish for. Liesbeth and Gaëtanelle, thank you for your restless guidance and your friendship!

I was also lucky to be a member of two research centers of the Linguistic Research Unit at the UCL: the CECL – Center for English Corpus Linguistics and Valibel – Discours et Variation, which both provided their share of inspiring seminars, friendly feedback and birthday parties. Many thanks to my colleagues of both sides. A special thought goes to the Fédération Wallonie-Bruxelles, which funded this research, and in particular to the members of the Concerted Research Action "Fluency and Disfluency Markers", both in Louvain-la-Neuve and Namur, who taught me the merits (and challenges!) of teamwork and whose contribution to the present book is substantial.

I thank Prof. Dr. Anita Fetzer for her guidance and efficiency as editor of *Pragmatics and Beyond New Series*, the reviewers of the manuscript for their careful and constructive suggestions, and Isja Conen from John Benjamins Publishing Company for her help with the publication. Any remaining errors are mine.

I am also particularly indebted to the COST network "TextLink", which gave me the opportunity to work with experts in my field all across Europe, in particular the professors Sandrine Zufferey and María-Josep Cuenca as well as the friendly advice of Pr. Ted Sanders and many other members of the Discourse Marker community here and abroad (Sílvia Gabarró-López, Elena Pascual, Karolina Grzech and more). Another group to have welcomed me at the beginning of my PhD was the MDMA team led by my colleague and friend Dr. Catherine Bolly, whose dynamism and *joie de vivre* never cease to amaze me.

A PhD is not just a professional adventure and I could not have made it to the end (relatively) sanely without the help of my friends. Cheers to the Big Five for our shared love of beer and boardgames. Much love to my Parisian Team

Rocket – always. And of course, to my French and Belgian families who loved and supported me, even though they did not always understand me. Last but not least, to my loving teammate, Kévin Libion, who suffered endless blabber about discourse markers, post-conference debriefs and office gossip with true stoic grace, cared for me, fed me, always believed in me, my most faithful supporter. Collector!

CHAPTER 1

Introduction

"Spoken language exists in time, not space"
Carter & McCarthy (2006: 193)

1.1 Fluency in time and space

Linguistic theory has made ample use of metaphors throughout the century of its existence to refer to otherwise complex mechanisms of production and perception, in agreement with their general function in our everyday experience (Lakoff & Johnson 1980). In the field of spoken language studies and in particular spoken fluency, one such popular metaphor is that of language as motion, more precisely "frictionless motion" (Ginzburg et al. 2014: 10) when referring to fluent speech. In the same line of thought, many definitions of fluency evoke the idea of fluidity, picturing (idealized) speech as the smooth unfolding of a stream of words (e.g. Crystal 1987; Koponen & Riggenbach 2000; Segalowitz 2010).

Although the notion of language-as-motion is compelling, as attested by its recurrence in many notable works in the field, I would like to introduce a new metaphor that helps better understand the dynamics and constraints of spoken language and provides a productive framework to investigate the concept of fluency, viz. the spacetime continuum: the (metaphorical) spatial dimension of speech (as in "drawing" parallels between utterances, "bridging" over a digression, "retracing" and "editing" a previous statement) can only be conceptualized in relation to how time pressures the production (avoiding long silences, managing working memory load) and leaves no hard trace but an evanescent product (*verba volant, scripta manent*: "spoken words fly away, written words remain"). In other words, while written texts can be primarily described as graphic (spatial) objects, speech is resolutely multidimensional, combining spatial-like moves with temporal constraints. As a result, comparing spoken fluency with motion shows the influence of writing-based accounts of language and might be overlooking core differences between the two modalities.

The spacetime metaphor is in fact motivated by the very phenomenological nature of speech as rooted in the present while at the same time constantly "moving" between retentions and protentions (Deppermann & Günthner 2015, quoting Husserl 1964). The introductory quote by Carter & McCarthy (2006) highlights the unique character of speech as opposed to writing and, to a lesser extent, sign

language: speakers and listeners cannot "rewind" nor "fast forward" but are stuck in the linear flow of speech. In this, speech contrasts (1) with written texts, which are not constrained by the same time pressure and where writers and readers are free to navigate across the different paragraphs (Danks & End 1987) and (2) with sign languages, which offer some simultaneity thanks to the relative autonomy of each hand, although limited to non-contradictory and non-independent content (Levelt 1981). Speech, on the other hand, is restricted to the linearity of the phonological channel and does not afford the same freedom of movement as graphic writing.[1] And yet, speech is still often evaluated against a "written language bias" (Linell 1982) of ideal linguistic output as a smooth, uninterrupted flow of words, completely denying the temporal nature of online production. Equating fluency with flawless fluidity is therefore not true to the cognitive processes of language production and perception, and particularly unrealistic for spontaneous, unplanned speech.

In this work, I will strive to show that non-standard structures such as so-called disfluencies are not systematically problematic (as opposed to what a writing-based standard of fluency would argue) but can actually create coherent and efficient discourse. Disfluencies have been extensively described in the literature as potentially strategic and discourse-functional, especially in recent frameworks (e.g. dialogic syntax, Du Bois 2014) where they are interpreted as productive, hearer-oriented uses of conversational grammar. In particular, several studies have repeatedly shown that clusters of disfluencies help identify major discourse boundaries (e.g. Rendle-Short 2004) and trigger other local structuring effects such as generating expectations (e.g. Arnold & Tanenhaus 2011) or creating lists (e.g. Auer & Pfänder 2007). In other words, disfluencies can be viewed as "tricks" that allow speakers to restitute a spatial dimension to the temporality of speech by manifesting the directionality of particular discourse moves. By pursuing such a growing line of research, this book answers Auer's (2009) call for more research taking the notion of temporality as central in the study of speech.

This approach focuses on (dis)fluency markers that have a direct impact on the structure of discourse, such as marking boundaries or connecting utterances. "Discourse markers" (e.g. Schiffrin 1987), i.e. pragmatic expressions such as *but* or *I mean*, fulfil this structuring role and are therefore the central focus of this study, which investigates their many forms and functions and studies their combination with other (dis)fluent devices such as pauses or repetitions in different

1. Co-verbal gestures are an important feature of face-to-face interactions and can convey some meaning which is not necessarily fully redundant or even compatible with the verbal content (Poggi & Magno Caldognetto 1996; Colleta et al. 2009; Bolly & Thomas 2015). However, gestures are only available in face-to-face interaction and will not be considered here as part of the spoken linguistic system *per se*.

configurations. The role of discourse markers in fluency and disfluency is particularly well illustrated outside academia by the many websites and tutorials giving advice on how to use or not use discourse markers. For example, a 2008 article from the LanguageLog website reports on US Senator Caroline Kennedy's receiving bad press during her campaign because of "some cringing verbal tics that showed her inexperience as a speaker", pointing out that she produced more than 200 *you know*s and many *um*s in a 30-minute interview.[2] By contrast, a Youtube video entitled "English fillers to speak fluently and confidently" gives advice on how to use some expressions – such as *you know* – to sound native-like and fluent.[3] Many similar online articles and videos point to a duality between disruptive (even annoying) vs. strategic uses of discourse markers, thus motivating a more thorough, scientific investigation of these varied expressions and their contribution to (dis)fluent discourse.

This very duality or ambivalence is central to the present approach to (dis)fluency insofar as the study does not restrict its scope to either "symptoms" (of production trouble) or "signals" (of an inference to be made). "Symptoms" and "signals" (Clark & Fox Tree 2002) are two sides of the same coin, and it is argued throughout this study that it is only through a cluster of contextual and linguistic variables that, for a single element, the diagnosis can be made. Most classification schemes (e.g. Shriberg 1994; Meteer et al. 1995; Strassel 2003; Besser & Alexandersson 2007) seem to draw the line between "fluent" and "disfluent" uses, excluding the former from their typology by arguing that, e.g., "fluent" pauses or discourse markers are supposedly part of the speaker's intention. Contrary to these *a priori* exclusions, the present approach aims at exhaustivity through the lens of functional ambivalence, a notion which provides a framework that can deal with both symptomatic and signposting effects of disfluencies.

The major challenge addressed by such a program is to create a scale of fluency against which local contexts of clustered disfluencies can be interpreted. However, a more realistic ambition will be pursued: to use the functional and positional features of discourse markers to interpret the relative fluency of the clusters they occur in, through the converging use of evidence from different types (formal, functional and contextual variables). Another source of information to feed this scale of fluency is to use frequency as a clue to the degree of cognitive entrenchment, and thus relate it to the ease of production and comprehension. The more frequent a certain pattern, the more accessible it is for speakers and listeners, following assumptions from usage-based linguistics. Since this research deals with native speakers, such an

2. http://languagelog.ldc.upenn.edu/nll/?p=964. Last accessed on Mar. 21st, 2017.

3. https://www.youtube.com/watch?v=tKmkB7OVO_M. Last accessed on Mar. 21st, 2017.

approach is compatible with the use of authentic data as "abstracted corpus norm", representative of the (dis)fluency standard in a given population (Esser 1993; Götz 2013), in this case speakers of British English and French. These two languages will be studied contrastively in order to identify both distinctive and shared features in how native speakers handle the "intrinsic troubles" of their mother tongue (Schegloff et al. 1977: 381).

In sum, the purpose of this research is to uncover the strategic uses of disfluencies in relation to discourse structure, here understood in a broad sense as local and global management of discourse units, through the specific lens of discourse markers (henceforth DMs) in English and French. In doing so, it will become clear how both "fluent" and "disfluent" uses can be combined in the same typology, and how they form a scale or continuum – to borrow the term of the spacetime metaphor – rather than clear-cut categories.

1.2 Background and objectives

Despite the relative novelty of this joint study of discourse markers and disfluency, it owes many of its conceptual foundations to previous research, especially in its concrete application to corpus data. Research on fluency has been a major trend in linguistics since the 1960s – the first crucial reference in the field being Maclay & Osgood (1959) – and is still growing today. Not only do the different works cover many types and subtypes of phenomena related to the abstract construct of fluency, but they are also very varied in terms of their theoretical and methodological frameworks. Within this literature, a number of well-researched topics will not be addressed in this book, such as pathological disfluency (e.g. Mahesha & Vinod 2012), non-native speech (e.g. Chambers 1997), temporal variables (e.g. Goldman-Eisler 1968), perceptual and psycholinguistic effects (e.g. Corley et al. 2007) or computational applications for detection and removal of disfluencies (e.g. Mieskes & Strube 2008).

Similarly, discourse marker research has expanded considerably since the 1980s to explore the many dimensions of their syntactic and pragmatic behavior in monolingual and multilingual, spoken and written data (see Fischer 2014 for a recent overview). While some studies aim at automatic identification and disambiguation of discourse markers, others, like the present one, are more descriptive and show the interplay of their many characteristics, which can also be relevant for computational purposes.

Against this background of existing research, a number of gaps in both fields need to be filled and motivate the present study. The major gap is probably the quasi-absence of crosslinguistic fluency research. Contrastive fluency has very rarely

been pursued at a large scale (with the exception of Eklund & Shriberg 1998), and never for the English-French pair since Grosjean & Deschamps (1975). A few contrastive case studies do exist and shed some light on individual fluency-related phenomena: O'Connell & Kowal (1972) on pauses; Fox et al. (1996) on syntactic repair; Fox Tree (2001) and Vasilescu et al. (2007) on fillers. By contrast, discourse markers have been widely studied crosslinguistically (e.g. the papers in the edited volume by Aijmer & Simon-Vandenbergen 2006), however not in direct relation to fluency and disfluency – many of them actually work on discourse markers in writing.

This book aims at addressing this double gap in the field, namely studying contrastive (native) fluency in English and French and relating discourse markers to their role in (dis)fluency. Considering discourse markers as full-fledged markers of (dis)fluency will reconcile mainstream DM studies with the research community on fluency, which acknowledges the role of discourse markers in speech quality (especially for naturalness and speech flow) but rarely includes them in their analyses, or only covers a selected few (e.g. Hasselgren 2002; Müller 2005; Denke 2009; Götz 2013). A third related goal of this research is to complement the numerous case studies on particular DM expressions with a more systematic, corpus-based investigation of the whole DM category, thus reconciling not only two objects of study but also two methodological trends, i.e. qualitative discourse analysis and quantitative corpus annotation. In this respect, the present research stands as rather innovative against both DM and fluency research.

In addition to supplementing the current state of the art, this study follows a three-fold empirical objective: (1) to identify and characterize different types of discourse markers in a comprehensive and fine-grained portrait of the whole category; (2) to describe how discourse markers combine with other disfluencies; (3) to interpret the relative fluency of different types of such combinations on the basis of their corpus distribution. The underlying methodological objectives are (4) to offer a reliable annotation model for these linguistic categories and (5) to test the limits of the insights that one can gain from such a statistical approach to a complex (and primarily perceptive) phenomenon. In the context of the "big data" trend in linguistics, this monograph advocates for manual, qualitative analysis of highly enriched datasets, provided they are combined with sound quantitative methods.

1.3 Preview of the book

This monograph includes seven chapters besides introduction and conclusion: two theoretical, one methodological and four empirical. The next chapter (Chapter 2) develops the present ambivalent and componential approach to spoken fluency and situates this study against the background of fluency research, focusing on

corpus-based works. It will appear from the literature review that the originality of the present framework lies in its inclusion of non-disfluent, functionally ambivalent elements of speech as potential markers of (dis)fluency. The assumptions of usage-based linguistics and their application to the present object of study will also be developed, pointing in particular to the role of co-occurrence patterns, context and frequency.

Chapter 3 will be dedicated to discourse markers, which are considered as one type of (dis)fluency marker. Among the vast body of research on this complex category, a selective review of the literature will identify the core features of definition as well as major annotation frameworks which were highly influential in the present methodology, focusing in particular on the functional spectrum of discourse markers. The formal-functional definition adopted in this study will be presented. The specific challenges of a contrastive bottom-up approach to the highly polyfunctional DM category will be discussed, taking stock of previous research targeting written language as well. The link between discourse markers and (dis)fluency will also be developed in light of the notion of functional ambivalence and in relation to the (relatively small) literature combining these two objects of study.

Chapter 4 presents the dataset and methodology, detailing the annotation schemes for DMs and (dis)fluency markers. The definitions and notions introduced in Chapters 2 and 3 will be operationalized.

In Chapter 5, a corpus-based portrait of the DM category in several registers of English and French will be drawn from a systematic analysis of all DM-level variables (part of speech, position, function, co-occurrence). Univariate and multivariate analyses will make use of a range of frequency-based and other statistical methods in order to test the centrality of DM features often mentioned in the literature such as initiality. In particular, the integration of positional and functional variables will uncover interesting form-meaning patterns. This chapter seeks to fill the gap in the bottom-up and functional description of discourse markers in spoken English and French, with no direct link to interpretations of relative (dis)fluency.

Chapter 6 reports on the distribution and combination of DMs and disfluencies in the subcorpus of interviews, where they both have been fully annotated. This chapter will answer the following question: what can we conclude about the (dis)fluency of DMs on the basis of corpus frequency and their clustering with other (dis)fluency markers in the typology? DMs will first be situated within the disfluency typology by identifying the rate and strength of association between different members. Interpretations of relative fluency and disfluency will then emerge from the converging evidence of corpus frequency and disfluency structure.

Chapter 7 integrates the results from the previous two chapters. The analysis of (dis)fluent clusters takes into consideration the syntactic and pragmatic features of DMs in order to identify more or less fluent DM configurations, with a focus on

their functions. The findings thus obtained build up a tentative scale of (dis)fluency on the basis of independent yet converging evidence.

Lastly, in Chapter 8, the annotations of discourse markers will be combined with a qualitative categorization of repair types strongly inspired by Levelt's (1983) model. This chapter focuses on the signaling role of DMs in sequences of overt repair, in order to identify associations between DM functions and degree of fluency, thus complementing previous corpus-based patterns. This analysis is designed to provide a more direct access to the interpretation of fluency and disfluency, pursuing the same overarching goal to rank DM uses on a scale of fluency.

The main findings of the study will be summarized and discussed in the conclusion, along with suggestions for further research avenues and implications of the present study.

CHAPTER 2

Definitions and corpus-based approaches to fluency and disfluency

The aim of this chapter is to define and discuss the theoretical notions, models and frameworks related to the concepts of fluency and disfluency. Different approaches to both definition and annotation of (dis)fluency will be systematically compared, before introducing the approach adopted in this study. Full review of all existing frameworks is beyond the scope of this chapter, given their number and diversity in fields as varied as second language acquisition, speech pathology or computational linguistics. This chapter rather focuses on works which are (1) relevant to the present study and (2) representative of theoretical differences in the terminology and definition of (dis)fluency, whether holistic or componential, and within the latter whether qualitative, quantitative or both. Furthermore, this work is theoretically embedded within the framework of usage-based linguistics: key notions and general assumptions are outlined along with a discussion of how they were applied to the present research purposes.

The relevance of the distinction between fluency and disfluency in spoken language might be questioned, especially to avoid any prescriptive judgement or generalization: utterances can be produced in non-standard ways (e.g. with repairs and edits) and still be well understood and perceived. Nevertheless, some working definitions can be preliminarily laid out in order to ease the reading process: *fluent* characterizes perceptively unmarked talk, which can be plain, eloquent or creative, albeit not necessarily flawless; *disfluent* applies to major breaks in the speech flow or in the syntax, leading to some sort of disruption; *a disfluency* (or *disfluencies*) refers to an actual occurrence of phenomena such as pauses, repetitions or truncations.

I will start by introducing Levelt's (1983, 1989) seminal model of repair, in order to address terminological issues and defining concepts which are still in use more than thirty years later, up to the present study (see in particular Chapter 8).

2.1 Disfluency or repair? Levelt's legacy

Like other fields in linguistics, fluency studies suffer from a lack of consensus at the level of definition, which is "notoriously difficult" to agree upon (Hasselgren 2002: 147), but also at the lower level of terminology. Research on fluency started with the study of pauses and other "hesitation phenomena" (e.g. Maclay & Osgood 1959; Goldman-Eisler 1968) before being taken up by conversation analysts who soon talked about "repair", as in Schegloff et al. (1977: 381): "An adequate theory of the organization of natural language will need to depict how a natural language handles its intrinsic troubles. Such a theory will, then, need an account of the organization of repair". Despite the rather negative connotation in the "repair" term (suggesting that something is damaged and needs repairing or correction), it has been used quite often in computational linguistics (e.g. Nakatani & Hirschberg 1994) and conversation analysis where it comes from (e.g. Auer 2005; Auer & Pfänder 2007). Although most of recent research now uses the – still connotated – term "disfluencies", the notion of repair remains important and relevant mainly for two reasons: (1) the notion covers different meanings, which need to be disentangled for the sake of clarity; (2) it is often associated with Levelt's (1983, 1989) larger model of speech production, which remains a seminal work in the domain.

In its first sense, *repair* is synonymous with *disfluency* and refers to instances of trouble in the linguistic production. Within this meaning, a further distinction has been made between a wide definition of repair, as in "instances in which an emerging utterance is stopped in some way and is then aborted, recast, continued, or redone" (Fox et al. 1996: 189), and a narrow definition where repair is roughly equivalent to reformulation, leaving out other types of interruptions labeled as "disfluencies".[4] In both cases (wide or narrow definition), repairs correspond to disfluent stretches of talk which may be labeled differently (e.g. filled pause, repetition, substitution, reformulation) depending on the typology.

In their second sense, repairs are synonymous with *reparans* and only correspond to one structural component of a disfluency, namely the last part, where fluency is resumed, that is, "the correct version of what was wrong before" (Levelt 1983: 44). In Levelt's (1983: 44) terminology, a repair (or *reparans*) is combined with a *reparandum* ("item to be repaired"), a moment of interruption ("the point at which the flow of speech is interrupted") and an editing phase (also called *interregnum* e.g. in Shriberg 1994) possibly containing an editing term (typically *uh*, *well*, etc.). This use of the term can still be found in more recent studies which investigate the structure of disfluencies such as Pallaud et al. (2013) or Dutrey et al. (2014).

4. See Section 8.1 for a detailed review of the relation and partial overlap between repair and reformulation.

Figure 2.1, borrowed and simplified from Levelt (1983: 45), illustrates this internal structure and the corresponding terms.

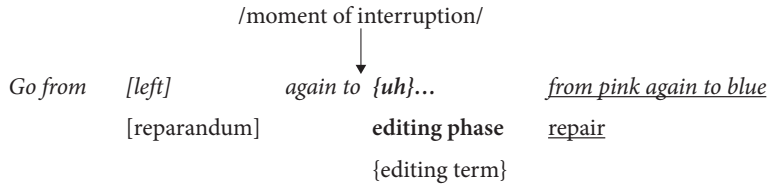


Figure 2.1 Levelt's (1983) terminology

The situation becomes quite confusing when authors (starting with Levelt himself) use "repair" to refer to both meanings at the same time, as in Eklund (2004: 164) who says that the notion of repair entails that "something needs to be corrected [first sense], and that there is a structure to repairs themselves, with a reparandum, and interruption (sometimes an editing term), and a/the repair (or reparans) [second sense]. A repair can include other phenomena, such as repetitions, substitutions, insertions, deletions and so on". In his view, simple elements such as filled pauses or prolongations can be incorporated in repairs but need not be (see also Postma et al.'s (1990) distinction between repairs and disfluencies). In other words, Eklund's (2004) view of repair is polysemous (a type of repair and a structural component) yet narrower than other definitions (cf. Fox et al. 1996).

Levelt (1983, 1989) uses both meanings of "repair" combined with the notions in Figure 2.1 (among others) and includes them in a larger "blueprint" model of speech production which has been re-used (and criticized, e.g. Seyfeddinipur 2006) many times since, thus explaining the fame of the "repair" term. Levelt takes as a starting point the notion of "monitoring", that is, the automatic process of comparing the linguistic output with the intended message, and generating adjustments when necessary. Monitoring is the last "processing componen[t] involved in formulating and repairing" (1983: 47) after the following other steps:

- message construction (ordering messages and ideas);
- formulating (retrieving word forms and phonetic strings);
- articulating (oral output);
- parsing (understanding the intended message from the output);
- monitoring (comparing output with intentions and language standards).

Levelt (1989) relates some of these components to two major cognitive processes, viz. macroplanning and microplanning, respectively dealing with (1) the selection of information relevant to the realization of the communicative intention, and (2) the information structure and style of the utterance. In his view, macroplanning

and microplanning differ in cognitive demands (higher for the former) and alternate in temporal cycles which correspond to stretches of hesitant vs. fluent phases.

Repairs are the results of monitoring, which can target anomalies at any step of the model developed above, and can take two main forms, namely overt vs. covert repairs. The former necessarily involves a change, addition or deletion of morpheme, while the latter merely constitutes an interruption point, such as pausing or repeating the same word with no change (e.g. "I went to to London"). Overt repairs correspond to the narrow definition of repairs presented above, while the wider definition includes both overt and covert repairs. This distinction can be found in many studies under different names, one interesting proposal being Ginzburg et al.'s (2014) "backward-looking" vs. "forward-looking" disfluencies: the former "refers back to an already uttered reparandum" while the latter refers to the "completion of the utterance which is delayed by a filled or unfilled pause or a repetition" (2014: 4).

Levelt's model includes more distinctions and subtypes of repairs, depending on their format (e.g. immediate or delayed) and on their source or motivation (e.g. error or inappropriateness, see Section 8.1.3 for the detailed typology). Levelt also insists on the versatility of repair, stating that "there are many repairs where there is nothing wrong to start with; also many repairs are not correct themselves, sometimes leading to a staggering of additional repairs" (1983: 44). In the case of covert repairs, it is impossible to identify the reason why the speaker interrupted their utterance: since no apparent change occurs in structure or content (as in "I went to to London"), the interruption cannot be reliably interpreted any further than an undefined case of hesitation, regardless of whether the speaker meant to say "Paris" instead of "London", or forgot the name of the capital, or was aiming for a more specific referent like "Greenwich". On the other hand, overt repairs provide more structural cues for their interpretation and analysis, as carried out in Chapter 8 of this book.

To sum up, Levelt's (1983, 1989) model takes scope over many features of repair which he understands in a broad sense, encompassing all the phenomena that will later be referred to as disfluencies. However, this original definition of repairs is not consensual and has tended to disappear from the literature, in spite of the quality of the overall model. Therefore, I will now focus on the concepts of fluency and disfluency, which do not necessarily entail erroneous or corrected language, as will be developed in the following sections. Levelt's (1983) model and the notion of repair will be central to the analyses in Chapter 8, where further details will be provided.

2.2 Holistic definitions of fluency

Many early definitions of fluency consider the concept as a holistic assessment or impression of the general production of a speaker, usually focusing on one aspect of language, although the specific aspect may differ from one definition to another. Holistic approaches do not investigate the components of fluency but instead target conceptually central features, however subjective they may be, in order to describe the global impression of fluency (and not its parts). Three of these central concepts emerge from the literature, namely automaticity, flow, and efficiency, which are all briefly reviewed in the following.

The origins of fluency research in pausology and second language acquisition explain why a number of authors associate fluency with automaticity and effortlessness, as in the following definitions: "smooth, rapid, effortless use of language" (Crystal 1987: 421); "automatic procedural skill" (Schmidt 1992: 358); "speed and effortlessness" (Chambers 1997: 535). No explicit reference is made to the content or structure of discourse, but mainly to the underlying cognitive processing which concerns all aspects of language at once, as stressed by Levelt (1989: 2) who considers production automaticity as "a main condition for the generation of uninterrupted fluent speech". This first group of holistic definitions is therefore strongly cognitive and speaker-oriented.

The second notion, which is also present is some of the definitions above, is that of flow or rhythm: according to Ejzenberg (2000: 287) for instance, fluency is "a component of overall language ability or proficiency that indicates the degree to which speech is articulated smoothly and continuously without any 'unnatural' breakdowns in flow". Similarly, Fiksdal (2000: 128) talks about "steady tempo": this phrasing emphasizes the temporal, almost musical character of idealized fluent speech, and reflects a focus on temporal variables (speech rate, pause duration). In a more syntactic sense, flow is also referred to in a negative way (i.e. absence of flow) in French works such as Blanche-Benveniste et al. (1990) who define disfluency as breaking the syntagmatic unfolding of the utterance, or Dister (2007) who uses the term "paradigmatic piling up" ("*entassement paradigmatique*" in the French original). This type of definition is appealing for its metaphorical and descriptive power, which also relates to the temporality of spoken delivery: while speech does have a regular cyclic rhythm which contributes to the ideal fluent melody, it is however not a continuous rhythm but much rather one of alternation between sound and silence. Definitions such as "steady tempo" might therefore not be accurate in this regard.

The last group focuses on efficiency and differs quite strongly from the previous two in that it includes an idea of relativity, either from a distributional viewpoint or a cognitive one: the issue is no longer to produce many disfluencies or none at all,

but to remain efficient despite the production of disfluencies. This line of reasoning is mostly found in studies on second language (henceforth L2) acquisition: for instance, Brumfit (1984: 57) defines fluency as "the maximally effective operation of the language system so far acquired by the students". Denke (2009: 15) goes one step further by taking into consideration not only frequency but also position and use of disfluencies: "being fluent in a language does not mean that there is a total lack of, e.g., hesitation, but rather that there are differences to be found between native and non-native speakers regarding how often and where it occurs". This functional view of disfluencies as being strategically distributed in the speech string seems promising and realistic beyond the realm of L2 fluency.

There is obviously some overlap between all these definitions, some of them including more than one of the aspects discussed above. Lennon (2000: 26) offers yet another example of a synthetic account which covers most or all of these notions: "the rapid, smooth, accurate, lucid, and efficient translation of thought or communicative intention into language under the temporal constraints of on-line processing". Such broad definitions are useful to capture the full scope of what fluency entails, especially when each element of the definition can be traced back to measurable observations, which is not always the case.

2.3 Componential approaches to fluency and disfluency

Despite their insightful resort to aspects of language such as flow or efficiency, which are indeed central to spoken fluency, holistic definitions have long been criticized in the literature, as early as Hieke (1985: 136) who regrets "their essentially subjective nature", which is why I will now focus on componential approaches. These will be subcategorized into qualitative and quantitative descriptions, where qualitative corresponds to features which cannot be measured but are rather perceived as a whole, as opposed to quantitative observations of discrete phenomena. These terms do not fully map with the holistic-componential distinction, which rather refers to the methodological approach, either combining non-defined variables into a global impression (holistic), or investigating specific features separately (componential).

2.3.1 Qualitative components of perception

If the literature on fluency was ranked on a scale with holistic approaches at one end of the continuum and componential approaches at the other, the works discussed in this section would be somewhat intermediary: the following definitions are componential, in that they acknowledge distinct groups of phenomena within fluency, yet

qualitative because these phenomena are not (all) measurable quantitatively. I will focus in particular on two authors, namely Fillmore (1979, 2000) and Segalowitz (2010), who do not provide full typologies of disfluencies but decompose their definition in different perceptive variables.

Fillmore (2000: 51) identifies four dimensions of fluency: (1) the "ability to talk at length with few pauses, the ability to fill time with talk", that is, a notion of rhythm possibly measured by temporal variables; (2) the "ability to talk in coherent, reasoned, and 'semantically dense' sentences", in other words fluent speakers "tend not to fill discourse with lots of semantically empty material"; (3) the "ability to have appropriate things to say in a wide range of contexts" (which relates to general world knowledge as well as Grice's (1957) maxim of relevance) and (4) the "ability some people have to be creative and imaginative in their language use, to express their ideas in novel ways, to pun, to make up jokes, to attend to the sound independently of the sense, to vary styles, to create and build on metaphors, and so on", in other words the aesthetics of one's language. He concludes that the ideal speaker should master all these aspects. Fillmore (2000) himself acknowledges that it is challenging to find operational measures matching this definition other than intuitive rankings. He fully embraces the difficulty of fluency assessment yet argues that this modularity in the definition is true to the many ways in which speakers can be fluent, depending on their vocabulary, creativity or general world knowledge. It seems that the third (and fourth, to a lesser extent) dimension(s) would be especially hard to measure.

Segalowitz (2010) proposes a three-fold definition of cognitive, utterance and perceived fluency: cognitive fluency is the "ability to efficiently mobilize and integrate the underlying cognitive processes responsible for producing utterances with the characteristics that they have" (hardly accessible to the analyst); utterance fluency corresponds to the actual observable features of an utterance such as temporal variables and repair characteristics; perceived fluency is the synthesis of the other two and concerns "the inferences listeners make about a speaker's cognitive fluency based on their perception of utterance fluency" (2010: 48). Only utterance fluency is fully measurable. Segalowitz (2010) claims that only by extending the definition of fluency to non-audible (i.e. cognitive) processes of both production and perception can we grasp the full nature of fluency. However, he acknowledges the methodological difficulty of interpreting data of so many different kinds, and therefore calls for multidisciplinary approaches to the issue. Segalowitz's (2010) definition certainly strikes one as very broad, combining aspects of both speaking and listening in a complex but partitioned model.

2.3.2 Quantitative components of production

In a corpus-based perspective, broad definitions such as the ones discussed so far are most useful when they are combined with a more fine-grained analytical grid, mapping each element of the definition with a discrete, measurable variable, in an operational typology directly applicable to corpus data. Such typologies of components of (dis)fluency are abundant in the literature and reflect a diversity of theoretical approaches and research agendas in many languages and data types. In the following, a selection of proposals are reviewed and grouped according to their underlying conception of fluency and disfluency, namely *disfluencies as removable errors* or *disfluencies as functionally ambivalent devices*.

2.3.2.1 *Disfluencies as removable errors*

The (chronologically) first group of annotation models adopts a rather negative perspective on the elements in their typology, which is reflected by the connotated use of the term "disfluencies": the phenomena under consideration need to be identified in order to be later removed for a variety of applied purposes such as automatic detection, assessment of proficiency or summarization. These works thus target rather disruptive features of spoken language and are influenced by their computational orientation. Despite the great number of existing proposals in this line of investigation, it is possible to identify commonalities, especially since many of the later references take up previous original typologies. Four of these seminal references will be discussed here: Shriberg (1994), Meteer et al. (1995), Strassel (2003) and Besser & Alexandersson (2007).

The first and major reference is Shriberg (1994), whose influence grew beyond her original framework to fluency research in general. Shriberg aimed at finding regularities in disfluencies in order to build (the first steps of) an encompassing theory. Despite this general purpose, Shriberg (1994: 1) states a number of restrictions to the scope of her typology: "The DFs [disfluencies] considered are cases in which a contiguous stretch of linguistic material must be deleted to arrive at the sequence the speaker 'intended', likely the one that would be uttered upon a request for repetition".

More specifically, this approach to disfluencies as removable elements excludes unfilled pauses, uncorrected prosodic errors, coughing or discourse markers (such as *well*, *like*) on the grounds that "they are arguably part of the speaker's intended utterance" (1994: 1). What it does include, however, are: repetitions, substitutions, insertions, deletions, filled pauses, explicit editing terms, some uses of discourse markers, coordinating conjunctions, word fragments, misarticulations, contractions

and syntactic incompletion.[5] Shriberg's model also provides a notation system for the different parts or "regions" of a disfluency, taking up Levelt's (1983) terminology: *reparandum*, interruption point, *interregnum*, repair or *reparans*.

Her corpus analysis showed that (1) disfluency rate is dependent on utterance length, an effect which is itself dependent on the corpus or interaction settings; (2) disfluencies mostly affect utterance-initial words; (3) there is a co-occurrence or attraction effect between initial and medial disfluencies in single utterances.

The main shortcoming of this proposal, with respect to the specific purpose of the present research, is its restriction in scope, which seems slightly contradictory with Shriberg's (1994) own endeavor to strive towards theoretical neutrality: deciding what is part of the speaker's original intention and building a theory on this basis seems a rather strong not-so-neutral position. Overall, this typology paved the way for later annotation models – including the one used in the present work – on many levels, namely classification of disfluencies in "orthogonal" (i.e. non-overlapping) categories, visualization of the internal structure of disfluencies and efficiency of the labeling system. In particular, it was taken up by Eklund (2004) who pursued a very similar endeavor in Swedish where he found comparable results (especially the correlation between frequency of disfluencies and utterance length) and an overall frequency of 6.4 disfluencies per 100 words. This rate is corroborated by Bortfeld et al. (2001) who carried out a sociolinguistic study of disfluencies in conversation, adopting a similar typology (although not explicitly related to Shriberg's (1994) original): they found a 5.97% rate of disfluencies, affected by planning demands (unfamiliar topic, longer turns) but not by speakers' age.

Another widespread framework is that of the Switchboard corpus of telephone conversations and the annotation model by Meteer et al. (1995). In the perspective of "cleaning" transcriptions, they provide a three-step annotation, covering (1) "non-sentence elements" (filled pause, explicit editing term, discourse marker, coordinating conjunction, aside), (2) "slash-units" (tagged as complete or incomplete in the case of mid-utterance interruptions) and (3) "restarts", directly based on Shriberg (1994) and including repetition, substitution, deletion and complex restarts. The Switchboard corpus and its disfluency annotation has been used by several authors focusing on different aspects of fluency, for instance Clark & Wasow (1998) on repetitions, which they consider to function as "initial commitments" used by the speakers to comply with the "temporal imperative" or planning

5. Shriberg (1994) includes discourse markers in her theoretical typology but restricts their identification in corpus to cases where they occur within another disfluency. Like many authors, she excludes them from most of her analyses because of the uncertainty of their "intentionality".

pressure. In that sense, Clark & Wasow (1998) are closer to functional accounts of disfluencies (see next section) rather than the disfluency-as-removable-error approach. Meteer et al.'s (1995) disfluency typology was also used by Zechner (2001), who was the first to tackle summarization of spoken dialogues. Mieskes & Strube (2008) then started from Zechner's (2001) version of the Switchboard annotation to train an automatic disfluency classifier in multi-party dialogues.

Following the same research agenda, viz. cleaning transcriptions of natural speech (here phone calls and broadcast news in English), Strassel (2003) developed SimpleMDE, a specification of "metadata" (her term) which covers "fillers" (filled pauses, discourse markers, explicit editing terms, asides and parentheticals), "edit disfluencies" (repetitions, revisions, restarts, complex disfluencies) and "semantic units" (defined as complete ideas). While the typology is fairly similar to others discussed above, Strassel's (2003) guidelines stand out as particularly operational and prescriptive. She dedicates a specific section to the complex disambiguation of discourse markers such as *like* or *so* and allows for a "difficult decision" label, thus prioritizing reliability over exhaustivity (leaving out cases of hesitation and complex structures). This model was taken up by Dutrey et al. (2014) in the perspective of automatic detection of disfluencies in French.

The last proposal to be discussed in this section is Besser & Alexandersson (2007), who claim to be more exhaustive than both Shriberg (1994) and Zechner (2001) with whom they share the same summarization purpose. Working with both native and non-native data (international business meetings in English, AMI corpus), Besser & Alexandersson (2007: 182) focus on "syntactic and grammatical errors according to standard syntax and grammar", that is, "phenomena that actually lead to the interruption of the syntactic or grammatical fluency of an utterance", thus leaving out stylistic or semantic considerations. They identify three groups of disfluencies based on their surface structure: "uncorrected" (mistake, omission, wrong order), "deletable" (hesitation, stuttering, disruption, slip of the tongue, discourse marker, explicit editing term) and "revisions" (deletion, insertion, repetition, replacement, restart, other). It does appear that this typology makes finer distinctions than others, all the while remaining reliable: the authors show outstanding inter-annotator agreement, with a Kappa score of $\kappa = 0.924$. Their approach is also much more normative than others in this section, which is probably due to the presence of non-native speakers in their data: the "uncorrected" category in particular reflects their view on grammaticality, which is absent from all other frameworks discussed so far.

To sum up, most of these typologies share a componential approach and are more or less rooted in Shriberg's (1994) legacy, in addition to their common view of disfluencies as rather disruptive and removable phenomena. Some differences

regarding the number and types of disfluencies included, as well as technical choices, remain due to the sometimes divergent research agendas (e.g. theory-building vs. automatic summarization).

2.3.2.2 *The functional ambivalence of disfluencies*

Pallaud et al.'s (2013) typology of "self-interruptions" in French shows a more nuanced, if not clearly positive view of disfluencies by acknowledging their functional ambivalence, from disfluent to more helpful and strategic uses:

> These interruptions and reorganizations do not seem, in the great majority of cases, to hurt the unfolding of the speech segment but rather to impose a rhythm that is specific to oral utterances. What is more, it seems that this oral-specific rhythm creates, on the contrary, the conditions for an optimal interaction insofar as, by triggering a reorganization of the utterance, it reduces the informational load of the utterance for the listener. (2013: 1, my translation)

As a result of this broader scope on non-problematic disfluencies, Pallaud and colleagues include unfilled pauses, in addition to the common core of disfluencies shared with other frameworks, but exclude considerations of grammaticality such as the "uncorrected" category in Besser & Alexandersson (2007). Apart from this major theoretical difference, the annotation system is fairly similar, although perhaps more oriented towards syntax and segmentation: disfluencies are annotated according to their structure in three parts, namely *reparandum*, *interregnum* and *reparans* (cf. Levelt 1983; Shriberg 1994), and further distinguished according to the grammatical class of the item affected by the interruption (word, determiner, phrase, etc.). The categories identified by Pallaud et al. (2013) are quite different from the majority of annotation frameworks: they do not refer to "repetition" or "substitution" but rather describe the type of syntactic effect triggered by the interruption. This different perspective on disfluency annotation takes up some of the formal variables identified by Levelt (1983) regarding the structure of repairs, such as way of restarting or type of unit at the moment of interruption. Although potentially more fine-grained than others, this model presents the disadvantage of having to combine multiple labels in lengthy, opaque tags which may render the annotation process cumbersome.

It remains that the more encompassing, functionally ambivalent view of disfluencies adopted by Pallaud and her colleagues is highly compatible with a wealth of experimental evidence suggesting that not all disfluencies are disruptive. In fact, only a handful of studies show negative effects of disfluencies. Fox Tree (2001) for example found that utterance-medial false-starts cause processing trouble especially when they co-occur with discourse markers. MacGregor et al.'s (2009) study

on repetitions provides more nuanced results, showing that repetitions do not have a detrimental effect but rather no effect at all on the processing of subsequent words. On the other hand, positive effects of disfluencies have been shown to concern both the speaker and the hearer and include reference resolution (Arnold et al. 2003), memory enhancing (Liu & Fox Tree 2012; Bosker et al. 2014) or expectation triggering (Barr & Seyfeddinipur 2010; Corley 2010).

Somewhat in-between these two extremes, Brennan & Schober (2001: 292) develop the claim of a "disfluency advantage" whereby "there is information in disfluencies that partially compensates for any disruption in processing". In a response-time experiment, they show in particular that disfluent utterances trigger faster responses than fluent ones, and that the presence of fillers (*uh*) in particular reduces erroneous responses, which they explain by the extra time a filler allows for processing. Brennan & Schober (2001: 295) conclude that "fluency is still desirable from a listener's perspective" but convincingly show some compensating effects which mitigate the opposition between the two accounts discussed so far, namely *disfluencies as removable errors* vs. *disfluencies as functionally ambivalent devices*.

Focusing on English filled pauses, Clark & Fox Tree (2002) summarize this divide in the literature by calling the negative approach *filler-as-symptom* (i.e. disfluencies are involuntary side-effects of a production problem), as opposed to the more positive *filler-as-signal* view (i.e. disfluencies are motivated by some kind of interactional intention, for instance to hold the floor). Other authors (e.g. Auer 2005: 100) even suggest that disfluencies are not "a remedial device correcting some deficiency […] but rather as part of the solution to this problem". The definition and approach taken in this book adopts a similar functionally ambivalent perspective.

2.3.3 Götz's qualitative-quantitative approach

Götz's (2013) comprehensive approach to both production and perception of English speech offers a recent milestone in the study of first- (L1) and second-language (L2) fluency. Her proposal ties together many aspects which are usually studied individually in other frameworks and shows a high degree of integration both at the theoretical and methodological levels. In particular, she combines a holistic definition with a componential typology, which is itself structured around both quantitative and qualitative variables or dimensions of fluency, respectively investigated through corpus analysis and experimentation.[6]

6. Götz (2013) refers to her own model as "holistic", which she uses as a synonym for comprehensive or integrated. According to the present definition of this term, her three-fold typology could be described as componential.

Her definition of native (L1) fluency is consensual and synthetic, largely borrowing from Lennon (2000) and other holistic definitions (Section 2.2): "speak smoothly, appropriately, correctly, with ease and effortlessness" (2013: 1). She further distinguishes production fluency (i.e. the aspects of speech that "enhance the speaker's ease and effortlessness in their speech production", 2013: 2) from perceptive fluency (i.e. the elements that establish the perception of a speaker's fluency). In Götz's view, production and perception are both associated to the speaker's perspective, and therefore potentially more accessible to the analyst. This stands in sharp contrast with other approaches such as Segalowitz's (2010) who also includes production and perception but reserves the latter to the listener's experience (cf. "the inferences listeners make about a speaker's cognitive fluency based on their perception of utterance fluency", 2010: 48), which is not directly observable.

One of Götz's (2013) main contributions to the present approach is terminological: she suggests the concept of "fluenceme" to refer to "an abstract and idealized feature of speech that contributes to the production or perception of fluency, whatever its concrete realization may be" (2013: 8). This term has a less negative connotation than "disfluencies" and thus well-suited to describe their functional ambivalence: the *–eme* suffix expresses the heterogeneity and potential of these elements to be used either fluently or disfluently. This term will henceforth be used in the remainder of this book.

Götz (2013) identifies three types of fluencemes: fluencemes of production, which can be related to issues of planning pressure; perceptive fluencemes, which usually attract the listener's attention; nonverbal fluencemes, which contribute to both production and perception depending on their functions. The elements under consideration cover almost every aspect of language (prosody, lexicon, discourse, pragmatics, nonverbal communication). This typology does not fully map with Segalowitz's (2010) own tripartite definition of fluency (cognitive, utterance, perceived): in Segalowitz's terms, all the features analyzed by Götz are produced by the speaker and therefore belong to "utterance fluency"; productive fluencemes could be considered to depend on "cognitive fluency" (but also gestures or sentence structure); perceived fluency builds on the information from all three components (productive, perceptive and nonverbal). While Segalowitz (2010) targets a cognitively valid – albeit abstract – model, Götz (2013) favors a speaker-based approach which only includes observable features of communication, however (non-)pervasive they may be.

It remains to be empirically tested whether her categorization is "orthogonal", to take up Shriberg's (1994) term, or whether some fluencemes could be considered to function both at the productive and perceptive levels: filled pauses, for one, have been repeatedly shown to affect speech perception and processing (e.g. Bosker

et al. 2014; Watanabe et al. 2008), while intonation could be strongly affected by difficulties of production.

To sum up, Götz's proposal is valuable for many reasons: she combines definition and typology in a single model; she encompasses quantitative variables of production together with more qualitative variables of perception; her model is compatible for both native and non-native speakers; her mixed-method approach is innovative and powerful; the term "fluenceme" succeeds in capturing the ambivalent nature and function of the phenomenon, as opposed to the more pejorative term "disfluencies". One limitation which prevents direct use of this typology is the lack of technical guidelines from the perspective of corpus annotation: Götz extracted a selection of features (semi-)automatically without directly annotating the data, a methodological choice which is time-saving but perhaps questionable from the point of view of replicability, exhaustivity and granularity. All in all, Götz (2013) provides an applied perspective and motivation to the present research (namely the study of L1 fluency to better understand L2 fluency), in addition to her theoretical, methodological and terminological contributions.

2.4 Synthesis: Definition adopted in this work

Against this rich backdrop, the approach adopted here aims at inclusiveness and combines elements from the different contributions reviewed so far. More specifically, my definition takes up (1) the overt-covert distinction from Levelt's (1983) model of repair, (2) the notions of flow and efficiency from the holistic definitions, (3) the interest for quantitative measures of production and their functional ambivalence from componential approaches to disfluency and (4) the notion of fluenceme in particular from Götz (2013). I therefore consider fluencemes to be discrete devices which function as signals of cognitive processes of speech production and perception. This definition follows the overarching claim that fluencemes (as a whole category and as individual members) are not necessarily problematic but rather reflect some cooperative (even listener-oriented) search for the optimal utterance. Fluency is the result of these signaling strategies: it does not equate to absence of fluencemes, but rather efficient (sometimes even creative) use of them, where efficiency is defined inter-subjectively and in context by the co-participants. Disfluency, on the other hand, corresponds to the perceptive effect of more symptomatic uses of fluencemes, through which the speaker is expressing some sort of production trouble, leading to major disruptions in the prosody and/or syntax.

This functionally-oriented definition is complemented with a typology of fluencemes which includes 10 primary phenomena ("fluencemes"), three secondary phenomena ("related elements") and three "diacritics", as can be seen in Table 2.1.

Table 2.1 Typology of fluencemes (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	explicit 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"
	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"
	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"

These elements were manually identified in the corpus, following operational definitions provided in Crible et al. (2016) (see Chapter 4). The typology covers both ambivalent devices (e.g. pauses, discourse markers) and others more conceptually related to disruptiveness and disfluency (e.g. false-starts, misarticulation), in line with the concept of functional ambivalence at the core of the present definition of (dis)fluency.

2.5 A usage-based account of (dis)fluency

Definitions and typologies, however broad they may be, are limited in their explanatory power insofar as they target general categories and phenomena, while language production and perception deal with successions of particular instantiations. It is a challenge specific to corpus linguistics to be able to derive theoretical models from a closed set of observations, since authentic data is subject to variation and factors which are not all under the analyst's control and cannot be accounted for in one corpus. Experimental studies usually work with an even more restricted lens (few stimuli, few participants) but benefit from a high degree of control on internal and external factors, which allows them to draw robust generalizations.

In the case of (dis)fluency, the limitations of corpus-based research to generate evaluations of (dis)fluency for each occurrence of fluenceme is not only due to the complexity of the phenomenon but more fundamentally to methodological monism (i.e. resort to a unique method) and to the relative absence of theoretical background against which observed patterns can be interpreted. The former will not be addressed in this research, mainly because, as mentioned earlier, experimental work imposes high restrictions in the dataset, which somewhat counters the present endeavor to study full categories (of fluencemes in general and of discourse markers in particular). The latter limitation (i.e. lack of theoretical background), however, will be partly overcome by the systematic reference to the framework of usage-based linguistics which provides relevant notions and methods to build the targeted model of (dis)fluency.

2.5.1 Key notions in usage-based linguistics

The usage-based approach emerged in the 1980s from functional and cognitive linguistics striving to bridge the gap between *langue* (grammar) and *parole* (usage). Authors such as Bybee (1985) or Hopper (1987) started seeing language as a dynamic system whereby units emerge from general cognitive processes (e.g. categorization, analogy) which are not only relevant for the linguistic system but also for other faculties such as vision or thought. Kemmer & Barlow (2000) offer a systematic review of the characteristics of usage-based models of language, of which I summarize the main points: both linguistic structures and linguistic theory are based on observations of repeated instances of language use; frequency is an important factor in cognitive entrenchment; variation and change should be accounted for; context has a crucial role in language processing and can even be integrated as part of the semantic-pragmatic meaning of an expression, thus considering language as context-bound and underspecified. These tenets of the usage-based framework, especially the central roles of frequency and context, lie at the core of the present study of (dis)fluency.

2.5.2 From schemas to sequences of fluencemes

In particular, the notion of schema, as developed in the usage-based framework of Cognitive Grammar (Langacker 1987, 1988), offers some potential for the analysis of fluency. Schemas emerge from repeated exposure to particular usage events which are then abstracted from their context of occurrence and become progressively "entrenched" as cognitive routines:

> The occurrence of psychological events leaves some kind of trace that facilitates their re-occurrence. Through repetition, even a highly complex event can coalesce into a well-rehearsed routine that is easily elicited and reliably executed.
>
> (Langacker 2000: 3)

Usage-based schemas have been identified and studied at different levels of language (phonology, syntax, discourse). For a particular unit or structure to be considered "schematic", it needs to meet some requirements such as high frequency and should integrate into a network with particular instantiations and other related schemas.

In the perspective of disfluency analysis, the notion of schema cannot be uniformly applied to all observed occurrences but should rather be reserved for recurrent patterns of combination. For this reason, I will rather use the term "sequence" until converging evidence reveal whether these sequences constitute schemas as well. Sequences refer to the co-occurrence of several fluencemes on the syntagmatic axis of the utterance. For example, a particular instance of filled pause followed by a word truncation (e.g. *the uh h- house*), an insertion embedded within a repetition (e.g. *I said when he asked me I said*), or two discourse markers in a row (e.g. *well I mean*) would constitute respective sequences. Sequences are not restricted in minimal or maximal length nor in content; even a single fluenceme occurring in isolation will be referred to as a sequence. This rather atypical use of the term allows me to refer to any stretch of talk consisting in or affected by a fluenceme regardless of its size, thus providing a constant unit of analysis.

The similarity between sequences and schemas mainly relies on their ability to build from particular instances into more and more abstract categories. For example, in an utterance such as "the uh you know the house is big", we can observe three fluencemes: (1) the identical repetition of "the", (2) the filled pause "uh" and (3) the discourse marker "you know". They instantiate the sequence "the uh you know the", which can be abstracted into the pattern [repetition + filled pause + discourse marker] but also [utterance-initial determiner repetition + embedded fluencemes], [repetition + *uh* + discourse marker], [non-isolated repetition], etc., depending on the degree of abstraction or granularity necessary or relevant for the analysis.

This interest in sequences is not only motivated by the conceptual similarity with usage-based schemas, but also by corpus-based and experimental evidence showing that fluencemes are more often combined than isolated. Grosjean & Deschamps (1975: 176) compared the clustering tendency of filled pauses in English and French and found that they often combine with other hesitations, especially silent pauses (68.29% of filled pauses in English against 47.26% in French), a tendency which they interpret as the speaker's need to stall for encoding purposes. They also found that repetitions are more often preceded by silent pauses in French than in English

(49% vs. 27%), which might be correlated to the longer size of French repetitions. Duez (1991) found that, on average, 60% of filled pauses in her French corpus are combined with another marker across different registers. In Candéa's (2000) corpus of French child language, 35% of lengthenings are isolated, against only 12% for repetitions of function-words (i.e. non-lexical such as prepositions).

To conclude, the present study will take sequences as the basic unit of analysis, following usage-based and other evidence of the importance of combinatory patterns, as opposed to a fluenceme-by-fluenceme approach which would be overlooking the actual context of occurrence of the tokens. I would argue that reports on the frequency and use of fluencemes that do not systematically account for their combination (or not) with others offer a distorted picture of the data. This strong position puts forward the hypothesis that a filled pause alone is not used and perceived in the same way as a filled pause clustered with a discourse marker, for instance.

2.5.3 Variation in context(s)

One of the key characteristics of the usage-based framework is that it takes into account the "crucial role of context in the operation of the linguistic system" (Kemmer & Barlow 2000: xxi). Thus, linguistic patterns undergo the influence of linguistic co-text and extra-linguistic context. Following this claim, items and structures integrate some information from their local (i.e. linguistic co-text) and global (i.e. communicative context) environment in their meaning and use.[7] As a result, the same pattern in different contexts could be used and perceived differently. This is especially true for expressions which show little or no lexical or propositional content and instead rely on pragmatic interpretation to resolve their ambiguity and underspecification. Fluencemes (and especially discourse markers among them) match this description. For this reason, fluencemes should be studied in relation to their co-text and context, that is, across registers.

Register variation can be seen as an overarching, more theoretical factor over co-textual variation, "filtering the choice of linguistic features from the language system" (Neumann 2014: 36). In a functional perspective on context, speakers' linguistic options are affected by recurring and conventionalized contextual configurations (Halliday & Hasan 1989). For instance, lengthy pauses are typical features of news broadcasts where they mark discourse structure and ease information

7. Cutting (2008) refers to these types of information as "co-textual context" and "situational context", respectively.

processing. However, in interactive settings, the longer the pause, the higher the risk of losing the floor and being otherwise perceived as hesitant.

Several studies have investigated the impact of register on the distribution of fluencemes, starting with Broen & Siegel (1972) who experimentally compared the production of participants in elicited tasks, namely television broadcasting, talking in front of an audience, conversing with the experimenter or speaking alone. The authors found a discrepancy between the participants' rating of their own fluency and their actual production: "in casual situations where speech is of no special importance, adults do not monitor their speech very carefully. They are neither especially aware of their disfluencies nor concerned to control them" (1972: 229). Broen & Siegel (1972: 229–230) conclude that "it is not the situation which induced greater or lesser disfluency. It is rather the subject's evaluation of the requirements of that situation which is crucial". Their seminal study therefore suggests that intermediary registers such as interviews or professional encounters can be expected to present a substantial frequency of fluencemes given the heightened attention of speakers to notice and correct their errors or imperfect structures, which would lead to an increase in interruptions and reformulations. This idea was also put forward by Halliday (1987: 68), who claimed that disfluent phenomena are more characteristic of "self-conscious, closely self-monitored speech" such as academic seminars than casual conversation.

Finally, I would like to extend the notion of context in order to incorporate crosslinguistic variation as an external factor impacting the distribution of fluencemes. Register and language comparison can hardly be carried out independently, as advocated by Neumann (2014: 40) who argues that register is "crucial as a component organizing usage-based contrastive studies" since it ensures comparability between the linguistic forms and uses investigated and between the data types. While fluencemes have been identified in many different languages such as English (Shriberg 1994), French (Pallaud et al. 2013), Swedish (Eklund 2004) or Japanese (Watanabe et al. 2008), very few studies have carried out large-scale crosslinguistic analyses of the full typology of fluencemes. This gap in the literature might be due to the lack of "universal" typologies, most proposals being language-specific, with the exceptions of Grosjean & Deschamps (1975), focusing on temporal variables, and Eklund & Shriberg (1998), who seem to have merged two pre-existing language-specific typologies. A number of studies have focused on specific types of fluencemes in several languages: in particular, filled pauses have been investigated crosslinguistically (e.g. Zhao & Jurafsky 2005 for the English-Mandarin pair; Crible et al. 2017 in English-French), revealing a great variety of forms, from vocalizations (English *uh*, French *euh*) to demonstratives (Spanish *este*, Japanese *eeto*) (see Clark & Fox Tree 2002). Discourse markers also benefit from a wealth of contrastive research which will be discussed in Chapter 3.

Overall, this state of the art seems to call for more research tackling the cross-linguistic and register variation of a broader range of fluencemes in an integrated approach. The present study will therefore pursue such an ambition by investigating three types of contexts: from different language systems (contrasting English and French) to registers within one system (e.g. conversation vs. news broadcast) to specific combinations and syntagmatic behaviors within and across particular texts.

2.5.4 Accessing fluency through frequency

Fluency is often defined as the impression of automaticity and effortlessness or, in other words, the ease of processing from the speaker's and listener's perspective. Langacker (1987 and onwards) associates such ease of processing with the degree of entrenchment of the particular unit at stake, given that highly entrenched units are more rapidly produced and retrieved. Entrenchment is itself a function of the frequency of the unit, since it is only through repetition that schemas are abstracted from their instantiations and shared in a community (e.g. Bybee 2006). At this stage, an over-simplistic conclusion would state that frequency creates entrenchment which facilitates processing and therefore contributes to fluency.

In this fluency-as-frequency view, frequent sequences are expected to be less cognitively demanding for both production and comprehension and trigger limited hesitations given their high accessibility for the participants. On the other hand, rare patterns would be expected to strike the listener as less automatic in production and unexpected in comprehension, especially if register variation is taken into consideration: a particular sequence can be relatively frequent in one register and rare or absent in another, thus rendering its occurrence in the latter context all the more surprising, out of place and potentially disruptive. This line of reasoning is particularly valid for fluencemes which are not typical of formal settings: for instance, the discourse marker *you know* might be perceived as more marked (possibly more disfluent) in a news broadcast than in a casual conversation.

Many authors (e.g. Chafe 1992; Schönefeld 1999; Gries & Stefanowitsch 2006) have advocated the compatibility between corpus linguistics and cognitive theory. Relating language and mind has been particularly promoted by Schmid (2000: 39) and his "from-corpus-to-cognition principle", whereby "frequency in text instantiates entrenchment in the cognitive system". This claim relies on the assumption that linguistic and cognitive categorization is exemplar-based (i.e. starts from concrete tokens of experience), a proposal which is highly compatible with the usage-based framework, as put forward by Diessel & Hilpert (2016: 3):

> If we think of grammar as a network of symbolic units, frequency does not only strengthen the cognitive representations of linguistic elements in memory (as suggested by exemplar theory), but also reinforces the associative connections between them. Other things being equal, the more often linguistic elements occur together in language use, the stronger is the associative bond between them in memory.[8]

This principle of cognitive corpus linguistics is especially interesting if we consider not only textual frequency (i.e. absolute frequency in a given corpus) but also "conceptual frequency", a distinction proposed by Hoffmann (2004) who defines the latter as the relative frequency of an item with respect to all its paradigmatic competitors. The study of individual phenomena in isolation from other members of their category (for instance, extracting only filled pauses or certain types of discourse markers and not the other fluencemes in the typology) would overlook the inter-relation between members and provide an incomplete picture of the broader phenomenon.

One limitation of the fluency-as-frequency approach concerns the impact of high frequency on perceptive impression forming. It seems that, in our daily experience as speakers and listeners, we tend to notice the pervasive presence of "tics" at a certain level of frequency, after which they are perceived as excessive and reflect poorly on the speaker's fluency (cf. the example of the American senator in Chapter 1). Wagner & Hesson (2014: 652) have also shown that frequency of marked language, that is, nonstandard or containing unexpected forms, influences listeners' impressions of the speaker through what they call a quantitatively sensitive "sociolinguistic monitor". Although the authors do not explicitly relate their study to the concept of (dis)fluency, it is clear that perception can be negatively affected by the high frequency of (some uses of) linguistic phenomena such as fluencemes.

To conclude, if fluency, through automaticity and ease of processing, can be related to entrenchment, then it can reasonably be expected to have some sort of relation with high frequency in use. Frequency information will therefore be treated as one factor (among others) of fluency in order to uncover the extent to which rare and frequent sequences can be ranked on a cognitive-functional scale of (dis)fluency.

---

8. This view of language reinforces the choice of sequences of fluencemes as the basic unit of analysis, as opposed to individual fluencemes.

2.6 Summary and hypotheses

In this chapter, the present definition and approach to (dis)fluency, centered around the notion of functional ambivalence, was developed in relation to previous research on fluency and disfluency and to the usage-based framework. It was made clear that sequences of fluencemes constitute the first level of analysis in this study, investigating in particular their combinatory patterns and contextual variation (across languages and registers). The aim is to uncover frequency-based tendencies which could be tentatively related to different ends on the fluency-disfluency scale.

At this level of analysis, a number of hypotheses emerge from each of the three notions borrowed from the usage-based framework, namely combination, variation and frequency. The first hypothesis concerns the clustering or combination of fluencemes and aims at testing whether fluencemes in general occur more frequently alone or combined with other members of the typology. Evidence from the literature tends to suggest that fluencemes do occur more frequently in clusters than in isolation.

Regarding variation, types and sequences of fluencemes which are specific to informal registers (unplanned interactive dialogues) and rare or absent from more formal registers will be considered as typically and relatively "disfluent", and vice versa (sequences specific to formal registers are relatively "fluent"). Fluencemes showing no significant difference across registers are expected to be more ambivalent, and in need of further investigation with additional sources of information. Furthermore, the situation might be more complex than a planned-unplanned divide, with the special status of speaking tasks at a mid-level of complexity. The combination of no or little preparation on the one hand with a heightened attention for self-monitoring on the other, in registers such as interviews or professional meetings, could lead to an increase in fluencemes. In casual and unplanned situations, on the contrary, fluencemes might be equally frequent but less marked and generally unnoticed. Therefore, I expect intermediary registers to be more similar to the unplanned settings in overall rate, if not also in terms of type distribution and clustering tendencies

It should be noted that this interpretation of the fluency of sequences can only be (1) relative to other sequences extracted from the data and (2) generalized, that is, not applicable for each occurrence in the corpus, given that such an ambition would require perceptive ratings or other experimental validation, which is impractical and irrelevant for the present corpus-based study. This line of investigation combines quantitative, qualitative and contextual evidence in the forms of frequency data and multivariate modeling, cognitive-functional interpretation of the patterns at the conceptual level and register expectations based on psycholinguistic research.

No specific evidence in the literature motivates any major expectation of cross-linguistic variation between English and French. The two languages will therefore be compared in a more exploratory manner, looking for any language-specific patterns and uncovering potential "universals" of (dis)fluency.

Lastly, the fluency-as-frequency hypothesis will be pursued as a methodological research question. I will be looking for converging evidence that support the proposed heuristic equivalence between high frequency and fluency (and its negative equivalent, i.e. low frequency with disfluency).

To sum up, the notions, theories and hypotheses developed in this chapter offer a flexible framework for the cognitive-functional study of fluencemes in a contrastive and usage-based approach. This study strives towards the overarching goal of modeling the typology of fluencemes across different registers of English and French, uncovering the inter-relation between its members and linking their most representative patterns to tentative interpretations of their relative (dis)fluency. However, analyses at this general level (which I refer to as sequence level) are limited to quantitative findings. The focus on discourse markers in this study provides a more thorough and qualitative level of analysis which brings us closer to the targeted cognitive-functional scale of fluency.

CHAPTER 3

Definitions and corpus-based approaches to discourse markers

Discourse markers (DMs) are highly central and relevant to the study of (dis)fluency as a functionally ambivalent phenomenon. One motivation for the investigation of DMs is their high informative value, relatively to the other fluencemes in the typology. Their lexical and propositional content, although limited to a semantic core and a procedural meaning (Schourup 1999), can serve as a useful basis for a number of more qualitative, functional and cognitive analyses than what would be available in a study focusing on the production of formal patterns only (e.g. truncations or pauses). In this respect, DMs are also more informative than the widely studied filled pause (*uh*) which conveys a much vaguer meaning, although it bears functional similarities with DMs (Swerts 1998). Pragmatic interpretation of DMs, combined with their syntactic behavior and co-occurrence with other fluencemes, will be taken as evidence of the relative (dis)fluency of various clustering patterns in the corpus.

The present chapter will define the category of discourse markers and their relation to (dis)fluency with respect to the wealth of previous works available on the topic and the specific research questions under scrutiny here. As is the case for fluency, there is little consensus on how DMs can be defined, which is why the first section of this chapter will first lay out the basic concepts and terms commonly used in the field and situate the present approach against this backdrop. A great deal of contrastive research has specifically tackled (groups of) DM expressions as crosslinguistic equivalents in different languages. The challenges and state of the art of contrastive DM research will be developed in Section 3.2. Section 3.3 will focus on corpus-based models designed to capture the polyfunctionality of DMs, thus discussing the merits and differences of each framework before introducing the selected taxonomy and how it handles the characteristics specific to spoken language. The relation of DMs to (dis)fluency (and its relative absence from the literature) will be developed in Section 3.4. Finally, Section 3.5 will lay out the research questions and hypotheses specifically related to the variation of DMs (analyzed in Chapter 5) and the inter-relationship between DMs and other (dis)fluent phenomena (Chapters 6 to 8).

3.1 From connectives to pragmatic markers: Defining the continuum

Discourse markers form a very slippery linguistic category which has been defined many times but still escapes consensus even after decades of research. The problem stems from the changing nature of language in general, the fuzziness of semantics and the variation of discourse in particular. Another reason for the lack of consensus concerns the many different frameworks within which the category has been investigated throughout the years, diverging either on theories, research agendas, methods or data types, perhaps to a larger extent than (dis)fluency research as we will come to see. For this reason, it has become standard practice in DM research to provide a long list of competing terms available in the literature to refer to the (apparently) same category of expressions (e.g. Brinton 1996: 29; Fraser 1999: 932; Müller 2005: 3; Aijmer & Simon-Vandenbergen 2006: 2). Beyond idiosyncratic preferences, terminology involves deeper theoretical disagreements on the definition and delineation of the DM category (see Fischer 2006 for an overview). I will start with the dichotomy between "discourse markers" and "pragmatic markers" (henceforth PMs). Hansen (2006: 28) assigns to PMs the status of an overarching category with a much broader scope, including *de facto* DMs:

> *Discourse marker* should be considered a hyponym of *pragmatic marker*, the latter being a cover term for all those non-propositional functions which linguistic items may fulfil in discourse. Alongside discourse markers, whose main purpose is the maintenance of what I have called "transactional coherence", this overarching category of functions would include various forms of interactional markers, such as markers of politeness, turn-taking etc. whose aim is the maintenance of interactional coherence; performance markers, such as hesitation marker; and possibly others.

In this view, PMs include various procedural 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: 10). 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. It is indeed far from obvious how items such as routines or vocatives are in any way similar to connectives or reformulation markers, however they might be defined. While abstract definitions can afford to be this inclusive and exhaustive, a more clearly defined category is necessary for methodological efficiency in the perspective of corpus annotation.

Another frequently encountered notion, especially in studies on writing, is that of "connectives", a term which stresses the relational meaning of expressions

connecting two segments (e.g. Fraser 1996; Prasad et al. 2008).[9] Connectives typically correspond to coordinating and subordinating conjunctions (e.g. *and*, *but*, *because*), as well as some prepositional phrases and adverbials (e.g. *so*, *in other words, however*). Relationality and the necessary presence of two abstract objects exclude from connectives non-relational markers such as *you know* or *sort of*, which only take scope over one unit. Authors working on writing tend to focus on connectives or relational DMs (see Section 3.3.1), while spoken studies tend to exclude conjunctions from the category on the grounds that they are more syntactically integrated and carry propositional information (e.g. Briz & Pons Bordería 2010; but see Cuenca 2013). Degand & Simon-Vandenbergen (2011), however, suggest that the divide should be more gradual and propose to locate different expressions on a scale from purely non-relational (e.g. *I think*) to purely relational (e.g. *because*) elements, thus acknowledging the polyfunctionality of the whole category and of its individual members. Similarly, in the present study, both relational and non-relational items are included in a broad category of "discourse markers", which is itself considered as a subtype of pragmatic markers.

Following many existing proposals, the DM category will be defined primarily on functional grounds, as in Schiffrin ("sequentially dependent elements which bracket units of talk", 1987: 31) or Hansen (2006: 25): DMs "provide instructions to the hearer on how to integrate their host utterance into a developing mental model of the discourse in such a way as to make that utterance appear optimally coherent". In order to provide more specific criteria for DM identification, I would like to suggest the following definition:

> DMs are a grammatically heterogeneous, syntactically optional, polyfunctional type of pragmatic marker. Their specificity is to function on a metadiscursive[10] level as procedural cues to constrain the interpretation of the host unit in a co-built representation of on-going discourse. They do so by either signaling a discourse relation between the host unit and its context, making the structural sequencing of discourse segments explicit, expressing the speaker's meta-comment on their phrasing, or contributing to the speaker-hearer relationship.

---

9. "Relational" and "relationality" are used in this study to refer to uses of discourse markers signaling a relation between two segments. In other words, relational discourse markers correspond to connectives. For lack of a better term, other types of discourse markers, which do not connect two segments but apply to one unit only, are termed "non-relational" discourse markers.

10. "Metadiscursive" is preferred over "discursive" since it better reflects the speaker's distance and subjectivity towards their discourse, in other words their "comments" on the message or form of the utterance.

This definition is (1) relative to other pragmatic categories, (2) formal-functional, with a primary pragmatic role constrained by syntactic filters and (3) very much indebted to previous proposals. The primacy of functional criteria over syntactic ones motivates the inclusion of both relational and non-relational DMs already mentioned above. This decision avoids any premature exclusion overlooking the many characteristics in common between relational and non-relational DMs (e.g. optionality, metadiscursive function). While the features mentioned in this definition are criterial, additional characteristics of DMs frequently found in the literature, such as "weak-clause association" (Schourup 1999: 232), short lexemes, prosodic independence or high frequency in speech (Brinton 1996), are only prototypical and therefore optional in the categorization of a candidate token as DM.

In other words, the present definition refrains from defining the category by its prototype, as prototypical definitions often have to deal with counter-examples and borderline cases. Instead, this study resorts to independent criteria which should be met by potential DMs in context (cf. the approach in Bolly et al. 2015, 2017). The combination of syntactic and pragmatic features is the result of working with authentic corpus data, where it soon appeared necessary. This to-and-froing between theoretical definitions and more practical considerations is, in my view, necessary early on in corpus-based pragmatics to ensure that the annotations match the definition, especially when dealing with such broad and complex linguistic categories (Crible 2017a). It also strives to answer Fischer's (2014: 274) call to "augment efforts of definition with efforts that further our understanding of the mechanisms that allow the items under consideration to fulfil their broad spectrum of functions" (see Section 3.3 for functional models).

To sum up, linguistic categorization is by no means theory-neutral, especially not when dealing with semantics and pragmatics. Terminological (and the underlying theoretical) choices are to some extent research-specific, and each have their own purpose and advantages: no one "best" proposal could or should be identified. Nevertheless, one can only deplore the confusion that this chaotic situation brings up, limiting the inter-operability and communication between different approaches, but overall reflecting the intrinsic complexity of the common object of study. As Maschler & Schiffrin (2015: 203) put it:

> Research on discourse markers has spread into many areas of linguistic inquiry, drawing scholars from many different theoretical and empirical orientations. Although this welcome diversity has led to an abundance of information about discourse markers, it has also led to knowledge that is not always either linear or cumulative. The result is that it is difficult to synthesize the conclusions of past research into a set of coherent and consistent findings and, thus, to integrate scholarly findings into an empirically grounded theory.

For the sake of readability and because of the motivations laid out above, I will henceforth consistently refer to "discourse markers", including in reviews of other proposals regardless of the original term, unless specified otherwise.

3.2 Discourse markers in contrastive linguistics

I will now proceed to a review of contrastive DM research, drawing a distinction between onomasiological and semasiological accounts. By onomasiological, I mean studies that are not based on closed lists of expressions which they categorize as DMs (i.e. semasiological), but rather start from the very definition of the category and observe what expressions meet the definition in context (I also refer to this type of approach as bottom-up, categorical or paradigmatic). The main point of this section is that DMs are very rarely studied in onomasiological approaches in spoken multilingual data, as opposed to the bulk of contrastive case studies, with a number of notable exceptions which will be discussed below.

DMs (or their translation equivalents: *marcadores discursivos, marqueurs du discours*, etc.) have been identified in many different languages from the Romance and Germanic families but also in Turkish, Hindi, Japanese or Quechuan languages, which would suggest the universality of the concept. However, most works focus on individual DMs or top-down selections of a handful of expressions, which by-passes the issue of categorical definition. Regarding the English-French pair, under scrutiny in this book, contrastive case studies include: Fleischman & Yaguello (2004) on *like* vs. French *genre*; Lewis (2006a) on *on the contrary* vs. French *au contraire*; Willems & Demol (2006) on *really* vs. French *vraiment*; Defour et al. (2010) on *in fact* vs. French *en fait, de fait* and *au fait*; Beeching (2017) on *just* vs. French *juste*. Given the paradigmatic scope of the present study, these references will not be reviewed in any more detail, since their results are not generalizable to the full category of DMs.

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 written data, such as causal connectives (Pander Maat & Degand 2001 on French and Dutch; Stukker & Sanders 2012 on French, German and Dutch) or reformulation markers (Rossari 1994 on French-Italian; Cuenca 2003 on English, Spanish and Catalan). In this line of works, a few French-English studies should be mentioned, namely Zufferey & Cartoni (2012), who compared causal connectives in the perspective of translation, and Dupont (2015), who focused on the position of adverbs of contrast in the framework of Systemic Functional Linguistics. 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) showed that English *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.

It appears that few studies aim at exhaustivity over the whole DM category across several languages. One of them is reported in Lopes et al. (2015) who used 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.[11] Zufferey & Degand (2013) carried out a multingual annotation experiment in a parallel newspaper corpus 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.[12] Other crosslinguistic studies have been working with spoken data as well. Kunz & Lapshinova-Koltunski (2015) compared connectives (along with co-reference and substitution) in English and German written and spoken registers and found that connectives are more affected by crosslinguistic than register variation, for instance with a higher variety of cohesive devices in German than in English. Lastly, González (2005) provides, to my knowledge, the only large-scale crosslinguistic study of spoken DMs, including both connectives (*so, anyway*) and speech-specific expressions (*well, I mean, you know*) in a corpus of English-Catalan oral narratives.

Overall, contrastive onomasiological studies covering a wide range of DMs remain extremely rare, especially in spoken data, and nonexistent for the English-French pair to date. I explain this gap in the literature by the lack of consensus regarding the definition of DMs, which I have already mentioned in the previous section. Blakemore (2002) even questions the very existence of a DM category because of this absence of consensual definition. Crosslinguistically, onomasiological approaches are even more challenging since the observed phenomena must be strictly comparable across the different linguistic systems, which explains the rarity of such approaches. Therefore, the methodological requirements of comparability will now be developed in particular relation to the DM category.

Seminal references in contrastive methodology are James (1980) and Krzeszowski (1981). The latter coined the notion of *tertium comparationis*, which was later applied to semantics and pragmatics by Jaszczolt (2003). A *tertium comparationis* is a

11. The Europarl corpus is available at http://www.statmt.org/europarl/.

12. Their corpus was gathered from the Press Europe website, available at http://www.presseurop.eu/en.

common platform of comparison which aims at optimal similarity across different languages through the use of criteria and features focusing on what is constant between systems. *Tertia comparationis* can be designed at any level of analysis and are usually research-specific in that (1) they depend on the particular languages targeted in the study and (2) they are relative to the agenda and objectives of the study. In the present case, for instance, the *tertium comparationis* is explicitly designed to encompass forms as different as *although* (typically used for discourse structure) and *you know* (typically used for intersubjectivity), as opposed to other authors who do not group them in the same linguistic category.

Connor & Moreno (2005: 155) develop a model of contrastive quantitative research, of which the *tertium comparationis* is the second step, immediately preceding the "operationalization of the textual concepts into linguistic features appropriate to each language". The authors further note that *tertia comparationis* should be functional rather than formal in order to account for grammatically distinct realizations of the same concept in different systems. This recommendation is especially relevant for DMs, which originate from a wide variety of grammatical classes. All in all, the relative absence of onomasiological contrastive studies of DMs could be very well explained by the challenges of designing a valid semantic-pragmatic *tertium comparationis*. Indeed, confusion on the boundaries of the DM category and the interplay of functional and formal features do not ease the defining task at the monolingual level, let alone crosslinguistically. This situation is reflected very directly in the literature, with a large number of monolingual case studies, some contrastive case studies, a few monolingual categorical studies and almost no contrastive categorical studies (at least for the English-French pair), as was shown above.

Given this absence of contrastive corpus research on the behavior of English and French DMs, expectations of differences are very limited. Some intuitive (and contradictory) insights are provided by contrastive stylistics: Guillemin-Flescher (1981) argues – without any empirical validation whatsoever – that coordination (as opposed to juxtaposition) is more frequent in English than in French, which could be related to a higher connective use; however, Vinay & Darbelnet (1995) claim the contrary. There is no further evidence in the literature that frequency of connectives and other DMs should be different between English and French, only that some preferences can be observed in terms of position (Dupont 2015) and meaning-in-context (Zufferey & Cartoni 2012).

3.3 Models of discourse marker functions

I hope to have made it clear so far that what lies at the core of the DM category is their pragmatic and interpretative function(s). In the previous sections, I focused on general definitions, in keeping with the inclusive scope of the present research. However, the complexity and polyfunctionality of DMs demand further investigation of detailed taxonomies, in order to better grasp what the general definition actually covers or excludes.

Given the profusion of works proposing DM-specific taxonomies in very fine-grained – not always replicable – methods, I will restrict the literature review to models targeting a broad coverage of DMs and functions. This section thus deals with functional categorizations in major corpus-based frameworks, either writing- or speech-based, and discusses their influence on the present approach, starting with the notion of discourse relation (Section 3.3.1) and moving on to more inclusive views of the functions and scope of DMs (Section 3.3.2).

3.3.1 Discourse relations in the Penn Discourse TreeBank 2.0

In this section, I present a major English framework, namely the Penn Discourse TreeBank 2.0 (henceforth PDTB, Prasad et al. 2008). This taxonomy is writing-based (i.e. originally designed for writing, although not restricted to writing) and includes different discourse relations such as cause, concession or condition in more or less fine-grained distinctions of meaning. It has been adapted to several languages (e.g. Oza et al. 2009 in Hindi or Zeyrek et al. 2013 in Turkish), and recent endeavors have started to transfer these taxonomies to spoken corpora (e.g. Tonelli et al. 2010 for spoken Italian). The PDTB, much like other writing-based models such as Rhetorical Structure Theory (Mann & Thompson 1988) or Segmented Discourse Representation Theory (Asher & Lascarides 2003), is application-oriented: it aims at high replicability and automatization, in order to be used for summarization or full-text segmentation purposes.

The PDTB 2.0 develops a lexically-based approach (i.e. start from connectives and annotate the related segments), where only relations between adjacent units are annotated, covering 43 different relation types. In their revised version of the PDTB model, Zufferey & Degand (2013) made a number of changes regarding the structure and content of the taxonomy, including the removal of all six subtypes of "condition", as well as those for "contrast", "concession" and "alternative". The revised taxonomy, represented in Table 3.1, was used as reference for the present design of relational functions of DMs.

Table 3.1 Revised PDTB from Zufferey & Degand (2013)

Level 1	Level 2	Level 3	Level 4
Temporal	Synchronous		
	Asynchronous	precedence	
		succession	
Contingency	Cause	reason	pragmatic
			non-pragmatic
		result	pragmatic
			non-pragmatic
	Condition	pragmatic	
		non-pragmatic	
Comparison	Contrast		
	Concession	pragmatic	
		non-pragmatic	
	Parallel		
Expansion	Conjunction		
	Instantiation		
	Restatement	specification	
		equivalence	
		generalization	
	Alternative		
	Exception	list	

Another interesting feature of the (original and revised) PDTB is the distinction between "pragmatic" and "non-pragmatic" (or "semantic") relations. It appears under different names in the literature, probably starting with Halliday & Hasan's (1976) "internal" vs. "external", later on "subject matter" vs. "presentational matter" (in Rhetorical Structure Theory, Mann & Thompson 1988), "content" vs. "epistemic" vs. "speech-act" (Sweetser 1990), "ideational" vs. "rhetorical" (Redeker 1990) or "objective" vs. "subjective" (Langacker 1990, applied to discourse by Pander Maat & Sanders 2000, 2001; Pander Maat & Degand 2001). These terms do not always fully overlap, as noted by Sanders (1997), but all roughly correspond to the writer's or speaker's degree of subjectivity involved in a particular discourse relation or connective. Semantic relations relate facts happening in the real world, whereas pragmatic relations are concerned with illocutionary force and structuring effects. Sweetser's (1990) tripartite theory is especially relevant for spoken language, since it further distinguishes a particular kind of subjective uses, viz. speech-act relations. The three types of relations (content, epistemic, speech-act) are illustrated by causal relations in Examples (1)–(3) borrowed from Sweetser (1990: 77–78).

(1) John came back *because* he loved her.

(2) John loved her, *because* he came back.

(3) What are you doing tonight, *because* there's a good movie on.

In (1), "because" relates the fact that "John came back" to its external/objective/content explanation ("he loved her"). In (2), however, there is no logical semantic relation between the two segments connected by "because", the relation rather stands between one fact ("he came back") and its subjective/internal/epistemic conclusion or interpretation, which could be reformulated by "John must have loved her" or "I conclude that John loved her". Lastly, in (3), "because" introduces a justification ("there's a good movie on", i.e. you should come to the theater) for the upcoming speech-act, here a question ("what are you doing tonight"), thus functioning at a different level of language (the *how* and not the *what*). This third type typically involves imperatives or interrogatives and is, by nature, more specific to speech than writing, although not impossible in the latter. For this reason, as well as because of its potential overlap with the "epistemic" type, speech-act relations will not be classified as such in the present approach but rather merged with "pragmatic" relations.

As mentioned earlier, the PDTB model has been adapted not only to other languages (e.g. Oza et al. 2009 in Hindi; Danlos et al. 2015 in French) but also to spoken data (Tonelli et al. 2010 in Italian conversations; Demirşahin & Zeyrek 2014 in Turkish). In these works, however, the original taxonomy is merely mapped onto the particular characteristics of spoken connectives (i.e. more polyfunctional, underspecified) and does not target speech-specific DMs such as *well* or *you know*, since they do not (always) meet the connectivity or relationality criterion. These works on spoken data therefore remain writing-based and the current state of the PDTB cannot account for more conversational or interactional DM functions such as turn-taking or monitoring for one's attention. The non-relational end of the DM category is left unattended by most of the literature originating from the study of discourse relations in written corpora.

Nevertheless, it has been widely acknowledged that both relational and non-relational uses of DMs co-exist in the category, and sometimes for a single DM lexeme (cf. Degand & Simon-Vandenbergen's (2011) scale of relationality). Examples (4) and (5) illustrate two more or less relational uses of the same DM, viz. French *alors* 'well'/'then'.

(4) Nietzsche le philosophe allemand parle de a une définition de l'art *alors* pas uniquement de l'art mais euh notamment de l'art (0.464) comme quelque chose qui serait du côté *alors* il dit de la santé il dit ch- surtout de la grande santé
Nietzsche the German philosopher talks of has a definition of art alors *'well' not only of art but uh among other things of art (0.464) as something which belongs to* alors *'well' he says to health he says above all to great health* (FR-clas-02)

(5) si nous savons les encourager (0.148) les libérer (0.895) *alors* euh oui la France sera bien partie pour le siècle qui vient
if we can encourage them (0.148) free them (0.895) alors *'then' uh yes France will be ready for the upcoming century* (FR-poli-02)

The two "alors" in Example (4) are not (entirely) relational in that the DMs do not connect one proposition or abstract object to another but merely punctuate the utterance and signal focus (cf. Vincent's (1993) French term *ponctuants* 'punctuators' for similar expressions). This is especially true for the second occurrence where "alors" is inserted within a prepositional phrase ("du côté de la santé") and functions as introducing reported speech ("il dit"). The first occurrence of "alors" in Example (4) is somewhat intermediary on the scale of relationality, since it both performs a punctuating function and expresses a slight reformulative or specifying meaning: it is a definition of art but not exclusively. By contrast, the "alors" in Example (5) expresses a full-fledged relation of condition (with temporal nuances) between the *si*-clause and the *alors*-clause: France will be ready for the next century if and when we can free them. Such examples of relational, non-relational and intermediary uses of a single DM motivate the inclusion of the full functional spectrum of DM expressions in a marker-based approach such as the present one, provided each use meets the criteria of the function-based definition.

3.3.2 The many scopes of DM functions

3.3.2.1 *Long-distance relations*

To address the limitations in coverage of writing-based models such as the PDTB 2.0, I would argue for a broader, more inclusive view of coherence whereby no use or scope of DMs is excluded unless it does not meet the definition criteria (i.e. procedural, non-propositional, optional, discourse-level scope). Such a perspective would therefore reconcile the relational view of connectives with more discourse-structuring uses of DMs functioning at other levels of discourse organization.

This view is supported by a number of authors working on spoken language, although not exclusively. Unger (1996: 409), for instance, acknowledges that "discourse connectives can have scope over an utterance or a group of utterances" and that connectives introducing new paragraphs minimize processing effort by signaling a change of context (paragraph breaks are equivalent to long pauses in speech, according to him). However, he admits that "though a paragraph break broadens the range of assumptions serving as candidates for the choice of a context, one particular utterance within a preceding paragraph may still be the most likely candidate as one yielding an interpretation consistent with the principle of relevance" (1996: 436). In other words, a DM at the beginning of a paragraph does

not necessarily take scope over the full previous paragraph but can be connecting a single, more adjacent segment. Crible & Cuenca (2017) discuss a complex case of *so* which illustrates such multiple interpretations. I report it here:

(6) <BB1> could you talk a little bit about the Wirral accent I I know that um (0.200) there's obviously quite a um range of accents in that part of the country <BB4> yeah (0.520) uh well I (0.290) consider myself to have a Cheshire accent because when I was born (0.300) and I lived in (0.110) on the Wirral (0.287) uh (0.333) i- (0.460) it was a Cheshire accent which is (0.440) the accent I have now though (0.270) there are overtones of (0.230) the Liverpudlian accent (0.290) however over the years certainly it has changed (0.270) and now it's very much (0.110) a Liverpool accent (0.340) and uh you know which (0.430) I'm not (0.300) I'm not saying I disapprove of it but I think it's a lazy speech and you need to (0.440) actually um (0.530) think about what you're saying I know my nephew sometimes'll to speak to me in the Liverpool accent (0.350) and I'll say please speak to me in English (0.160) but it's things like "yeah" and "you what" and (0.230) whereas you know mine is "yes" "pardon" or whatever <noise/> I'm a bit old-fashioned in that way *so* I do find the accent (0.440) is a bit harsh and it's interesting that actually that accent is spread out into the (0.270) uh (0.390) the parts of north Wales that are very near to the Wirral… . (EN-intf-03)

It is argued that "so" introducing the segment "I do find the accent is a bit harsh" can either be interpreted as (1) connecting it to the immediate co-text ("I'm a bit old-fashioned in that way"), thus signaling a relation of objective consequence; (2) acting as a conclusion to the anecdote about the nephew; (3) referring back to the previous evaluative segment ("I'm not saying I disapprove of it but I think it's a lazy speech"); or 4) introducing an answer to the interviewer's (<BB1>) original question. In Examples like (6) and others, it is not always possible to determine which scope is more prevalent in context, nor whether one is necessarily more relevant than the others, in keeping with the polyfunctionality of DMs.

Lenk (1998: 208) terms this variability "local" vs. "global" scope: "discourse segments can also be connected to other segments that are not immediately adjacent, but that were mentioned earlier in the discourse, or that a speaker intends to include later on". She further argues that this difference in scope is scalar, relative and not absolute: "local discourse markers probably represent one end of the continuum where utterance relations are marked, whereas global discourse markers represent the other end of the continuum where topic relations are marked" (1998: 211). We see that topic relations are fully considered as part of DM functions in this perspective, unlike in the PDTB 2.0, where they are not included in the taxonomy.

3.3.2.2 *Co-occurrence of discourse markers*

Multiple levels of discourse coherence are particularly relevant in the case of co-occurring DMs, which are argued to be a strong tendency of unplanned spoken discourse in Crible & Cuenca (2017). In the speech string, DMs tend to aggregate in clusters of two or more separate DMs forming one new complex unit, depending on their degree of fixation and semantic non-compositionality. The phenomenon of DM co-occurrence and combination has been studied by a number of authors, especially in French (e.g. Luscher 1993; Razgoulieva 2002; Waltereit 2007; Dostie 2013; Crible 2015). According to Cuenca & Marín (2009), DMs can combine either as juxtaposition (different functions with different scopes), addition (different functions with the same scope) or composition (the DMs now form one complex unit with a single function).[13] Although combined DMs are often language-specific (e.g. French *bon ben* 'well'), they have been identified in many languages and are sometimes shared cross-linguistically (e.g. *or else*, French *ou sinon*, Spanish *o sino*; *and then*, French *et puis*), which points to the universality of this tendency to cluster independent expressions into combined units with a more or less complex meaning-in-context.

3.3.2.3 *Utterance-final discourse markers*

Besides the tendency of spoken discourse to produce clusters of DMs, another explanation for the use of combined DMs is the ability of DMs to occur in final position of the utterance, especially in spoken language, which can in turn be related to the temporality of this modality. Given the low planning in speech production, not every word in an utterance is pre-planned and speakers sometimes have to add after-thoughts or backward-looking information (such as DMs) in order to improve the connectivity and coherence of the on-going utterance in a retrospective manner. Utterance-final DMs range from relational (e.g. *though*) to non-relational uses (e.g. *you know*), and sometimes combine with DMs in different syntactic positions, as in Example (7).

(7) I took lots of photographs I don't know if they're any good *though but* we shall see (EN-conv-07)

In this example, *though* connects two segments which are both antecedent in a concessive relation ("I took lots of photographs" but "I don't know if they're any good"), while *but* is initial with respect to the next utterance which it introduces

13. Luscher (1994) makes a very similar distinction between additive and compositional sequences of connectives. In the former, the co-occurring DMs share the same syntactic scope yet convey different instructions; in the latter, the DMs share the same syntactic scope and convey partially common instructions, one of them strengthening the other (1994: 221–222).

("we shall see"). Unlike this example, most DMs in final position tend to perform hearer-oriented functions where they call for attention and check the hearer's comprehension on the utterance just produced (Degand 2014). In sum, there is no one-to-one mapping between DM function, position, relationality and scope.

3.3.2.4 *Speech-based models and present taxonomy*

To sum up, so far, DMs function at many levels with different scopes and directions, sometimes simultaneously, as in the case of complex DMs, combining relational with non-relational, retrospective and prospective functions. This flexibility and variability motivates the choice of functional taxonomies which include more functions than the purely relational (writing-based) model of the PDTB 2.0. Indeed, most proposals specifically designed for spoken DMs or spoken language in general cover more functions than mere discourse relations, starting from Halliday's (1970) seminal distinction between ideational, textual and interpersonal functions. His third domain appears as specific to the spoken mode and typically targets expressions such as *you know*. Interpersonal functions are also found in other proposals such as Schiffrin's (1987) "participation framework" or the "modal" functions in the Val.Es.Co model (Briz & Val.Es.Co Group 2003; Briz & Pons Bordería 2010) and in Cuenca (2013). Topic relations and higher-level discourse organizing functions are also accounted for by the "textual", "exchange structure", "structural" and "sequential" domains in Halliday (1970), Schiffrin (1987), Cuenca (2013) and Redeker (1990), respectively (see Maschler & Schiffrin 2015 for a comparative review of three functional approaches to discourse markers).

The present approach to the functions of DMs is strongly rooted in Redeker (1990), and more precisely its adaptation by González (2005), who developed a fine-grained taxonomy of about twenty functions grouped in four components of discourse structure, namely *ideational*, *rhetorical*, *sequential* and *inferential*. This four-fold model takes up terms and notions which should be familiar by now: the distinction between objective (ideational) and subjective (rhetorical) functions, the inclusion of topic or text-structuring (sequential) functions and that of typically non-relational hearer-oriented (inferential) functions. González (2005) combines functions which are specific to spoken language (e.g. playing for time to think, face-threat mitigation) with typical discourse relations (e.g. conclusion, addition) that also exist in writing.

Although innovative and fairly exhaustive, this proposal leaves some room for improvement, in particular regarding the operational definition of the functions and their classification in coherent categories. The final model, revised from González's (2005) original taxonomy, is developed in Crible (2017a) and reported in Table 3.2 (detailed definitions of each label are provided in Crible 2014).

Table 3.2 Present taxonomy of DM functions

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 organized in four domains (generic function labels) and thirty functions (specific function labels). These domains and their functions were elaborated and tested on corpus data to make sure that the labels are operational and cover all possible types of DMs. The resulting corpus-based taxonomy is thought to match the present functional definition of DMs, and to adequately reflect the polyfunctionality of the DM category in the perspective of corpus annotation. More details on this taxonomy and how it was applied in the present study are provided in Section 4.2.2 and in Crible (2014).

3.4 "Fluent" vs. "disfluent" discourse markers

I will now address the link between DMs and fluency in order to (1) describe the specific contribution of DMs to fluency and disfluency and (2) situate them within the typology of fluencemes. I will first lay out some of the features that make DMs relevant to (dis)fluency research. I will then discuss the rare studies which tackled the relation between DMs and (dis)fluency and try to explain their relative absence from the literature.

3.4.1 DM features and (dis)fluency

Before the corpus era, DMs were sometimes mentioned rather indirectly in relation to fluency and disfluency, with studies pointing at the syntactic or pragmatic characteristics which make them relevant for the quality (or failure) of language production and perception. Starting with non-empirical, somewhat outdated reports, DMs used to be stigmatized as "a sign of dysfluency and carelessness" (Brinton 1996: 33) resulting from "unclear thinking, lack of confidence, [or] inadequate

social skills" (Crystal 1988: 47) by authors such as O'Donnell & Todd (1980: 67) or Ragan (1983: 166), who attribute their use to "unskilful speakers" and "powerlessnes", respectively. As Gilquin & De Cock (2011) report, DMs were often termed negatively ("exasperating expressions" in Stubbe & Holmes 1995; "throwaways" in Erard 2004; "pollution" in Boula De Mareüil et al. 2005: 27) until corpus studies uncovered their many functions and more positive roles.

Other qualitative accounts of DMs have identified a number of areas where DMs are potentially beneficial for the participants. Ejzenberg (2000) mentions their use in turn-taking and common ground. Hasselgren (2002: 143) describes DMs as "a system of signals bringing about smoother communication". Götz (2013) associates DMs to the perception of naturalness, especially for non-native speakers (see also De Cock 2000: 52). The connectivity of DMs is even part and parcel of Pawley & Syder's (1983) definition of nativelike fluency as "the native speaker's ability to produce fluent stretches of spontaneous connected discourse". All in all, the functional ambivalence of DMs is reflected in rather opposite (qualitative) accounts of the negative and positive roles of DMs, without providing more quantitative evidence of the proportions and conditions under which they are used more or less fluently.

In my view, DMs reflect and support cognitive processes of production and comprehension. Figuratively, it could be said that their multiple scopes and crucial role in planning processes add some spatiality to the temporality of speech: DMs segment spoken discourse much like punctuation marks and paragraph breaks segment written texts. In this way, they allow both speakers and hearers to backtrack or project their attention along the string of words, in a relative freedom of movement without which communication cannot seem to be performed efficiently.

3.4.2 Previous corpus-based accounts of DMs and disfluency

3.4.2.1 *Exclusions based on DM polyfunctionality*

Discourse markers are usually absent from fluency research or included only selectively on rather arbitrary closed lists. Authors tend to motivate this exclusion by various reasons. One recurrent justification is found in approaches to disfluencies as removable errors (cf. Section 2.3.2.1), where DMs are discarded on the grounds that they might be intentional, and thus non-removable. This approach suggests a distinction between some uses of DMs which are disruptive and removable, and others which are more clearly fluent and useful (henceforth not disfluencies). This view is represented notably by Shriberg's (1994) model and its adaptation to Swedish by Eklund (2004). In her typology, Shriberg (1994) distinguishes "discourse markers" from "coordinating conjunctions". She groups the former with filled pauses and explicit editing terms in the category of "extra-syntactic-words", while the latter are

categorized as "inter-sentence-words". However, they are excluded very early on in her thesis: "Other elements that have been grouped with filled pauses as 'fillers' in some accounts – in particular discourse markers ('well', 'like') – do not fall under the category of 'disfluency' in the present work because they are arguably part of the speaker's intended utterance" (Shriberg 1994: 2). She further specifies that, unlike other disfluencies, DMs are not deleted from transcriptions (cf. the cleaning objective of her research) and are only annotated when they occur within another disfluency (i.e. in the editing phase).[14]

Similarly, Eklund (2004) acknowledges that some DMs could be considered as disfluencies, yet he excludes them with no further justification. Shriberg's (1994) analogy between discourse markers and filled pauses or "fillers" is quite frequent in the literature (e.g. Swerts 1998; Pawley & Syder 2000; Tottie 2015) and is not always explicitly defined, which makes a precise literature review hardly achievable in this regard. In Bortfeld et al. (2001), however, DMs are neither considered as a type of disfluency, nor grouped with fillers. The authors specifically oppose other accounts such as Broen & Siegel (1972) who include them "despite the possibility that these discourse markers have quite distinct discourse functions" (Bortfeld et al. 2001: 141).

Overall, in this first line of research, authors motivate the exclusion of DMs by acknowledging their polyfunctionality. The ambivalence between "fluent" and "disfluent", intended or unintended DMs is incompatible with the rather negative view of disfluencies in these works. A related perspective is taken by studies on L2 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. Although the underlying assumption of the polyfunctionality of DMs is compatible with the approach taken in this study, I do not share either of these objectives (summarization or L2 fluency) and aim at a more bottom-up selection, so that I will not pursue the discussion of these works.

3.4.2.2 *Exclusions for methodological validity*

Another, more practical reason for the usual exclusion of DMs in the studies currently available 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). The former opt for a number of fine-grained distinctions between conceptually related categories, which are probably detrimental to the reliability of the overall method, while the latter chooses

14. A similar restriction can be found in Pallaud et al. (2013: 10), who found DMs to be included in 10% of "disfluent interruptions".

to work with a closed-list approach to avoid the complex identification process: "Because of the many uses of DMs in speech, and the resulting complexity of defining and identifying them, we will annotate only a limited set of discourse markers" (2003: 6), namely *actually, anyway, basically, now, see, so, I mean, let's see, like, well, you know* and *you see*. A potential problem of this list is the absence of more generic conjunctions which are extremely frequently used as DMs, such as *and* or *but*. Strassel (2003) even specifies that the subordinating and connective uses of *so* are excluded from the annotation.

Although quite restrictive, such a method is preferable over vague definitions which do not clearly state out the bottom-up criteria or top-down selections used during the annotation. One such example is Besser & Alexandersson (2007), who include a DM category in their typology of disfluencies but with a rather restrictive definition ("giv[e] the speaker time to think of what to say next and to hold the turn") exemplified by *I mean, so, well, you know, like*, while it is obvious that these DMs also perform many more functions than stalling for planning.

3.4.2.3 *Treatment of DMs and disfluencies as distinct categories*

The next trend of research contains works that do study both DMs and disfluencies in rather inclusive approaches, but treat them as two distinct phenomena, thus opposing the present view of DMs as one type of fluenceme. Two such works from the French literature can be identified, namely Beliao & Lacheret (2013) and Boula de Mareüil et al. (2013). Their framework and results will be developed so as to provide a comparative basis to the present analysis. Starting with Beliao & Lacheret (2013), the authors distinguish between prosodic (lengthening) and morphosyntactic disfluencies (interruptions, repetitions), which they term "discursive markers" that "come as series of impaired verbal constructions, such as *uhu, well, uh, so, hem*, etc. These units […] are equipped with an illocutionary operator but they do not convey information content" (2013: 6). In their corpus study, they found that DMs are more often combined with disfluencies than the opposite (proportion of disfluencies combined with DMs) and that disfluencies are overall more frequent than DMs. Furthermore, they found an association between the joint presence of both DMs and disfluencies on the one hand, and discourse type on the other, with lower frequencies in planned public speech. The conclusion of their study highlights the need to combine prosody (disfluencies) and syntax (discursive markers) to better understand spontaneous speech.

In Boula de Mareüil et al. (2013), the authors focus on the interaction between DMs, disfluencies and overlapping speech. DMs are acknowledged in their polyfunctionality, from stalling to more structuring uses with expressions such as French *alors* 'well', *donc* 'so', *mais* 'but', *enfin* 'I mean', *et* 'and' or *je crois que* 'I think that'. To this rather wide view of the DM category, the authors add three subtypes

of disfluencies, namely filled pauses (*euh*), repetitions and revisions. Overlaps are themselves divided depending on whether the overlap leads to a change of speaker (turn stealing) or not (backchannelling). In a corpus of political interviews, they found more filled pauses and repetitions in journalists, whereas guests produce more revisions and DMs. The most frequent DM expressions can roughly be classified as structuring (typically *alors*), stance-taking (typically *je crois que*) or interactional (typically *hein* 'right'). DMs are twice as frequent before a disfluency than right after. Although potentially more inclusive than Beliao & Lacheret (2013), this study still provides a coarse-grained picture of DM behavior, only taking into consideration positional information, local co-text and participants' status, instead of more pragmatic variables such as DM functions, except for the three general types which they identified and which seem to be strongly based on the semantics of the expression.

Another study which treats DMs and disfluencies as separate categories is Denke (2009), who focuses on a shortlist of three DMs and compares their production across native and non-native English speakers. This work stands out from the other L2 studies as well as from the previous two references (Beliao & Lacheret 2013; Boula de Mareüil et al. 2013) in that it includes a much more qualitative analysis of the DM functions. She takes up Erman's (2001) three domains of use, namely text-monitors (coherence-building, encoding and editing a text), social monitors (interactive and comprehension-securing) and metalinguistic monitors (attitudinal, commitment to truth and importance of the message). Textual markers seem to have a special relation to fluency through their editing functions, often connected with corrections, restarts and word search. Denke (2009) found that this text-monitoring function is the most common in both native and non-native speakers, which she explains by the monologic nature of her data (seminar presentations). She also identified a higher polyfunctionality of *you know*, compared to *I mean* and *well*, with uses across all three domains, although it appears to function predominantly as text-monitoring, especially in non-native English. She concludes that there are no major differences between native and non-native speakers when looking at the generic domain only (textual vs. social vs. metalinguistic), but that preferences emerge in specific functions (e.g. native use of *you know* and *well* as markers of reported speech; non-native use of *I mean* as marker of specification). The major limitation of Denke's (2009) contribution is the absence of integration between her analysis of DMs and that of repairs and repetitions, which are all considered individually without a synthesizing approach.

Overall, it appears that the great majority of fluency research makes some rather strict restrictions on their inclusion of DMs, whether on practical or more theoretical grounds. By contrast, DMs are here considered as full-fledged markers of (dis)fluency, in keeping with the assumption of functional ambivalence (symptom

vs. signal) underlying the present research. To the best of my knowledge, the present study is the first attempt to combine these two levels of analysis, namely an intensive and extensive annotation of DMs with a word-level tagging of fluencemes, thus reconciling the two fields of study and filling the gap on (crosslinguistic) onomasiological investigation of these phenomena.

3.5 Summary and hypotheses

DMs are defined in the present study as a subtype of pragmatic markers including both connectives and speech-specific expressions characterized by their syntactic optionality and polyfunctionality. DMs in speech appear to perform a wide array of functions at different levels and scopes of discourse and therefore require speech-specific models to be analyzed in their full expressive potential, beyond what writing-based models can provide. DMs are considered to be one type of fluencemes, alongside pauses or repetitions. The present corpus-based approach to DMs aims at filling a gap in onomasiological contrastive studies (which are almost inexistant in multilingual spoken corpora, as opposed to the bulk of case studies) and in fluency research, where DMs are often excluded or highly restricted.

From the literature review and theoretical background developed in this chapter, a number of research questions and hypotheses emerge regarding the behavior and variation of DMs. Three major sets of analyses will be carried out in separate chapters. Firstly, an exploratory investigation of the positional and functional behavior of DMs across registers and languages will uncover typical configurations in different contexts of use and potentially universal patterns (Chapter 5). Secondly, the combination of DMs with the other fluencemes in the typology will be analyzed, striving towards a cognitive-functional scale of (dis)fluency (Chapters 6 and 7). Thirdly, the (dis)fluency of discourse markers will be tackled from a more qualitative angle through the investigation of their role in the context of overt repairs, thus further distinguishing fluent from disfluent uses of DMs (Chapter 8). These three steps also correspond to a difference in analytical levels, viz. DM-based, sequence-based and repair-based. They are further distinguished by their explanatory power: Chapter 5 is purely descriptive, taking annotations and metadata as main evidence for the interpretation of the results; Chapters 6 and 7 strive towards more theoretical explanations of the observed patterns, in light of the usage-based framework developed in Section 2.5; Chapter 8 combines quantitative and qualitative methods to provide more direct interpretations of the (dis)fluency of DMs. The specific hypotheses of Chapter 8 will be developed in their own Section (8.1.4) given that they are somewhat independent, although pertaining to the same general approach and integrating results from Chapters 5 to 7.

First, in line with hypotheses on fluencemes (Section 2.6), previous contrastive research does not suggest any expectation of differences between French and English DMs at the basic level of frequency. Similarly, I expect the most frequent DM expressions to be semantically equivalent. By contrast, given the strong connection between DM use and planning pressure, I expect to find relatively more DMs (higher frequency and greater variety of DM expressions) in spontaneous discourse than in registers with a higher degree of planning. Furthermore, in light of the hearer-oriented uses of DMs and their role in turn-taking and turn-holding, interactive registers (i.e. dialogue, free exchange) are expected to show a higher frequency and diversity of DMs.

The present onomasiological study will allow us to confirm the centrality of some DM features often mentioned in the literature, namely the grammatical heterogeneity of the category, the initial position of DMs, their prominent structuring function and their tendency to co-occur with one another. Any meaningful co-variation of features will be statistically identified (for instance, position by function or function by register), thus illustrating the analytical potential of a paradigmatic, bottom-up approach to the category (as opposed to the more restricted lens of case studies). I will further use corpus data to test hypotheses gathered from previous works, and explore any further interaction between variables, thus providing an exhaustive portrait of DMs in English and French.

Multifactorial models will then be computed to integrate the variables from the sequence level (fluencemes) and the DM level. In this regard, I expect a high attraction between text-structuring DMs and pauses, given their connection with discourse planning and unit boundaries, significantly more so than other functional domains. Any register-specific or language-specific association of domain and sequence type will be identified at various degrees of abstraction.

Furthermore, I would like to propose a subset of so-called "Potentially Disfluent Functions" (henceforth PDFs), which correspond to uses of DMs conceptually related to (dis)fluency, namely *monitoring* (checking for understanding, calling for help), *punctuation* (stalling, planning) and *reformulation* (paraphrase and actual corrective relations) (see Crible 2014 for the precise definition of all functions in the taxonomy). Given their semantics, DMs expressing PDFs should be particularly associated with rather disfluent sequences of fluencemes, that is, "symptom" rather than "signal" uses.

Overall, the type of control and perceptive validation which psycholinguistic experiments can provide will not be met by the present corpus-based research. However, the number and diversity of fine-grained variables of analysis, combined with considerations of language and register and related to cognitive assumptions of the usage-based framework, all vouch for a robust methodology which should uncover interesting results regarding the production of (dis)fluent discourse.

CHAPTER 4

Corpus and method

The present approach to DMs and (dis)fluency in speech is a usage-based, empirical study of language in use and thus requires authentic data as working material to test the hypotheses presented above, in keeping with the strong tendency towards corpus approaches to cognitive linguistics and pragmatics (e.g. Gries & Stefanowitsch 2006; Schmid 2012). The contrastive and variationist perspective of this research, as well as its focus on the production of language, further call for a corpus-based methodology that will allow us to compare distributions of observed phenomena in different settings, thanks to a comparable corpus design with informative metadata and a valid *tertium comparationis* (Krzeszowski 1981; Connor & Moreno 2005).

In this chapter, I will first describe the structure and content of *DisFrEn*, the dataset which was used for the present research. I will then move on to the two annotation protocols that have been elaborated and applied to *DisFrEn*, following a number of technical and methodological instructions provided by the coding schemes (Crible 2014, Crible et al. 2016). The particular, more qualitative methodology used for the analysis of DMs in repairs will be presented in Chapter 8, where it is relevant.

4.1 The *DisFrEn* dataset

The data used for this research does not consist of newly collected texts recorded for the present purposes but rather of a compilation of already existing transcriptions, following selection principles which meet the research questions in this book. They were gathered from available source corpora in French and English in a comparable corpus design and underwent a uniform technical formatting.

4.1.1 Source corpora

Spoken corpora and databases are not as large and as available as their written counterparts, for obvious reasons related to the more intrusive technique to collect the data (cf. "observer's paradox", Labov 1972) and the human- and time-cost of its encoding into a digital format. As a consequence, large and freely available banks of

spoken data, which also provide the audio files and cover a wide array of situational settings, are rather scarce even to this day, especially for lesser-known languages. This is however not true for English, since corpus pioneers mostly came from Great Britain and the United States, thus providing the English language with several reference spoken corpora, e.g. the British National Corpus (BNC Consortium, 2007) or the Santa Barbara Corpus of Spoken American (Du Bois et al. 2000–2005).

Spoken French is less well documented, with smaller or more specific corpora that do not meet the requirements of size and representativeness of a reference corpus, although several projects are currently working towards a reference corpus for French – see the Orfeo project (http://www.projet-orfeo.fr/). As a result, most corpora of spoken French are built for more specific research purposes and often comprise either many different registers in small quantity (e.g. the C-PhonoGenre corpus, Goldman et al. 2014; the Louvain Corpus of Annotated Speech, LOCAS-F, Degand et al. 2014) or one (usually experimental) speaking task in larger quantity (Phonologie du Français Contemporain, PFC corpus, Durand et al. 2002, 2009; Corpus of Interactional Data, CID corpus, Bertrand et al. 2008). The other type of resource in French is databases (e.g. Corpus de langue parlée en interaction, CLAPI, Balthasar & Bert 2005; VALIBEL, Dister et al. 2009) which differ from corpora in that they are collections of texts from multiple sources as opposed to a single well-defined design. Databases are therefore not always easily accessible (different authorship restrictions for their different parts or collections) and often present a heterogeneous format (e.g. audio files not always retrievable for the whole database, inconsistent metadata). Nonetheless, they are very valuable because of their size and the diversity of text types they include.

The absence of a reference corpus for spoken French and the difference between English and French in this matter is reflected in the number of source corpora across French and English in the present dataset *DisFrEn*. The English texts in *DisFrEn* come for the most part from the British component of the International Corpus of English (ICE-GB, Nelson et al. 2002), a one-million-word corpus of written and spoken British English structured by situational metadata. Despite the age of this corpus (recordings date back to the 1990s) and the technical limitations that came with it (no word-to-sound alignment, poor quality of the audio files), ICE-GB was chosen for practical reasons, namely the availability of both the transcript and the soundtrack, and the structure of the corpus by situational features which roughly correspond to the metadata system adopted here. Although speakers' age is a well-known relevant variable in discourse marker use (e.g. Andersen 1997), it is not available in the metadata of this corpus. However, since sociolinguistic variables are not the focus of this research, this shortcoming has a limited impact for the present purposes.

The remaining English data comes from the Backbone project (Kohn 2012), which consists of freely available video recordings of interviews in several languages (including English and French) from the years 2009–2011. This more recent data was used to address the absence of face-to-face interviews in ICE-GB. Other texts from Backbone were also used as a pilot corpus for the design and testing of the DM-level annotation protocol: 27 transcripts of interviews (no sound) amounting to about 28,000 words and 2.5 hours in English and in French, which are not found in the final corpus *DisFrEn* to avoid any training effect during the annotation.

Turning to the French subcorpus, as mentioned before, the situation is slightly more chaotic. Sampling from several source corpora was therefore necessary. The prime resource was the VALIBEL database (Dister et al. 2009), a collection of (partly aligned) transcripts recorded from the 1990s to the present day in French-speaking Belgium. VALIBEL comprises a range of different types of interactions from which were selected conversations, face-to-face interviews and news broadcasts, amounting to more than 40,000 words. The contributions of the other French source corpora are much smaller but necessary to fill some gaps in the corpus structure when a particular data type was not available in VALIBEL or not in sufficient quantity. These resources are the following: CLAPI (the "Artisans" and "Assureurs" corpora of phone calls, Palisse 1997); C-PhonoGenre (sports commentaries and political speeches, Goldman et al. 2014); LOCAS-F (news broadcasts, political speeches, radio interviews, Degand et al. 2014); French Corpus of Humorist Speech (C-Humour, radio interviews, Grosman 2016) and Rhapsodie (a treebank for multiple interaction settings collected from other corpora, Lacheret et al. 2014). Similarly to the English subcorpus, no particular attention was paid to sociolinguistic variables such as age or language variety, with productions from both French and Belgian speakers. It is rather the content of these resources in terms of registers and speaking tasks which was considered, as detailed in the following section.

4.1.2 Comparable corpus design

The dataset resulting from the sampling described above can be characterized as a comparable bilingual corpus balanced across eight interactional settings. In this section, the internal structure of *DisFrEn* will be presented with its dual metadata system that refers to the speaking tasks at hand in two different ways. Priority was given to the balance between languages for each register, rather than the balance between registers within each language. The ideal of perfect balance between each subcorpus (e.g. the subcorpus of French conversations is as large as that of English interviews) was not met due to the scarcity of certain data types (e.g. classroom lessons).

First, if we refer to the subcorpora in terms of register or task labels, *DisFrEn* represents eight different settings: free conversations, phone calls, face-to-face interviews, radio interviews, classroom lessons, sports commentaries, political speeches and news broadcasts. It amounts to 896 minutes in total (about 15 hours of recordings), giving an average of 56 minutes for each register in each language. The actual internal structure can be seen in Table 4.1.

Table 4.1 Words and minutes per register per language in *DisFrEn*

Subcorpus	English		French		Total	
	words	min.	words	min.	words	min.
conversations	17479	88.70	17432	89.81	34911	178.51
face-to-face interviews	17055	92.60	18043	103.87	35098	196.47
radio interviews	8773	41.97	8416	38.83	17189	80.80
phone calls	9747	44.22	6783	30.97	16530	75.19
classroom lessons	9425	64.84	3723	23.24	13148	88.08
political speech	8650	60.93	7824	59.18	16474	120.11
news broadcast	7046	39.89	6788	36.47	13834	76.36
sports commentaries	8237	39.34	6279	41.10	14516	80.44
Total	**86412**	**472.49**	**75288**	**423.47**	**161700**	**895.98**

It clearly appears that two registers emerge as the most represented in the corpus, namely face-to-face interviews and conversations. This difference with the other registers is voluntary and reflects an interest for spontaneous language use, as opposed to more formal registers with a more conventional and fixed setting such as news broadcasts. Most other settings comprise around 40 minutes in each language, except for political speeches and English classroom lessons which are slightly larger with 60 minutes.

The most striking limitation of this design is the difference between English and French classroom lessons, which is due to the scarcity of this data type in French resources (only two texts were found in VALIBEL and Rhapsodie). Consequently, relative frequencies per thousand words will be used instead of raw frequencies to ensure the comparability of the different subcorpora, even those which share a similar number of words. Each text in *DisFrEn* underwent a similar (semi-)automatic technical treatment, resulting in a sound-aligned, word-segmented corpus following uniform rules and transcription conventions.

With more than 160,000 words, *DisFrEn* is the largest spoken corpus fully annotated for discourse markers. Nevertheless, it is much smaller than other written corpora investigating similar research questions: for example, the PDTB corpus (Prasad et al. 2008) contains one million words from the Wall Street Journal (about 40,000 annotated items); its French counterpart, the French Discourse Treebank

(Danlos et al. 2015), is based on the French Treebank corpus (Abeillé et al. 2003) and amounts to 535,000 words. The RST corpus (Rhetorical Structure Theory, Carlson et al. 2002), however, is comparable in size to *DisFrEn* with 176,000 words. Turning to spoken corpora, the LUNA corpus of dialogues (Tonelli et al. 2010), for instance, was partially annotated for discourse relations (under the PDTB 2.0 framework) on a sample of 25,000 words. In sum, in spite of its small size, *DisFrEn* is relatively large compared to other (spoken) discourse-annotated corpora, especially considering the amount of qualitative annotations it includes (see Sections 4.2 and 4.3 below).

4.1.3 Corpus structure in situational features

Structuring a corpus in terms of speaking tasks or registers is the traditional, most direct method of description. However, task labels are neither interoperable nor fine-grained enough to contrast the different registers between themselves. The variationist hypotheses presented in Section 2.6 require more accurate metadata which allow to rank the registers against qualitative scales of spontaneity or preparation, variables which are highly relevant to the study of (dis)fluency. A scalar approach to registers in degrees and features, as proposed here, offers complementary information and vouches for better comparability with other corpora. This refined system is inspired by Koch & Oesterreicher (2001) and was elaborated in the framework of my doctoral research (Crible 2017b) jointly with the other project members (A. Dumont, I. Grosman and I. Notarrigo).

Six situational features can be distinguished. The first feature is elicitation and refers to the presence and weight of an experimental protocol constraining the interaction. In *DisFrEn*, only two levels are represented: natural (authentic production free from any experimental protocol, not generated for specific research purposes) and semi-structured (natural production in the framework of a flexible experimental protocol, monitoring the choice of topic but allowing the speaker to choose their wording, e.g. a sociolinguistic interview). Natural registers are much more frequent in the corpus, restricting semi-structured data to face-to-face interviews.

The second feature is the number of speakers actively taking part in the interaction, thus excluding by-standers and silent participants. It is fairly basic and distinguishes monologues, dialogues and multilogues (anecdotal in the corpus).[15]

The next feature is the degree of preparation or the extent to which the speaker prepared their speech. It distinguishes between: spontaneous settings, where the

15. The term "dialogue" is used in this study to refer to interactions between two participants. It can be distinguished from "multilogue" (more than two speakers) and from "conversation", which refers to a specific type of interaction (spontaneous, interactive), regardless of the number of speakers involved.

speaker conceptualizes as they speak; semi-prepared settings, where the speaker has prepared the general frame of their speech with a possible visual support (e.g. written script, slides); prepared settings, where both content and form of the speech have been fully scripted. In *DisFrEn*, spontaneous and semi-prepared settings have a fairly similar distribution, while fully scripted settings are less represented.

Another feature closely related to the situational hypotheses on fluent and disfluent speech is that of interactivity, i.e. the speaker's ability to adapt their speaking behavior to the other speaker's with respect to what is expected from their status in the interaction. Interactive registers are characterized by a symmetrical relationship between speakers where all speakers are allowed to hold the floor. Semi-interactive registers show an asymmetrical relationship where one speaker holds the floor more than the others without excluding sporadic interventions from secondary speakers. Non-interactive registers correspond to communication settings where one speaker keeps the floor nearly continuously without leaving turn-taking opportunities to the other participants (if any). All three levels are represented in similar proportions in *DisFrEn*.

Then, the feature of media coverage defines the extent to which broadcasting is the main goal of the interaction, with three levels again: broadcasting is the main aim of the interaction; the interaction is broadcast but would have taken place even without broadcasting; the interaction is not broadcast. In *DisFrEn*, the intermediary level is only represented by English political speeches which consist of parliamentary debates, thus differing from their French counterparts which are recorded to be TV-broadcast.

The final situational feature is a binary category that specifies whether the interaction is caused by one speaker's professional activity or not. This basic distinction is an indirect measure of the formality of the setting, assuming that professional encounters are more formal than private interactions, thereby avoiding the complex and rather subjective task of defining levels of formality as such. In *DisFrEn*, for reasons of availability, professional settings are much more frequent than non-professional settings.[16]

The crosslinguistic differences between the different levels are generally minor and correspond to previously mentioned gaps in the corpus structure (e.g. fewer classroom lessons in French). To sum up, situational features are mostly referred

16. Face-to-face interviews were categorized as professional registers since the interviewer's motivation for the interaction is scientific, therefore professional. In addition, interviewees were mostly recruited because of their profession (e.g. to talk about their job). However, the interaction is "less" professional than others in the corpus such as classroom lessons or news broadcasts, and a different categorization (i.e. as non-professional) would considerably improve the balance in the data.

to for their analytical power, their precision and interoperability as a bridging tool between different registers and corpora. They will be used in conjunction with task labels depending on the particular research question at stake.

The comparability between the English and French subcorpora mainly rests on the similarity of the interaction types which they include, described either in terms of task labels or as situational features. Nonetheless, *DisFrEn* presents several caveats which need to be considered when evaluating the generalizability of the corpus findings: as mentioned before, not all registers are equally represented within and between languages; the age of the source corpora varies greatly within and between languages, from the early 1990s to the 2010s; the quality of the audio recordings, of the transcriptions (e.g. different transcribing conventions) and of the segmentation (none to phoneme) all undermine the technical comparability of the corpora, which might have consequences for the analysis and interpretation. It remains that multilingual spoken corpora are very rare and challenging resources, especially when covering as many registers as is the case in *DisFrEn*, so that its value lies in the representativity of the corpus design and in the richness of the annotations, as developed in the following sections.

4.2 Discourse marker annotation

As mentioned in Chapter 3, several reasons motivated the elaboration of a specific coding scheme for the annotation of discourse markers in speech, motivations that are briefly summarized here:

- the lack of consensus in the field in defining and annotating DMs;
- the relative absence of frameworks specifically designed for the study of spoken language, as opposed to writing-based definitions and taxonomies;
- the ambition of this study to cover the whole DM category, adopting an inclusive definition, thus grouping discourse-relational devices with non-relational DMs.

Another major difference between the present approach and existing proposals in the literature is that this annotation targets DMs (i.e. explicit words) and not discourse relations, which can be both explicit or implicit. Definition criteria, functional values and overall annotation procedure differ greatly between studies focusing on relations and those focusing on the DMs that express these relations, with complexities and specific challenges in each group.

In this section, I will report on the coding scheme and annotation procedure at DM level. For reasons of space, complete details of each variable accounted for

in the coding scheme will not be repeated here (see Crible 2014) but rather the decision-making process and the main criteria that were used during the annotation of *DisFrEn*.

4.2.1 Identification of DM tokens

As mentioned before, the literature presents conflicting definitions of what is to be included in the category of DMs. Therefore, before turning to the actual annotation of DMs, it was necessary to specify the elements the protocol applies to, thus addressing the issue of DM identification. I report here the definition of the DM category as introduced in Section 3.1:

> DMs are a grammatically heterogeneous, syntactically optional, polyfunctional type of pragmatic marker. Their specificity is to function on a metadiscursive level as procedural cues to constrain the interpretation of the host unit in a co-built representation of on-going discourse. They do so by either signaling a discourse relation between the host unit and its context, making the structural sequencing of discourse segments explicit, expressing the speaker's meta-comment on their phrasing, or contributing to the speaker-hearer relationship.

Discussion of this definition with another annotator with a different expertise (Crible & Zufferey 2015) revealed that the definition was too weakly prescriptive to be entirely replicable. As a result, a list of criteria was added to restrict individual biases and the inherent ambiguity of speech as much as possible, thus striving towards an operational identification process. These additional features are pragmatic and syntactic, and state that the selection of potential DMs is first and foremost based on functional grounds, i.e. the item must fulfill at least one function from the four domains identified in the definition. They must be highly grammaticalized, following the requisites of fixation and semantic bleaching (e.g. Hopper & Traugott 2003; Dostie 2004). DMs are also strictly syntactically optional, thus excluding phrases such as *because of* since their removal would leave the utterance ungrammatical without a change in phrasing. Another feature related to the preceding one is the syntactic and semantic autonomy of the unit the DM applies to, where autonomy is defined as the presence of a finite predicate, including subclauses (as in Example (1) below) but excluding a number of constituents such as relative clauses, infinitive, nominal and prepositional phrases (as in Example (2)) except when these are acting as a-verbal predicates.

(1) the other thing with with it is that (0.180) *because* we're a comprehensive (0.270) community school (0.450) um (0.800) part of the funding (0.450) is to develop (0.190) relationships with the community (EN-intf-06)

(2) we have a gorge just at the back of us which […] is famous (0.310) not just *because of* its uh (0.220) high (0.750) uh sides *but* also for climbing and things like that (EN-intf-06)

The "because" in Example (1) introduces a subordinate clause with its own predicate ("because we're a comprehensive community") while being governed by the following main verb clause ("part of the funding is to develop relationships…"). In Example (2), however, the "because" in the prepositional phrase "because of" cannot be removed from the utterance (**the gorge is famous not just its high sides*), while "but" introduces another prepositional phrase without a predicate ("for climbing and things like that"). Some of these decisions are relatively specific to the research questions of this study and may therefore not directly suit other purposes. Another consequence is that the findings of this research are not completely comparable to other corpus-based studies using different identification criteria. However, such restrictions and exclusions are necessary to guarantee the consistency of the onomasiological identification procedure, provided they are motivated and documented.

DM tokens were identified entirely manually by one annotator (the author), bearing these criteria in mind while listening to the corpus recordings and reading the transcripts: any item in a given context that met the definition was selected and tagged as a DM. This bottom-up process did not resort to a closed list of pre-selected expressions (as is most often the case in other studies). The number and different types of DMs encountered in the corpus were therefore not limited nor planned in advance in any way.

The implementation of this definition to corpus data encountered the special case of "complex" DMs, i.e. more than one graphic and/or lexical unit co-occurring together as a grammaticalized, fixed form with a unique meaning. Diachronically, many present-day DMs originated from multi-word units (e.g. French *parce que*) and this fixation process might still be on-going for some contemporary DMs. The limit between mere co-occurrence and fixation is however subtle and partly based on frequency criteria: the more often two items appear jointly, the more fixed their respective position becomes. Another criterion is functional and states that it is neither possible nor relevant to assign a function to the elements of a complex DM taken separately (Waltereit 2007; Cuenca & Marín 2009; Crible 2015). Therefore, in a limited number of cases, such "complex" DMs were annotated as one item. In order to remain consistent during the final annotation round, a closed list of complex DMs was elaborated from the different testing phases on the pilot corpus: occurrences that met the criteria described above were selected and included in the closed list which was then used throughout the annotation of *DisFrEn*. The list comprises: English *and then*, French *mais bon*, *et puis*, *bon ben*, *eh ben* (and variants) and *ou sinon*.

Finally, testing phases as well as consideration of other proposals in the literature allowed me to identify borderline elements that are problematic to categorize, usually because they share some (but not all) characteristics of DMs as they are presently defined. These types of expressions, which are all specific to spoken language, have been explicitly addressed in the protocol, stating the theoretical reasons to exclude them from the category and the conditions under which some of them could be integrated. They consist in fillers (*uhm*, *euh*), interjections (*ah*, *Gosh* – sometimes included), answer particles (*yes*, *no* – sometimes included), epistemic parentheticals (*I think*), general extenders (*and so on* – sometimes included), tag questions (*isn't it*) and explicit editing terms (*sorry*, *I don't know*). Readers are referred to Crible (2014, 2017b) for more details and motivations.

4.2.2 Functional taxonomy

While the present annotation protocol covers several syntactic and contextual features of DMs (see next sections), its major contribution lies in the proposal of a functional taxonomy structured around four "domains" (Sweetser 1990) and thirty function values which were specifically designed for spoken language and in accordance with the definition of the category provided above (cf. Table 3.2). This taxonomy is best described as a combination of, on the one hand, the format and partial content of the PDTB's annotation guidelines (PDTB 2.0, Prasad et al. 2008) in terms of the operationalization of definitions and, on the other hand, the four-fold structure and speech-specific functions found in González (2005). I borrowed from the former the style of their definitions which are organized in a systematic way with clear terms and examples. I selected from the latter the function values that were missing from the PDTB 2.0 because these values occur only in spoken data. The taxonomy was designed in order to meet the balance between an extensive and exhaustive coverage of all possible functions of DMs in speech and, on the other hand, the intensive and operational definition of the different values in the taxonomy with no or little conceptual overlap between values. The following four domains and thirty functions have been selected:

- *ideational* domain: relations between real-world events; includes cause, consequence, contrast, concession, condition, alternative, temporal order, exception;
- *rhetorical* domain: relations between epistemic and speech-act events, and metadiscursive functions; includes motivation, conclusion, opposition, relevance, reformulation, approximation, comment, specification, emphasis;
- *sequential* domain: structuring of discourse segments; includes opening boundary, closing boundary, topic-resuming, topic-shifting, quoting, enumerating, punctuating, addition;

– *interpersonal* domain: interactive management of the speaker-hearer relationship; includes monitoring, face-saving, agreeing, disagreeing, ellipsis.

In this system, domains and functions are inter-dependent, insofar as one function value systematically belongs to a given domain and each domain contains a fixed number of possible function values. For instance, the relation of semantic cause, tagged *cause* in *DisFrEn*, belongs to the ideational domain, while its pragmatic equivalent *motivation* belongs to the rhetorical domain. This aspect constitutes the main difference with the PDTB 2.0 (and Zufferey & Degand's 2013 revised version) which places at the highest level four general meanings (i.e. *temporal*, *contingency*, *comparison* and *expansion*) which are then categorized as semantic or pragmatic (for some of them only; cf. Table 3.1). Although each system has its own pros and cons, the present approach in domains was chosen for its capacity to summarize the functions of DMs in a more informative way regarding the semantic-pragmatic distinction.

The definition of the four domains made heavy use of existing definitions in the literature and therefore did not lead to many revisions during the elaboration of the coding scheme. In contrast, defining the functions and categorizing them in a particular domain was a complex task, although many values were inspired by previous taxonomies. The major difficulty came from the adaptation of writing-based taxonomies to account for the particular characteristics of the spoken mode, which involved two types of revisions: either simplification of previous distinctions to avoid ambiguity and over-specification, or re-categorization of functions in different domains. For instance, in González (2005), "evidence" and "justification" are categorized in two separate domains (rhetorical and inferential, respectively) although the labels suggest a strong conceptual similarity.

Furthermore, earlier versions of the present protocol underwent several stages of similar revisions itself, in this case relying substantially on the annotations of the pilot corpus. These corpus-based revisions paid particular attention to making every decision explicit and replicable in the disambiguation process. Some major changes brought up by the testing phases and implemented in the final version of the protocol include a more detailed definition of all the values in the protocol, with additional criteria, prototypical paraphrases and examples, and the addition of two focus-sections in the guidelines dedicated to frequent polyfunctional DMs as they emerged from the pilot study and to the mapping of semantic and pragmatic equivalents (see Crible 2014).

Crible & Degand (in press) conducted an annotation experiment applying this taxonomy to samples of conversational data in French and English (55 tokens in each) and reported the following scores for inter-rater reliability: acceptable agreement for the identification of domains with $\kappa = 0.563$ (Fleiss' Kappa) and

70.9% of relative agreement between two expert coders; lower results for functions (κ = 0.406, 44.5%) but easily explained by the high number of possible values, the presence of rare values and the overall complexity of the task of pragmatic disambiguation (Spooren & Degand 2010). For these reasons, perfect agreement between annotators can never be achieved in qualitative annotations at discourse level, unlike in other tasks or linguistic domains (e.g. part-of-speech tagging) with fewer values and less ambiguity.

Intra-annotator agreement was also tested on a stratified sample representing one text per register per language (i.e. about 15% of the whole corpus in terms of duration and word count) The sample contained 1,194 instances of DMs (i.e. again about 15% of the whole dataset). For functional domains, the agreement is substantial (κ = 0.779, 84% of relative agreement) regardless of the particular value at stake. At the more fine-grained level of specific function values, the agreement is lower (κ = 0.74, 75.8% of relative agreement) but still substantial and much higher than the results for inter-annotator agreement. Functions with a larger proportion of disagreements than agreements are rare in the data (e.g. *comment*, *emphasis*); other notable problematic values are *opposition*, *contrast* and *consequence* which show around 40% of disagreements. Overall, the present state of the functional taxonomy remains challenging to annotate yet reliable enough to be used (after heavy training) in this research, bearing the necessary limitations in mind.

To sum up, the elaboration of this functional taxonomy followed a strict corpus-based methodology, with constant back-and-forth movement between theory and data, strongly rooted in the line of reference models (Halliday 1970, PDTB 2.0) and extensively tested on authentic data.

4.2.3 Three-fold positioning system

The next three variables are closely related and provide complementary information about the position of the DM. They differ in the size and type of unit that they refer to: the micro-syntactic unit, or minimal clause the DM belongs to; the macro-syntactic unit, or dependency structure with all its constituents; the turn-of-speech, a larger interactional unit defined by a change of speaker. The annotation of these three variables is independent and will be presented separately, starting with the position within the turn. The annotation of this variable uses a fairly straightforward system based on the exchange structure and the turn breaks as they are represented in the transcriptions. This feature, inspired by the Model for Discourse Marker Annotation (MDMA) project (Bolly et al. 2015), consists in four values: (absolute) turn-initial, (absolute) turn-final, turn-medial (any other position in the turn) and independent turn (when the DM constitutes the whole turn itself, including co-occurrences or repetitions of DMs).

Then, for the macro-syntactic position, I relied on the framework of Dependency Grammar developed by Tesnière (1959), with minor adjustments suggested by the terminology in German linguistics (Auer 1996; Lindström 2001). This level takes as a reference unit a main clause and all the subclauses or other adjuncts it governs (cf. Hunt's (1965) T-Unit adapted for speech by Foster et al. (2000)). The challenge of describing the position of DMs in traditional grammar terms is that most DMs do not occur within well-defined slots such as predicate, arguments and adjuncts, but mostly outside of them. I therefore chose to adopt a strictly linear and "topological" approach where no functional considerations are involved in the annotation of macro-syntactic position, leaving out distinctions such as governed vs. non-governed DMs, which partly overlap with ideational vs. rhetorical, respectively. Macro-syntactic position thus only locates the DM in five slots which are represented in Figure 4.1.

In this system, there is a first divide between elements comprised in the dependency structure (predicate and complements) and those outside of it, in the "periphery" of the utterance. Peripheries are subdivided in two slots depending on their respective position, viz. pre-field (initial position, "PRE") or post-field (final position, "POST"). Three governed slots are then distinguished within the scope of the main verb: left-integrated position ("LEFT"), that is, any integrated element before the main verb construction; middle field ("MID"), i.e. within the main verb construction; right-integrated ("RIGHT"), that is, any integrated element after the main verb construction.[17] Annotating the macro-syntactic position of DMs using this grid therefore consists in locating the DM in one of these five slots which are segmented based on syntactic considerations of dependency.

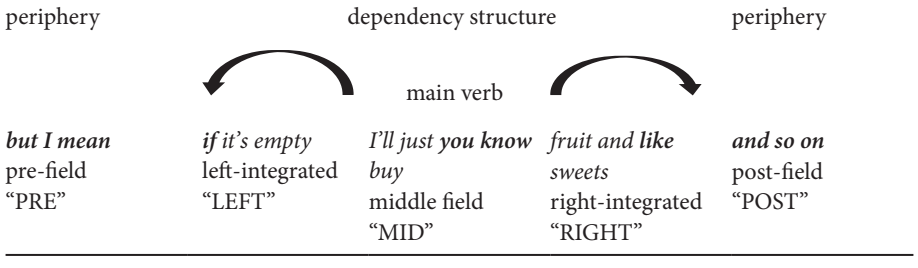


Figure 4.1 Macro-syntactic segmentation for DM position

17. The terms "left" and "right" are to be understood in a linear sense, with respect to the main predicative verb. This spatial terminology is somewhat inadequate to describe spoken language, but is used for reasons of consistency with the literature (e.g. Beeching & Detges 2014).

For instance, medial DMs such as "like" in Figure 4.1 are considered "right-integrated" since they occur within elements which are governed by the main verb ("I'll just buy fruit and sweets"). In other words, the position of the DM does not depend on whether or not the DM itself is governed or integrated in the dependency structure, but rather on whether the unit in which it occurs is governed or not. Macro-syntactic position is not related to the scope of the DM either: in the example, "like" targets "sweets", yet it is annotated with respect to the unit or slot it occurs in. Annotating DM scope, although potentially more insightful than this system of macro-syntactic position, is a highly complex task which includes functional considerations and is bound to lead to disagreements and ambiguous cases, especially in spoken data (cf. the long-distance relations discussed in Section 3.3.2.1). This second type of position therefore provides a detailed yet reliable view of the mobility of DMs that refrains from mixing different types of information (position and function are kept as distinct variables).[18] Detailed criteria and special cases are provided in the guidelines (Crible 2014), taking up the lessons from the tests on the pilot corpus.

The third type of position, in the micro-syntactic unit, is more straightforward and takes into consideration the position of the DM within its minimal syntactic unit, starting from subordinate clauses and larger. This variable provides useful information that completes the macro-syntactic variable, especially in cases where a DM is at the right of the governing verb ("right-integrated") but in initial position with respect to its own subclause, as in (3). This variable consists of five values: initial, medial (preceded and/or followed by non-optional elements), final, independent, interrupted.

(3) it's good for us *because* it puts us into a marketplace
(*Backbone* corpus, en011)

In this example, the *because*-clause depends on the main clause "it's good for us" to which it appears at the right (macro-syntax: "right-integrated") but this "because" is also initial with respect to the subclause it introduces ("it puts us into a marketplace"), hence "initial" in the micro-syntax. This flexibility in the annotation of DM position allows for more precision than a single-layer system by zooming in and out of the host-unit of the DM (from turn to dependency structure to clause or subclause). As opposed to other syntactic models that either take into account functional roles (e.g. the Val.Es.Co model, Estellés & Pons Bordería 2014; Blanche-Benveniste's (2003) proposal) or require heavy semi-automatic syntactic annotation (e.g. Basic Discourse Units, Degand & Simon 2009), this three-fold positioning system is both informative and operational, involving few theoretical notions and remaining independent from the annotation of DM functions.

18. For a proposal of DM annotation targeting scope (i.e. function and position mixed in one variable), see the Val.Es.Co segmentation model (Briz & Val.Es.Co Group 2003).

4.2.4 Other variables

Besides functions and positions, two other manually assigned variables are covered in the protocol. Firstly, a part-of-speech tag (POS-tag) was assigned to each DM, be it a single- or multi-word unit. In the latter case, only one tag was assigned to the whole expression. These tags do not refer to the syntactic behaviour of the expression in context, as traditional POS-tags do. Instead, they are allocated systematically (i.e. always the same tag for one DM expression) on an etymological and lexicographic basis: for instance, *so* is always labeled as an adverb, and *you know* is always a verb phrase, regardless of the way they are actually used in the data. This variable aims at documenting the grammatical diversity of the DM category in English and French. A similar approach is taken by Pitler & Nenkova (2009), who refer to this type of POS-tag as "self category": "the highest node in the tree which dominates the words in the connective but nothing else" (2009: 14). The list of tags is mostly borrowed from the PDTB's guidelines in Santorini (1990). The final list of POS tags, restricted to values that can apply to DMs based on the pilot corpus study, can be found in Table 4.2 with examples.

Table 4.2 List of all part-of-speech tags for DMs

CC	coordinating conjunction	*and*, *but*, *or*…	*et, mais, ou*…
RB	adverb	*so*, *actually*, *now*…	*donc*, *enfin*, *alors*…
VP	verbal phrase	*you know*, *I mean*…	*tu vois*, *je veux dire*…
SC	subordinating conjunction	*because*, *if*, *although*…	*parce que*, *même si*…
WP	pronoun	–	*quoi*, *un*, *et tout*
NN	noun phrase	*sort of*	*genre*
JJ	adjective	*right*	*bon*
PP	prepositional phrase	*in fact*, *for example*…	*au fond*, *par contre*…
UH	interjection	*okay*, *yeah*, *oh*…	*hein*, *ben*, *ouais*…

Cuenca (2013) presents an alternative approach to DM categories by taking into consideration both syntactic and functional features. She distinguishes between conjunctions, parenthetical connectives, pragmatic connectives, interjections and modal markers. This cognitive-functional approach is more explanatory and economical than the nine POS-tags showed in Table 4.2. However, Cuenca's (2013) model combines different types of features and considerations into one complex variable, while the present approach strives to keep variables as independent as possible for more powerful (statistical) analyses.

Secondly, the last variable at DM level is a contextual feature that accounts for the immediately contiguous presence of another DM (according to the same definition). In the case of co-occurrence, the annotation specifies the periphery

in which the other DM appears (left, right or both), following the MDMA model (Bolly et al. 2015).

In conclusion for this protocol, the fact that many authors have tried to describe DMs illustrates the discrepancy between the complex mechanisms responsible for language production and the rigidity of corpus annotation. The present proposal hopefully contributes to this issue by striving to respect both the intrinsic nature of language and the categorizing needs of linguistic description, in line with the ambitions of cognitive pragmatics (Schmid 2012).

4.2.5 Annotation procedure

4.2.5.1 *Software*

The annotation of *DisFrEn* was conducted under the EXMARaLDA annotation tool (Schmidt & Wörner 2012), an open-source software package designed for multi-layered annotation of spoken data with enriched metadata. Its annotation interface Partitur Editor makes it possible to manually or semi-automatically encode annotations over many different layers applying to different cell sizes, with the possibility to merge several cells together, as is the case in Figure 4.2, where the DMs and pauses are merged into one cell labeled DM+UP.

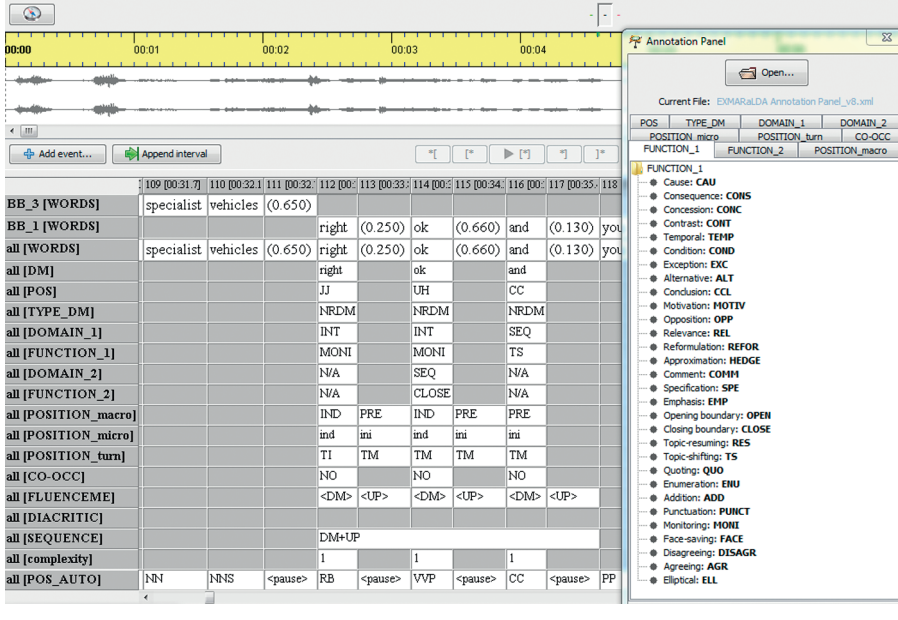


Figure 4.2 Partitur Editor annotation interface

4.2.5.2 *Disambiguation method*

The annotation protocol for DMs specifies a number of conventions regarding the disambiguation procedure. The annotation was only carried out once by a single annotator (the author) for the whole corpus, given the time cost and expertise it requires (cf. Section 4.2.2 for an assessment of replicability). The resort to the audio file was systematic and necessary, in line with the finding in Bolly & Crible (2015) that it significantly improves the accuracy of the annotation (see also Zufferey & Popescu-Belis 2004). In addition, any disambiguation technique was used to resolve functional ambiguity, with no particular instruction or restriction: anything helpful in context is welcome, be it substitution tests, translation equivalents, or the criteria in the protocol itself.

Up to two function values can be assigned for each DM, when a particular DM appears to express two functions, either from the same domain and type or from two different ones. This option is not meant as a solution to ambiguous cases (which should be resolved as much as possible) but for the quite frequent cases of multifunctional DMs. Simultaneous functions can be equally salient or not, but for operationalization purposes such a distinction is not made in *DisFrEn*. In fact, it is not always relevant to determine which function prevails over the other, and whether there is such prevalence at all: "no one function is necessarily predominant in a particular context" (Brinton 1996: 35).

4.3 Disfluency annotation

The flourishing literature on (dis)fluency results in a panel of annotation protocols, which are however rarely comparable or generalizable to data of a different type. While a componential approach to (dis)fluency is generally shared amongst authors (e.g. Shriberg 1994; Götz 2013; Moniz 2013), the scopes and formats of the annotation often differ. More specifically, the differences between frameworks include the number and categories of observed phenomena, data type (languages and modalities), technical choices such as labels and extraction method, and possibly others. Overall, most protocols present a number of drawbacks, be it on practical aspects (replicability of the annotation, efficiency of the quantitative treatment) or theoretical ones (validity of the categories, robustness of the criteria, cognitive-pragmatic relevance of the model).

In this perspective, the protocol described here (Crible et al. 2016) is a proposal to address some of these issues with a highly flexible, multilingual and multimodal approach to fluencemes. This work is collaborative (with A. Dumont, I. Grosman and I. Notarrigo) and benefits from the input of various frameworks,

thus overcoming methodological and theoretical monism. This section does not address the annotation of repair categories, which follows a more qualitative methodology based on Levelt (1983) and which is described in Chapter 8, where it is used.

4.3.1 Simple fluencemes

Simple fluencemes are composed of only one part (which can itself be a phrase in the case of discourse markers and editing terms). These phenomena can occur in isolation, juxtaposed with another, or embedded in compound fluencemes.

4.3.1.1 *Silent pauses*

The first simple category is that of silent pauses (tagged "UP"), defined by an interruption of the sound signal lasting more than 200 milliseconds, following Candéa (2000). This threshold is fixed and does not take account of speaking rate or speaking style variation, due to the very limited potential of *DisFrEn* for prosodic analysis. No distinction is made between the duration of silent pauses, be it as a continuous (seconds) or discrete variable (categories of length), following Little et al. (2013). Silent pauses in *DisFrEn* will not be investigated any further than their presence and surrounding context. More thorough prosodic analyses have not been pursued.

4.3.1.2 *Filled pauses*

Filled pauses ("FP") consist in vocalizations characterized by their conventional and neutral phonetic form (e.g. "euh" in French) and their function as supporting or maintaining on-going speech (e.g. Clark & Fox Tree 2002). Since final-vowel lengthenings are not annotated in *DisFrEn*, they have been categorized as filled pauses when hesitation was possible (especially for final schwa in French). This definition of filled pauses excludes backchannelling devices usually transcribed as "hm hm" or "mm". In the English data, spelling variation was reduced to the two forms "uh" and "uhm", replacing other variants (e.g. "er") when necessary.

4.3.1.3 *Explicit editing terms*

Explicit editing terms ("ET") cover any lexical expression by which the speaker signals some production trouble and which are not identified as DMs or filled pauses, such as *what is it* or French *comment* in certain contexts. Editing terms are only annotated in the vicinity of other fluencemes. The difference between DMs and explicit editing terms can be subtle and relies on the following criteria: editing terms must be explicit references to lexical access trouble, with a low grammaticalization degree (free juxtapositions and semantic transparency) and must have propositional content. Borderline cases are phrases like *if you will*, *I don't know* or French *je dirais* 'I would say', showing a high degree of fixation but directly referring

to the act of speaking or thinking. These will be considered DMs if they meet the criteria for this categorization in context.

4.3.1.4 *False-starts*

False-starts ("FS") are interruptions that leave a segment syntactically and/or semantically incomplete and where no elements from the previous, abandoned context are taken up in what follows (Pallaud et al. 2013), as in Example (4).

(4) for women possibly to *have* you know (0.231) they're getting (bb_en014)

Only the last word of the interrupted is labeled "FS" ("have" in this example). If any lemma is repeated (even modified) in the next segment, it is categorized as a repetition and/or substitution (see below).

4.3.1.5 *Truncations*

Finally, truncations ("TR") are interruptions that only apply to words and not segments as in false-starts (Example 5).

(5) so *wh-* how often do you play (EN-conv-02)

If the fragments are repeated and/or completed, the truncation becomes a compound fluenceme, since it becomes structured into several parts. As soon as the first phoneme of a truncation is repeated within the next words, it is considered completed, unless there is clear evidence to the contrary in the audio context. A truncation can be completed after the insertion of other fluencemes, as in Example (6).

(6) and *po-* after that it's partnership (bb_en009)

This example shows a case of compound truncation where only the first phoneme /p/ is repeated and completed after lexical insertions (see Section 4.3.3).

4.3.2 Compound fluencemes

Compound fluencemes function with a structure in at least two parts, namely the *reparandum* and the *reparans* in Levelt's (1983) or Shriberg's (1994) terms. Compound fluencemes include two types of repetitions and two types of substitutions (as well as completed truncations).

4.3.2.1 *Identical repetitions*

Identical repetitions ("RI") include any words formally similar to each other and directly contiguous, whether intentionally (e.g. because of an overlap, as in the constructed Example 7) or not (Example 8), so that we avoid any judgment as to their function and relative fluency at this stage.

(7) <spk1> I've been to [*see my*] *see my* grandmother
<spk2> [where?]

(8) it's just you know *the the the* qualities that spring to mind (EN-conv-03)

The only exclusion is the case of semantic repetitions which have some propositional content, usually in the form of an intensification (as in *I'm very very happy*). Repetitions can only apply to complete lexical elements. Truncated words and filled pauses are not included.

4.3.2.2 *Modified repetitions*

Modified repetitions ("RM") cover words belonging to a segment that is partially repeated but with a change in content, either by a substitution, a truncation, a deletion, or a lexical insertion, as in (9) where "tour" is inserted to specify a type of "coach".

(9) from the coach from the from the *tour c- tour* coach (EN-intf-02)

This type of repetition is thus less strict than the previous one since it admits syntactic-semantic modifications. It is very often found in the context of substitutions.

4.3.2.3 *Morphosyntactic substitutions*

Morphosyntactic substitutions ("SM") correspond to any morphological modification in a complete lemma (excluding truncations), be it an addition or deletion of a morpheme such as number marking or elisions. They often involve modified repetitions, as in (10).

(10) well I *wasn't* driving well I *was* driving partly on the road (EN-phon-02)

4.3.2.4 *Propositional substitutions*

Finally, propositional substitutions ("SP") correspond to any segment replaced by another one which introduces a semantic nuance or modification. The difference between false-starts and propositional substitutions lies in the fact that the *reparans* of a SP is the continuation of the previous utterance as in (11), while the segment next to a FS has no syntactic connection with the previous one.

(11) anything that *will* (0.200) *could* possibly go wrong (EN-intf-02)

All these definitions strive towards a purely formal and objective approach to fluencemes that does not require an interpretation of relative fluency or disfluency of the annotated segment. As a result, this protocol covers more phenomena than those traditionally included in other frameworks, with no additional complexity

for the annotators. The flexibility of our approach builds on the identification of reliable surface features that considerably minimise subjective considerations of semantic-pragmatic interpretation in the annotation process. In our view, this precaution is necessary since it keeps the different analytical steps (i.e. annotation, hypothesis-testing, interpretation) separate and independent, thus vouching for the methodological soundness of this approach.

4.3.3 Related phenomena and diacritics

Other categories are defined in the annotation protocol which are not fluencemes but related phenomena that either apply to an existing fluenceme (diacritics) or participate in their structure (insertions and deletions).

Lexical insertions ("IL") are propositional elements integrated into modified repetitions or truncations. They modify the content and are sometimes the very motivation for the repetition or truncation, as in (12).

(12) the monitors go off wh- *even* when we put our hands in (EN-intf-03)

Parenthetical insertions ("IP") are propositional segments functioning as a "parenthetical aside" (Shriberg 1994: 61) located in a sequence of fluencemes to which it adds some background information without directly modifying the content of the utterance (Example 13).

(13) I was sort of saving it for a rainy day and the rainy (0.250) well *touch wood* the rainy day's never come (EN-conv-07)

As we can see in this example, parenthetical insertions are not syntactically integrated, which is the main difference with lexical insertions. Another difference is their secondary informational status.

Deletions ("DE") mark the removal of a propositional element and induce a change of content (Example 14).

(14) you *can* do A lev- you do A-levels in music and dance (EN-intf-06)

Diacritics form the last category of annotated items. Diacritics only apply to fluencemes and cannot be annotated on their own. In *DisFrEn*, three categories have been used: "within", misarticulations and change of order. First, the "WI" tag (for "within") is applied to any simple fluenceme occurring within the structure of a compound fluenceme (e.g. a pause within a repetition, a discourse marker within a completed truncation). This information distinguishes isolated vs. embedded contexts of simple fluencemes such as pauses or DMs.

Misarticulations ("AR") apply to any element identified by the speaker as different from a "correct" pronunciation according to their own standard. It must be explicitly noticed by the speaker in the form of a fluenceme (editing term, DM, etc.) as in (15), otherwise it is not annotated in order to avoid any reference to a norm.

(15) any *uninamity* (0.413) any unanimity (EN-news-07)

Lastly, changes of order ("OR") indicate the syntagmatic re-ordering of repeated elements otherwise identical, with no propositional change (Example 16).

(16) *normally would take you* before you're a fully qualified solicitor *would normally take you* a minimum of (bb_en009)

All these additional phenomena can be the object of particular research questions but mostly serve to complete the description of fluencemes in context, in order to be exhaustive and make finer distinctions between different types of repetitions or substitutions. More specifically, they can prove very useful in the analysis by pointing to surface features that potentially explain the different patterns observed for the same type of fluenceme and their relative (dis)fluency rating.

4.3.4 Annotation procedure

4.3.4.1 *Technical format*

Categories of (dis)fluent phenomena have been extensively studied in the past twenty years, including in corpus-based studies. The content of the present protocol is strongly based on this prior work, borrowing many definitions from the literature (albeit with a number of revisions). The originality of our protocol therefore lie in its technical and quantitative treatment of the internal structure of fluencemes.

Firstly, in a sequence of fluencemes (i.e. a span of text covered by one or several fluencemes), all annotations are assigned at word level, with tags for every graphic unit categorized as fluenceme. Each annotated word has a two-letter tag corresponding to a type of fluenceme (e.g. "DM" for discourse marker). If this word is the sole element of the fluenceme, it will get opening and closing brackets such as "<DM>", thus marking its simple structure. However, if the fluenceme is complex and comprises several elements, the presence and side of the bracket will specify the position of the word in the internal structure. In addition, numbers can be added to tags in the case of compound fluencemes to identify their different parts (more details and examples can be found in Crible 2017b).

This flexible system combining letters, brackets and numbers makes it possible to account for very complex patterns of different sizes and types, with embedded phenomena and multiple tags for the same word. More details and examples can

be found in Crible (2017b) and in the annotation protocol (Crible et al. 2016). The guidelines also specify in detail the criteria for all fluencemes in different contexts, with some examples of problematic cases found while testing this protocol on different corpora.

4.3.4.2 *Scope of the disfluency annotation*

In all subcorpora except for radio interviews and face-to-face interviews, only sequences containing at least one DM were annotated. In other words, each time a DM was identified, all fluencemes in its context were tagged until no other adjacent fluenceme could be found. In the interviews data, all fluencemes were systematically identified regardless of the presence of a DM. In these two fully annotated subcorpora (radio and face-to-face interviews), information about the position of silent and filled pauses has been added, following a coding scheme similar to that of the three-fold position of DMs. This syntactic information was used for the analysis of clusters of discourse markers and pauses, carried out by Crible et al. (2017).

4.3.4.3 *Replicability of the disfluency annotation*

To evaluate the replicability of this protocol, inter-annotator agreement was computed on a sample of about 7,000 words of French radio interviews which was annotated independently by two (expert) annotators following the same guidelines.[19] We reached an agreement of $\kappa = 0.67$ which includes disagreements on boundaries and identification in addition to disagreements on fluenceme types. When restricting the dataset to "true positive" cases, i.e. words that were tagged by both annotators (although not necessarily with the same fluenceme type), the agreement increases to $\kappa = 0.79$. Given that the kappa-metric is sensitive to rare values, we simplified the dataset by excluding very rare labels (fewer than 10 occurrences). With this simplification, the agreement reaches $\kappa = 0.82$ on "true positives" (i.e. when both annotators have assigned a label).

All these scores range from "substantial" to "almost perfect" according to recognized scales (e.g. McHugh 2012), which is very encouraging and reflects well on the operationality of the guidelines. A more thorough analysis of inter-annotator agreement for the fluenceme level is provided in Crible (2017b).

---

19. I wish to thank my colleague Iulia Grosman, who was the second annotator and who carried out the statistical analysis reported in this section.

4.3.5 Macro-labels of sequences

Once the annotations were extracted from the corpus, a number of modifications were made to filter the many values of certain variables and summarize the annotations in different ways. Most modifications are related to fluenceme sequences and involve not only practical aspects of merging and summarizing but also more conceptual considerations for the design of valid and theoretically relevant categories. These new variables are therefore called macro-labels because of their categorizing function, beyond purely technical purposes.

One such macro-label is referred to as "sequence category" in the dataset and narrows the number of sequence types to only six possible values which are defined by roughly grouping the fluencemes they contain by complexity and function. These macro-labels reflect the focus on DMs in this research and are hierarchically ordered in terms of their impact on the linguistic context. All types, except for the first level, can include the fluencemes of other "inferior" types. The values are:

- D – the sequence contains only discourse marker(s);
- P – the sequence contains (silent and/or filled) pauses and may contain DMs;
- F – the sequence contains truncations and/or false-starts and may include the contents of "D" or "P";
- R – the sequence contains identical and/or modified repetitions and may include the contents of "D" or "P";
- S – the sequence contains propositional and/or morphological substitutions and may include the contents of "D", "P" or "R";
- Z – the sequence includes the combination of "F" with "S" and/or "R", and may include the contents of "D" and "P".

Examples (17)–(19) illustrate cases of F-, S- and Z-sequences, respectively.

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

(18) is point (0.680) is the regeneration then just in terms of the schools *or is (0.400) are* other projects (0.940) uhm being undertaken (EN-intf-05)

(19) they're *all uh and so the the the* joke is that (EN-clas-05)

In (17), the F-sequence includes a false-start on "might", a DM "for example" and a filled pause "uhm". The S-sequence in (18) includes a DM "or", a morphological substitution of "is" by "are" and a silent pause. Example (19) is a Z-sequence containing a false-start on "all", a filled pause "uh", two DMs "and" and "so" and an identical repetition of "the".

In a slightly different perspective, the second set of macro-labels describes the internal structure of the elements in a sequence and looks at three types of information: (1) whether the sequence contains simple or compound fluencemes; (2) whether the sequence contains one or several fluencemes; (3) whether the sequence containing compound fluencemes also contains simple fluencemes, and the position of the latter with respect to the former. This category has 10 different values that cover any type of sequence. The values and examples are provided in Table 4.3.

Table 4.3 Macro-labels for the internal structure of the sequence

one simple	<DM>
multiple simple	<DM> <UP> <FP>
one compound	<RI0 RI1>
one compound with embedded simple (WI)	<RI0 <DM> RI1>
one compound with peripheral simple (PE)	<UP> <RI0 RI1>
one compound with WI + PE simple	<UP> <RI0 <DM> RI1>
multiple compound	<RM0 <SP0 RM1> SP1>
multiple compound with WI	<RM0 <SP0 <UP> RM1> SP1>
multiple compound with PE	<DM> <RM0 <SP0 RM1> SP1>
multiple compound with WI + PE	<DM> <RM0 <SP0 <UP> RM1> SP1>

Finally, a three-fold category called "cluster" applies to DMs and indicates whether they form a sequence by themselves ("alone"), a sequence with other DMs and no other fluencemes ("with DM"), or a sequence with other types of fluencemes ("in sequence"). This variable offers a broad filter, a first answer to the hypothesis that DMs occur more frequently in sequences than in isolation.

4.4 Summary

In this chapter, the data and methodology of the present research have been presented in detail, with its strong corpus-based foundation. The key points of this chapter are the following:

- the comparable design of *DisFrEn*, balancing eight registers across English and French and amounting to 161,700 words and 15 hours of recordings;
- the definition of discourse markers and its bottom-up application to corpus data;

- the operationalization of variables describing the syntactic and pragmatic behavior of DMs, with a particular emphasis on a functional taxonomy specifically designed for spoken DMs and covering thirty values grouped in four domains;
- the word-level annotation of fluencemes, reproducing with great precision the internal structure of complex sequences;
- the assessment of these two annotation protocols by annotation experiments showing satisfactory inter- and intra-annotator agreement.

The contribution of this dataset lies in the rich annotations that were manually added to the original texts, following innovative yet operational procedures. Shortcomings are mostly due to practical considerations which, as we know, often interfere with theoretical ambitions in empirical studies. Nevertheless, the numerous revisions based on a pilot study as well as consideration of the literature and discussion with experts in the field of discourse annotation, reduce the number of major pitfalls and make *DisFrEn* a reliable – if relatively small – dataset for the study of discourse markers as fluencemes.

CHAPTER 5

Portraying the category of discourse markers

This first analytical chapter reports on corpus-based results regarding the syntactic and pragmatic behavior of DMs across registers in English and French. Starting from individual variables extracted from the annotations, it progressively incorporates information from multiple sources (frequency, language and register variation, form-function patterning) in order to draw an exhaustive portrait of the DM category, thus meeting the ambition of exploratory research and answering some of the hypotheses laid out in Section 3.5.

The overall frequency of DMs will first be compared across languages and registers (Section 5.1). Syntactic position will then be investigated in order to test the extent to which initiality is indeed a representative criterion for the whole DM category (Section 5.2). The two functional variables (domain and function) will be individually detailed (Section 5.3) and mapped onto register and positional preferences (Section 5.4). The different configurations of co-occurring DMs are examined in Section 5.5 in combination with both syntax and functions. Finally Sections 5.6 and 5.7 summarize and discuss the main results of this chapter.

5.1 Distribution across languages and registers

Due to the lack of large-scale contrastive research on DMs in spoken English and French, no hypothesis on quantitative differences were formulated. This gap in the field was explained by the profusion of contrastive case studies examining restricted groups of DM expressions in multilingual data, a limitation which can now be addressed by the present categorical approach to English and French DMs. Frequency of DMs by register, on the other hand, is highly documented and major effects of the degrees of preparation and interactivity are expected, following hypotheses on fluencemes in general and DMs in particular: high frequency and variety of DMs are associated with spontaneous discourse and interactive registers, given the role of DMs in planning processes and interpersonal strategies of exchange management.

5.1.1 General frequency

In *DisFrEn*, 8,743 DMs were identified and annotated. Table 5.1 reports on the distribution of DMs in the *DisFrEn* corpus in raw and relative frequency per thousand words (henceforth ptw).[20] A first general observation concerns the higher frequency of DMs in French than in English (about 60 DMs ptw in French, 49 in English). A test of log-likelihood (henceforth LL) shows that this difference is statistically significant, with a score largely above the admitted 3.84 threshold for $p < 0.05$ ($LL = 101.76$, $p < 0.01$).

Table 5.1 Raw and relative frequency of DMs by language and register

	English		French		Total	
	DMs	**ptw**	**DMs**	**ptw**	**DMs**	**ptw**
conversation	954	54.58	1520	87.20	2474	70.87
phone	609	62.48	530	78.14	1139	68.91
interview	1069	62.68	1299	71.99	2368	67.47
classroom	517	54.85	188	50.50	705	53.62
radio	479	54.60	441	52.40	920	53.52
sports	330	40.06	246	39.18	576	39.68
political	193	22.31	158	20.19	351	21.31
news	98	13.91	112	16.50	210	15.18
Total	**4249**	**49.17**	**4494**	**59.69**	**8743**	**54.07**

We see that the overall frequency of DMs across all eight registers is rather high, which suggests a highly pervasive and prominent use of DMs in spoken language. Given the wide coverage of the present DM annotation, this result is hardly comparable with previous works which usually target a restricted number of DM expressions (speech-based studies) or do not include non-relational, interactive uses of DMs (writing-based studies). Even González (2005) only reports frequency information for a selection of DMs in English and Catalan, despite her broad definition of the category and inclusive functional taxonomy (cf. Section 3.2).

20. Relative frequency is sometimes called "normalized frequency". The basis for normalization is usually either per thousand or per million words. Following the standard recommendation in corpus linguistics to use a common base which is closest to the corpus size (e.g. Biber et al. 1998: 264), results will here be reported in frequency per thousand words.

5.1.2 The status of tag questions

One possible explanation for the observed difference between English and French regards a theoretical and methodological decision on definition and identification of DM tokens. I mentioned in Section 4.2.1 some exclusions from the DM category, one of them being tag questions such as *isn't it*. Previous studies on tag questions have shown their relatively high frequency – with 11.26 occurrences ptw in English in Gómez Gónzalez (2014), for instance – and functional similarity with DMs. For example, Reese & Asher (2007) provide an analysis of the prosody and functions of tag questions under the Segmented Discourse Representation Theory (Asher & Lascarides 2003), which was originally developed for discourse relations and is still currently used in DM studies (e.g. Urgelles-Coll 2012). Furthermore, Pichler (2016) worked on the phonologically reduced tag *innit*, for which she found occurrences in utterance-initial position, supporting her classification of this form as a discourse-pragmatic variable.

However, tag questions were excluded from the present approach to DMs since they do not meet the syntactic criterion of fixedness: the form of tag questions varies depending on the main verb construction, tense and polarity of the utterance they are attached to (e.g. *isn't it*, *do you*, *wasn't she*), as opposed to the invariability of DMs and their independence from the syntactic structure. In addition, it is not always clear whether a tag question is used pragmatically as a discourse marker, checking for the interlocutor's attention, or whether it is an actual propositional question asking for an answer. This fuzzy distinction would have made it challenging to reliably identify discursive uses of tag questions in the present corpus-based approach. The fact remains that the inclusion of tag questions in the DM category would have considerably affected (i.e. smoothed out or even reversed) the frequency results and quantitative difference noted above.

5.1.3 Register variation

Table 5.1 also provides some elements supporting the hypothesis on register variation and the impact of preparation and interactivity. We see that the overall ranking (English and French combined) confirms the hypothesis, with the highest relative frequency in conversational genres (private conversations and phone calls, around 70 DMs ptw overall), closely followed by interviews (67 DMs ptw) and, to a lesser extent, classroom lessons and radio interviews (54 DMs ptw). This result seems to support Brinton's (1996: 33) association between DMs and "the informality of oral discourse and the grammatical 'fragmentation' caused by the lack of planning time". The temporal on-line nature of speech is both reflected and supported by time-buying devices such as DMs, thanks to their automaticity and limited

production cost for working memory. The overall distribution of DMs across registers seems to follow a decreasing cline from informal to increasingly formal contexts, which tends to corroborate the connection between DMs and informality. However, the situation is not as straightforward as it may appear when considering situational features instead of registers, as can be seen in Table 5.2.

Table 5.2 Distribution of DMs across degrees of preparation and interactivity

	English		French		Total	
	DMs	ptw	DMs	ptw	DMs	ptw
Preparation						
spontaneous	1893	29.95	2296	40.31	4189	34.86
semi-prepared	2065	58.58	1928	63.88	3993	61.02
prepared	291	18.54	270	18.48	561	18.51
Interactivity						
interactive	1563	57.41	2240	77.43	3803	67.72
semi-interactive	2065	58.58	1740	65.76	3805	61.66
non-interactive	621	25.95	514	25.83	1135	25.89

The most striking difference with Table 5.1 is that DMs are no longer the most frequent in spontaneous settings (conversations, phone calls and sports commentaries) but rather in semi-prepared registers (interviews and classroom lessons), which points to the special effect of intermediary contextual features, as suggested by the register hypothesis for general fluency. Speaking tasks such as interviews combine an intermediate degree of preparation (especially low for the interviewee) with a heightened attention for self-monitoring, which results in an increase of interruptions, reformulations and speech-supporting devices (cf. Broen & Siegel 1972; Section 2.5.3).

Interactivity, on the other hand, complies with the hypothesis of decreasing frequency (i.e. the less interactive the setting, the fewer DMs) overall and in French, while the difference between interactive and semi-interactive registers is small and non-significant in English. Table 5.1 already showed that English DMs are most frequent in interviews and phone calls (63 DMs ptw), followed by conversations, classroom lessons and radio interviews (55 DMs ptw), which contradicts the cline of formality observed in French and when both languages are combined. This first crosslinguistic observation suggests a stronger impact of interactivity on DM use in French, whereas this feature is only relevant to set apart non-interactive settings (e.g. sports commentaries) from (semi-)interactive ones in English. One should note that these contrastive effects are significant for the *DisFrEn* data but might not necessarily be generalizable to English and French as a whole, given the limitations of corpus comparability mentioned earlier (cf. Section 4.1.3).

5.1.4 A greater effect of register over language?

A number of additional crosslinguistic differences can be observed for register variation in Table 5.1, where a clear divide appears between, on the one hand, considerable gaps in frequency across English and French in the top three registers (conversations, phone calls, face-to-face interviews) and, on the other, a striking similarity between the remaining five registers and their lower frequencies of DMs. The difference between English and French in the former is always significant and in favor of French (LL = 132.17, 14.07 and 11.3, $p < 0.001$ for conversations, phone calls and interviews, respectively). The preference is less clear for registers with lower frequencies of DMs: no difference is significant and DMs are only slightly more frequent in English for all these speaking tasks except news broadcasts.

Such a quantitative similarity stands in contrast with Kunz & Lapshinova-Koltunski (2015), who found a greater impact of language (German vs. English) over register (e.g. fictional texts, corporate websites, academic speeches, interviews) on the frequency of discourse relations. In *DisFrEn*, both language and register seem to have an effect on DM distribution, either simultaneously (e.g. DMs are more frequent in conversations than in phone calls and more frequent in French conversations than in English conversations) or separately (e.g. DMs are equally frequent in English and French sports commentaries; DMs are equally frequent in English conversations and classroom lessons). The similarity of the English and French data in *DisFrEn* will be illustrated in many ways throughout the following analyses.

Similarities can also be observed within each language, especially in English where three out of eight registers show an equal relative frequency of DMs around 55 occurrences ptw, namely in conversations, classroom lessons and radio interviews. This finding suggests a partial agreement with Kunz & Lapshinova-Koltunski (2015): register variation is not always a relevant factor in DM distribution, especially in English. However, languages are not always sharply distinguished (cf. the five registers with low frequencies of DMs) and French DMs vary greatly under the effect of register, as already shown for the feature of interactivity.

5.1.5 DM expressions in contrast

To get a more concrete grasp of the data and the extent to which English and French differ, we can zoom in on the most frequent DM expressions, where both differences and commonalities can be found across languages and registers. The top five DMs are semantically and pragmatically equivalent in English and French, as can be seen in Table 5.3 with all registers combined, and relatively stable across registers, at least for some items (see Appendix 1 for the same table with register information).

Table 5.3 Top five most frequent DMs in English and French

	English	French
(1)	*and*	*et* 'and'
(2)	*but*	*mais* 'but'
(3)	*so*	*donc* 'so'
(4)	*well*	*alors* 'well/then'
(5)	*you know*	*hein* 'right'

In English, the generic conjunction *and* is invariably the most frequent DM across all registers except in phone calls where it is slightly topped by *but*. These two DMs are always included in the top five of all registers, usually as first and second most frequent expressions. In French, we find a similar prevalence of *et* 'and' in all registers with the same exception of phone calls (3rd position), where it is considerably less frequent than *donc* 'so' and *alors* 'so/well'. Another resemblance with English concerns *mais* 'but', which is particularly prominent in conversations and interviews and generally enters the top five of most registers (except for its 6th position in phone calls).

We see that the most frequent English and French DMs are not only semantic-pragmatic equivalents, but they also follow the exact same ranking. Boula de Mareüil et al.'s (2013) ranking is confirmed by the presence of *et, mais* and *alors* in this top-five. Many more observations could be made regarding DMs which are shared across or specific to a particular language and/or register. For instance, French *quoi* 'you know' is almost only used in conversations (216 occurrences out of 239), which points to its interactive function of sharing knowledge or perspective (Chanet 2001). Beeching (2007) further indicates that *quoi* is rather stigmatized as youth talk yet conveys a sense of solidarity between speakers, which is consistent with its frequency in the casual register of private conversation.

Another interesting observation concerns the subordinating conjunction *if* and its French equivalent *si*. They are respectively the third and second most frequent DM in political speeches, although only sixth and tenth in the general ranking across all registers, which could reflect politicians' tendency to make hypothetical and causal assertions. Similarly, some DMs such as *indeed*, *however*, *for* or *meanwhile* are more frequent in news broadcasts and political speeches than in any other register (although quite rare overall) and can therefore be considered as formal DMs. All DMs found in *DisFrEn*, their frequency and annotated functions can be found in Appendix 2.

5.1.6 Diversity hypothesis

So far, the hypothesis of higher frequency in spontaneous and interactive registers has been confirmed, with the nuance brought about by intermediary settings such as interviews (especially in English). However, the second aspect of this hypothesis, which not only concerns frequency but also diversity, is not confirmed by the data, as we can see in Table 5.4, where the ratio of DM types by DM occurrences or tokens (type-token ratio) is reported across languages and registers.

Table 5.4 Type-token ratio of DMs

	English		French	
	DM types	**Ratio**	**DM types**	**Ratio**
conversation	59	6.18	74	4.87
phone	41	6.73	45	8.49
interview	50	4.68	77	5.93
classroom	51	9.86	39	20.74
radio	42	8.77	54	12.24
sports	23	6.97	28	11.38
political	38	19.69	29	18.35
news	21	21.43	23	20.54
Total	**40.63**	**10.54**	**46.13**	**12.82**

We see that high frequency is not associated with high diversity, but rather the contrary: registers with small numbers of DM tokens (political, news) show the highest ratio of DM types, which reflects the high degree of planning in these speaking tasks and the resulting ability of speakers to vary their discourse-structuring devices, as opposed to more spontaneous discourse where the same multi-purpose DMs are often repeated (cf. the lowest type-token ratio for conversations, interviews and phone calls). French classroom lessons stand out with a particularly high ratio, in comparison to English classroom lessons and to other intermediary registers. This result may be due to cultural differences, such as a more formal academic style in French and a more interactive one in English. However, it is hardly interpretable given the low size of this subcorpus (cf. Section 4.1.3), which might over-generalize the observed frequencies.

Although the hypothesis of diversity is not confirmed at the level of the particular DM expressions, it is well attested at the level of grammatical categories, namely part-of-speech tags (henceforth POS-tags). Figures 5.1 and 5.2 represent the proportions of POS-tags in news broadcasts and conversations, in the two languages combined.

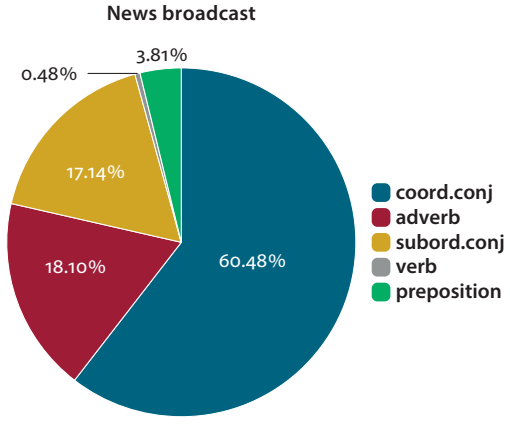


Figure 5.1 Proportions of part-of-speech tags in news broadcasts

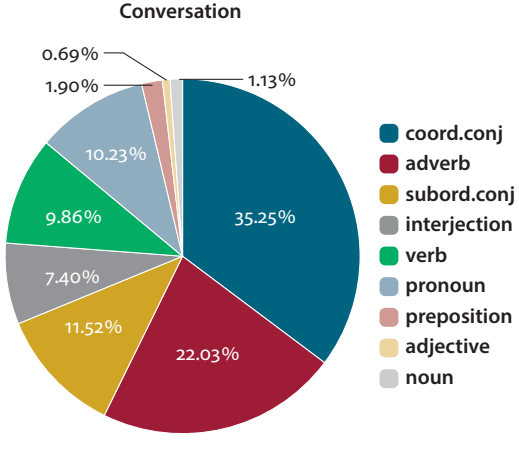


Figure 5.2 Proportions of part-of-speech tags in conversations

We see that these two registers, which stand on opposite ends in terms of DM frequency (cf. Table 5.1), are also contrasted in grammatical diversity of the DM category. Only five different POS-tags are used in news broadcasts, with an overwhelming majority of coordinating conjunctions, against nine types in conversation, where conjunctions take up a much smaller proportion. Still, coordinating conjunctions, mostly represented by *and* / *et* and *but* / *mais*, do appear as the most frequent class of DMs, followed by the much lower proportions of adverbs and subordinating conjunctions.

Overall, at this first level of observation, English and French do not strongly differ in terms of distribution and most frequent DMs, which confirms previous

contrastive research (e.g. Zufferey & Cartoni 2012; Dupont 2015). In line with this literature, differences are expected to be found at more subtle levels of analysis, i.e. when considering more qualitative variables of their behavior and meaning in context.

5.2 Position of DMs: Initiality in question

Apart from general contrastive results on frequency and diversity, another aspect where a categorical approach to DMs can prove enlightening is their positional behavior, in particular their supposed tendency towards utterance-initial position, as often claimed in general DM definitions.

5.2.1 Clause-initial DMs

As explained in the description of the methodology (Section 4.2.3), the position of DMs is annotated according to a tripartite system which distinguishes three reference units relevant to the behavior of DMs, namely the clause (a minimal propositional unit, including subclause), the dependency structure (a main clause and its constituents, roughly corresponding to an utterance) and the turn (the span between two changes of speaker). The main hypothesis in this regard is that initial position is expected to be the most frequent slot for DMs in English and French, although not to the same proportion across the three types of unit. Starting from the smallest unit, the hypothesis is largely confirmed at clause level, as shown in Table 5.5.

Table 5.5 Position in the clause (micro-position) by language

	English		French		Total	
	DMs	**ptw**	**DMs**	**ptw**	**DMs**	**ptw**
initial	3639	42.11	3417	45.39	7056	43.64
medial	292	3.38	223	2.96	515	3.18
final	245	2.84	730	9.70	975	6.03
independent	65	0.75	118	1.57	183	1.13
interrupted	8	0.09	6	0.08	14	0.09

It clearly appears that initial position is indeed the most typical use of DMs, with a very high frequency in both languages (over 40 occurrences ptw). Around 3 DMs ptw occur in final position in English, and in a similar frequency in the medial position of both English and French. Crosslinguistically, however, we see a sharp gap between English and French final DMs: the latter are more than three times

more frequent, a significant difference ($LL = 323.81$, $p < 0.001$) which can be partly explained by the exclusion of tag questions discussed in Section 5.1.2. The relative absence of final DMs in English might be the result of a pragmatic specialization of this syntactic slot to the occurrence of tag questions such as *isn't it*, which are presently excluded from the DM category although they express similar meanings (cf. Section 5.1.2). Further research is needed to delineate the distribution and uses of tag questions with respect to their DM rivals.

Another explanation for this crosslinguistic difference is related to the high frequency of the typically final French DMs *quoi* and *hein* identified in the previous section. In fact, these two DMs respectively take up 24% and 21% of all final DMs in *DisFrEn*, both languages combined (33% and 28% in French only).

The prevalence of initial position is somewhat nuanced by register variation: while initial DMs always take up the majority, the proportion varies from 74% (in conversation) to 94% (in political speech), with a general cline towards larger proportions in formal (professional, broadcast) registers, which is further proof of the restrictions on DM use in these speaking tasks.

5.2.2 Utterance-initial DMs

These findings can be refined by examining the more precise slots of macro-position, which not only distinguish the periphery (left or right) but also the (non-)integration of the DM with respect to the governing verb. Figure 5.3 represents the distribution of DMs across macro-syntactic slots in *DisFrEn*. The first observation is the striking similarity of English and French for this variable, with the notable exception of post-field DMs, whose higher frequency in French can be related to the previous finding at micro-syntactic level.

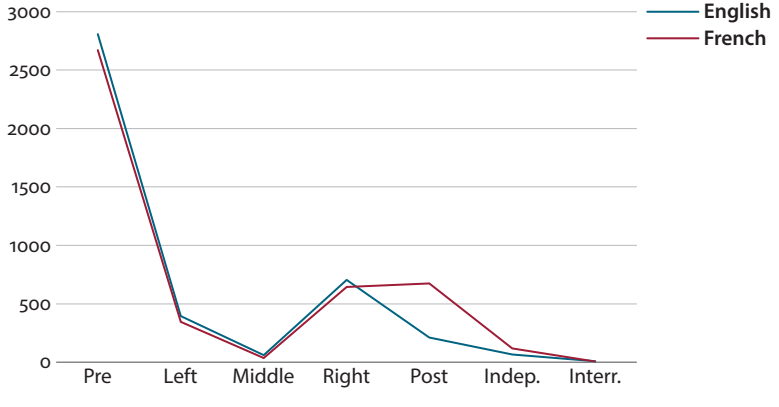


Figure 5.3 Macro-position (dependency level) of DMs

We see that, after the pre-field slot (i.e. initial, not integrated in the dependency structure), the second most frequent slot is "Right", that is, DMs occurring after the main verb yet integrated in its dependency (typically subordinating conjunctions such as *although* or *if*). Both pre-field and right-integrated DMs are initial in the sense that they introduce (different types of) units, namely whole utterances and subclauses, respectively, as illustrated by *so* in Examples (1) and (2).

(1) we will be examining the paradigm shift that's actually occurring (0.100) uh *so* (0.507) we've got a whole lot of uh clergy scientists poets (EN-phon-01)

(2) I like things also with a fantastic element to them *so* they stretch the imagination a bit which is what I've always liked (EN-phon-01)

In a micro-syntactic sense, both occurrences of "so" in these examples are initial: they introduce a (sub)clause, that is, a grammatical unit expressing a proposition and containing at least a predicate and its subject ("we've got a whole lot of uh clergy scientists poets" and "they stretch the imagination a bit"). *So* is one of the rare DMs which can occur both in integrated and non-integrated contexts (although not with the same function), while most DMs tend to specialize, usually as a consequence of their original grammatical class (subordinating conjunctions such as *although* are mostly integrated). In sum, this refined view of position converges with most DM definitions and confirms the central status of initial DMs, although initiality does not systematically imply that the DM occurs at the onset of a whole utterance.

The only notable effect of register variation concerns the higher relative frequency of left-integrated DMs in formal registers, where they occur as frequently as right-integrated DMs: 16% in political speech, around 10% in news broadcast, interview and classroom lesson, against around 6% in all other registers. In other words, both left- and right-integrated slots seem to be attracted to formality. This result evokes Pawley & Syder's (2000) notion of integration, which they associate with high levels of planning, as opposed to the less demanding mode of clause-chaining. Following their view, connecting segments by DMs at the left- and right-integrated macro-syntactic positions is cognitively costlier yet more "fluent" in that it reflects complexity and the efficiency of planning processes, which they in turn consider to be the basis of fluency defined as "the native speaker's ability to produce fluent stretches of spontaneous connected discourse" (Pawley & Syder 1983: 191).

5.2.3 Turn-initial DMs

Lastly, at turn level, initial position of DMs is no longer the most frequent slot, even in interactive registers, where turns are taken and given between speakers more rapidly than in other settings, where one speaker tends to hold the floor primarily (e.g.

face-to-face interview) or exclusively (monologues, e.g. political speech). Figure 5.4 shows the proportions of turn-initial DMs in these three degrees of interactivity. We see that DMs are used turn-initially in 15% of all occurrences in interactive settings such as conversations, against only 1% in non-interactive monologues, where they only correspond to contexts where the journalist resumes their speech after a documentary or a reporter's intervention during a news broadcast.

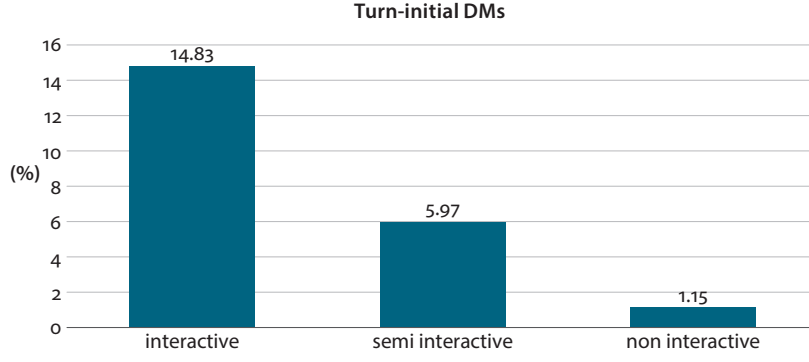


Figure 5.4 Proportions of turn-initial DMs by degree of interactivity

A similar observation can be made for turn-final and whole-turn DMs, which are also associated to interactive contexts (6% of final DMs and 1.42% of whole turns) and excluded from monologal registers. All in all, the initiality of DMs does not apply at turn level, even in registers where turns are a relevant structural unit. Nevertheless, in interactive settings, where DMs do occur at the beginning or end of turns (e.g. conversations), turn-initial DMs are always more frequent than turn-final DMs, which suggests a more prominent role of DMs in taking a turn (and holding it turn-medially) rather than giving it away.

The varying proportion of turn-initial DMs within (semi-)interactive situations could serve as an indicator of the mean length of turns. For instance, the difference in degree of interactivity between interviews (semi-interactive) and conversations (interactive) is reflected in the significantly higher proportion of turn-initial DMs in the latter (6% vs. 14%, respectively; $z = -8.89$, $p < 0.001$), which suggests longer turns in interviews, thus fewer occasions for turn-initial (or turn-final) DMs.[21]

The only crosslinguistic difference at turn level is qualitative and concerns the types of DMs each language uses primarily in turn-initial position: while the most frequent English expression is the speech-specific *well* ($N = 164$), French speakers

21. The z-ratio is used to test the significance of the difference between two independent proportions.

tend to start their turns with the more polyfunctional *et* 'and' ($N = 138$); French equivalents of *well* are much less frequent (*ben*, $N = 65$; *alors*, $N = 45$) while the English basic conjunction *and* also ranks lower ($N = 40$).

Overall, this higher unit of talk is mainly affected by register variation, in particular the effect of degree of interactivity, and only relates to the particular settings of the interaction, as opposed to the other two levels (i.e. micro- and macro-syntactic position) which allow further interpretation of their typicality in the category, their variation across registers and languages as well as their link to complexity and cognitive efficiency. This latter aspect is investigated in more detail in the next section.

5.2.4 Non-initial DMs

5.2.4.1 *Typical patterns*

So far, the prevalence of initial position has been established and nuanced by the variation in language, register and unit type. The present inclusive approach to the DM category allows us to complement its portrait by describing other patterns besides the initial position. Cross-tabulating macro-syntactic position with POS-tags offers a first filter into initial and non-initial patterns, as represented in Figure 5.5 (the vertical order of the values in the legend corresponds to that of the boxes on the barplot).

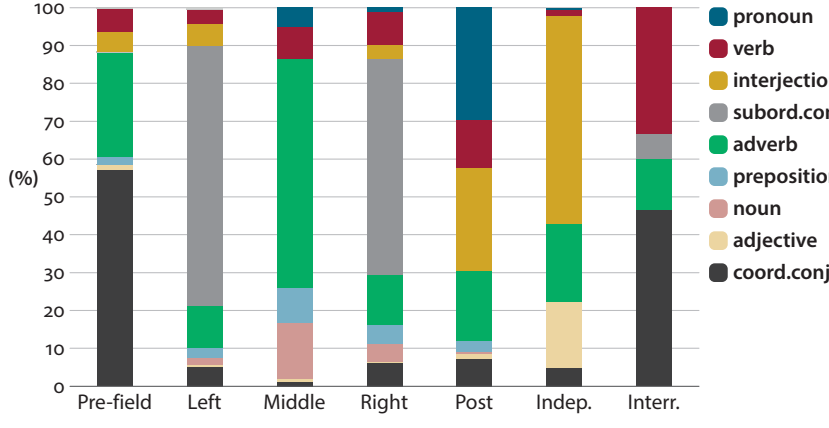


Figure 5.5 Proportions of POS-tags across macro-syntactic positions

Four favorites emerge from this graph: coordinating conjunctions in pre-field, subordinating conjunctions in both left- and right-integrated positions, adverbs in middle-field and interjections as independent units. These four patterns, which are the same in English and French, are illustrated in the following examples:

(3) they know that they're going to need these services as well (0.730) *and* also you can bring up pools (0.150) um using databases (EN-intf-08)

(4) *since* you're not having anything else you can have two of everything (EN-conv-05)

(5) the larger you get you can *therefore* make economies of scale (EN-clas-02)

(6) it's actually a proper increasing function (2.830) *okay* (1.730) so for example if you wanted to supposing you're looking… (EN-clas-04)

Example (3) corresponds to the generic use of DMs as inter-sentential connectives, where the related segments are at both sides of the DM. In (4), "since" is integrated in the syntactic structure of the main clause ("you can have two of everything") which is connected by a causal relation, both segments being located to its right. The medial position of "therefore" within the verb phrase "can make" in (5) is typical of more formal (even written) registers. Lastly, the pattern illustrated in (6) is the rarest one: stand-alone interjections tend to combine a hearer-oriented meaning, as in this example, with a punctuating or stalling function. This pattern is only instantiated by a handful of DM expressions in the corpus, namely *yeah*, French *bon* 'well', *hein* 'right' and *okay* in the two languages.

5.2.4.2 *Utterance-final DMs: Formal variation*

Non-initial positions can be expected to be more restricted in terms of DM variety, given that they are less frequent and less central in the category. We see in Figure 5.5 above that, contrary to expectation, less typical positions such as middle-field ("MID") or post-field ("POST") are not particularly more restricted in the types of POS-tags which can occur in these slots.

In particular, post-field DMs stand out from the other positions in that no POS-tag takes up the majority of occurrences, as opposed to all the other values which clearly favor one grammatical category over the others (interrupted units "INT" are also more balanced, yet their very low frequency precludes further discussion). No favorite POS-tag can be distinguished in the post-field slot, as opposed to the four patterns exemplified in the previous section. Utterance-final DMs appear to be more varied and balanced across five main possibilities, namely pronouns (30%), interjections (27%), adverbs (19%), verbal phrases (13%) and coordinating conjunctions (7%), leaving the remaining 4% to anecdotal cases of prepositional phrases, adjectives or noun phrases.

However, this greater formal variety of post-field DMs should be nuanced by taking into account the specific expressions each POS-tag covers. In fact, the occurrences of post-field pronouns are exclusively represented by French DMs, either *quoi* 'right' and variants (*voilà quoi*, *ou quoi*) or *et tout* 'and everything' and variants

(*tout ça, et tout ça*). Such pronominal DMs are specific to the post-field position (262 occurrences out of 293 in *DisFrEn*). Similarly, post-field noun-based DMs are only represented by three English expressions, viz. *and that kind of stuff*, *and things* and *or something*, while adjectival DMs only correspond to *right* and French *bon* 'right' in this final slot.

Therefore, the information of POS-tags can be refined by a second measure of formal diversity inspired by the so-called "standardized type-token ratio", which I adapted by computing the ratio of DM types (i.e. expressions) by macro-syntactic slot on a random sample of 100 DMs in each position and language. This ratio thus neutralizes differences in the overall frequency of DMs by position. Focusing on the opposition between pre- and post-field, this ratio shows a large contrastive effect on formal diversity: while the English data corroborates the higher formal diversity of post-field DMs shown in Figure 5.6, with 21 different DM types vs. only 10 in the pre-field slot, the French data shows a slightly reversed tendency, with 21 DM types in pre-field vs. 18 in post-field.

In sum, the utterance-final (post-field) position is not particularly more restricted in terms of formal diversity of syntactic classes than the typical pre-field slot, although a finer analysis of DM types nuances the difference between pre- and post-field, especially in French where the former covers more different DM types than the latter.

5.2.4.3 *Clause-medial DMs: Potential disfluency?*

Clause-medial DMs mostly correspond to adverbs, which are known for their syntactic mobility (41% in English and 35% in French, against 24% and 22%, respectively, in initial position). They are followed by verbal phrases (e.g. *I mean*), which cover about 20% of medial DMs in English and French, and prepositional phrases (e.g. *for example*) which are particularly frequent in the French data (25% vs. 5% in English).

With a view to evaluating the (dis)fluency of DMs, medial position could be associated with intrusive or interrupting uses of DMs which disturb the syntactic structure of the utterance. However, qualitative observation of the data excludes such generalization, or at least nuances it depending on the POS-tag of the DM: adverbs, verbal and prepositional phrases are not equally related to intrusiveness, as illustrated by the typical patterns in Examples (7)–(9).

(7) is there quite a high demand *then* for um care (0.260) nowadays (EN-intf-04)

(8) she was in the film within *you know* d- a day or two (EN-intr-04)

(9) we're building another care home *for example* in Yeovil at the moment (EN-intf-04)

Formally, we see similarities between the DMs in Examples (7) and (9), which both occur before a prepositional phrase ("for care", "in Yeovil"), while "you know" in (8) is phrase-internal ("within a day or two"). The compliance with linguistic boundaries, albeit local, could be seen as a first sign of the higher fluency of adverbial and prepositional DMs in medial position. Additional (functional) variables are needed to support this interpretation.

5.2.4.4 *The case of hedges*

Besides adverbs, verbal and prepositional phrases, the most striking distinctive feature of the clause-medial position is the occurrence of noun-based DMs in English. They take up 29% of all English medial DMs, never occur in French and only correspond to the *sort of – kind of* pair. These DMs, which are sometimes classified as hedges or mitigators (e.g. Brown & Levinson 1987; Miskovic-Lukovic 2009), do not seem to have a French equivalent, yet they meet the criteria of DM definition (procedural meaning, grammatically optional, metadiscursive, formally fixed). In *DisFrEn*, *sort of* and *kind of* mostly occur in clause-medial position, with rare initial occurrences as in (10); no occurrence of final position was found in the corpus, although it seems that this use might be developing, as attested by several examples found online, such as (11) coming from the title of a thread on a fansite.

(10) I'd dearly love to uh you know to be spending time writing poetry and fiction and *kind of* this last year's been (0.960) been uhm kind of commissioned work (EN-phon-01)

(11) I was happy… *sort of*![22]

The rarity of initial contexts exemplified in (10) and the absence of final contexts in the corpus point to the attraction of this DM pair to clause-medial position. The English specificity of noun-based DMs, as shown by the absence of this POS-tag in the French component of *DisFrEn*, should however be nuanced by the French DM *genre* 'like', which is strikingly similar to *kind of* both formally (same grammatical class and positional behavior) and semantically (originating from a word meaning "type, sort"). Only one occurrence of the DM *genre* was found in the corpus, reported as Example (12), which can possibly be explained by the fairly recent development of this DM in rather informal conversations between younger speakers (Fleischman & Yaguello 2004), data which were not available at the time of corpus collection.

22. http://m.downatthemac.proboards.com/thread/12655/happy-sort

(12) <VAL_16> ah une pièce de théâtre?
<VAL_15> oui ou bien un spectacle tu sais *genre* un mime ou je sais pas un petit spectacle
<*VAL_16*> *ah a play?*
<*VAL_15*> *yes or a show you know* genre *'like' a mime show or I don't know a little show* (FR-conv-03)

The effect of mitigation brought about by *sort of, kind of* or French *genre* 'like' seems to suggest a pragmatic specialization of clause-medial DMs to epistemic or sense-altering functions, whereby DMs are used to discard a literal interpretation of the host-utterance. However, these three DMs only represent 17% of all clause-medial DMs in the corpus (29% of all English medial DMs) and further functional analysis should confirm whether similar pragmatic uses apply to other clause-medial forms as well (see Section 5.4).

5.2.5 Interim summary on position

To sum up, the general distribution of DMs across languages and registers was refined by the analysis of their positional behavior in three types of units. The main results of Section 5.2 include:

- the prevalence of initial position, especially in formal registers, except at turn level even in interactive situations;
- the higher frequency of final DMs in French than in English;
- the higher frequency of syntactically integrated DMs either at left or right periphery of the main verb (i.e. subordination) in formal registers;
- four patterns of POS-tags by position, namely coordinating conjunctions in pre-field, subordinating conjunctions in both left- and right-integrated position, adverbs in middle-field and interjections as independent units;
- the higher formal variation of DMs in post-field than in initial position (especially in English), against our hypothesis;
- language-specific DMs, namely pronoun-based DMs in final position in French and noun-based DMs in medial position in English (and the quasi-absence of the French equivalent *genre* 'like');
- the potential disfluent character of medial position, related to intrusiveness and sense-altering DMs (such as hedges).

So far, the analysis was mainly descriptive and quantitative, based on purely formal variables. Independent and univariate investigation of syntactic and positional features of DMs is necessary because of the high variation of their behaviors, which requires careful step-by-step description – an endeavor which has never been

undertaken before on such a large scale on spoken English and French. I will now turn to the main contribution of this study, which is the functional analysis of DMs and the integration of syntax and pragmatics across languages and registers.

5.3 Domains and functions: Frequency and diversity

Functional variables are divided into two levels which vary in their degree of granularity, from four domains to 30 functions. Each level will be analyzed separately, using more elaborate statistical tools and integrating previously discussed variables in order to obtain comprehensive, multivariate models of DM behavior in various registers of English and French.

5.3.1 Single domains

As a reminder, the taxonomy of DM domains and their respective function values is reported below as Table 5.6. As mentioned in the methodology, a DM can be assigned up to two simultaneous domains and functions. Double tags only concern 350 DMs occurring mostly in phone calls and conversations, and will be treated separately (Section 5.3.3) given that they involve a slightly different annotation procedure and cannot be analyzed with the same method. The large majority of DMs in *DisFrEn* were only assigned one domain label and one function label. The analyses in this section deal with the distribution of these 8,393 occurrences in terms of language and register variation, as well as additional observations of association patterns.

The more coarse-grained functional variable is that of the domain of use, a term taken from Sweetser (1990) which refers to the level of discourse targeted by the DM. In this work, I distinguish four domains, namely ideational (content,

Table 5.6 Taxonomy of DM domains and functions

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		

objective relations), rhetorical (speaker's attitude, subjective relations), sequential (turn exchange and topic structure) and interpersonal (speaker-hearer relationship). In the following, mentions of "ideational DMs", for instance, will refer to ideational *uses* of DMs or DMs with an ideational function, as they were manually disambiguated in context. These shorthand terms are not meant to suggest that the same DM is always used in one domain or another: domains and functions are always assigned to individual occurrences in the data, in order to account for the great polyfunctionality of some DMs.

Based on Denke's (2009) corpus findings, DMs are expected to attend primarily to discourse structure, in other words, the sequential domain is hypothesized to take up the majority of DM uses in *DisFrEn*. Additional hypotheses of register variation further suggest that the sequential domain is favored in monologic situations (based on Denke 2009) and that the ideational domain is prevalent in factual discourse types (news broadcast, political speech, classroom lesson) given its very definition. The effect of register on domain distribution can also be reflected in a higher internal variation and diversity of DM domains in intermediary registers (e.g. interviews) as opposed to discourse types at either extreme of a formality scale. In particular, informal registers (e.g. conversations) are hypothesized to be strongly associated to interpersonal DMs, whereas formal registers (e.g. news broadcasts) are expected to show a high proportion of ideational DMs. Lastly, no specific hypothesis was formulated regarding crosslinguistic differences between English and French as far as DM domains are concerned.

5.3.1.1 *Domains across languages*

The data is reported in Table 5.7. It appears that sequential and rhetorical DMs occur in very similar (not significantly different) rates, with about 15 DMs ptw in English and 18 DMs ptw in French. Another similarity is found with ideational DMs in English and in French (about 13 DMs ptw, $LL = 3.28$, $p > 0.05$). Interpersonal DMs appear to be much less frequent than the other three domains, especially in English where they barely amount to 4 DMs ptw (8% of all English single domains).

Table 5.7 Distribution of single domains by language

	English		French		Total	
	DMs	**ptw**	**DMs**	**ptw**	**DMs**	**ptw**
Sequential	1269	14.69	1411	18.74	2680	16.57
Rhetorical	1319	15.26	1331	17.68	2650	16.39
Ideational	1144	13.24	920	12.22	2064	12.76
Interpersonal	322	3.73	677	8.99	999	6.18
Total	4054	46.91	4339	57.63	8393	51.9

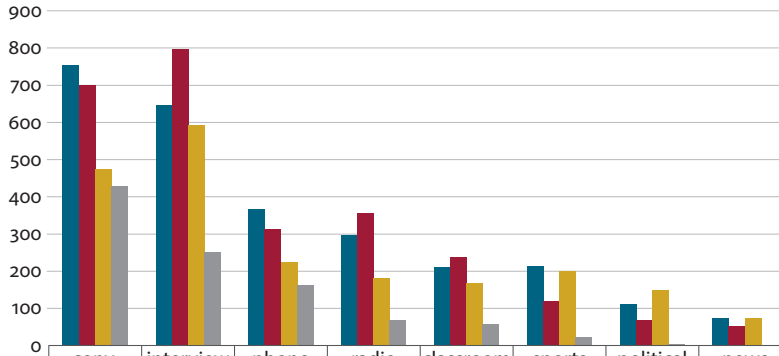
Overall, the data confirms the high frequency of sequential (text-structuring) DMs, although the difference with the rhetorical domain is very small (most frequent domain in English), while interpersonal DMs are the least frequent in the category, especially in English. This last observation can be interpreted in two different yet related ways. Methodologically, the interpersonal domain includes fewer functions than the other three domains (cf. Table 5.6) and thus offers fewer possibilities for DMs to function at this level of discourse. Theoretically, this is in turn related to the peripheral status of interpersonal functions in the DM category, as opposed to the other domains which are more representative of typical DMs and not (all) restricted to spoken language. However, neglecting the interpersonal domain altogether would overlook 12% of the DMs as broadly defined in *DisFrEn*.

The minimal role of language variation in domain distribution can be explained by the fact that DMs are presently defined through a functional *tertium comparationis* which strives to overcome the distinctive features of English and French,[23] while register can be expected to show stronger effects. This is confirmed at a very general level by a random forest analysis, computed with the cforest function from the {party} package (Hothorn et al. 2006) in R-Studio, an open-source statistical software. Random forests try to replicate the observed data in a very large number of decision "trees" and make it possible to evaluate a measure of distance or error between observed and predicted values, as well as the most relevant factors in the decisions. With both language and register as factors, the random forest analysis relies more strongly on the effect of the latter to train the algorithm and predict the domain value for each DM, which points to a larger discrepancy between registers than between languages.

5.3.1.2 *Domains across registers*

The distribution of DM domains across registers is provided in Figure 5.6. This graph clarifies the complementarity between the sequential and rhetorical domains, which are each preferred in different registers. The sequential domain is most frequent (although by very little) in spontaneous settings such as conversations, phone calls or sports commentaries, whereas rhetorical DMs are most frequent in both face-to-face and radio interviews and, to a lesser extent, in classroom lessons. This latter group of registers might be characterized as an argumentative discourse type where speakers tend to convince and develop their point of view.

23. But see above for the exclusion of English tag questions.



	conv.	interview	phone	radio	classroom	sports	political	news
Sequential	755	648	367	297	211	215	112	75
Rhetorical	701	799	312	356	239	121	70	52
Ideational	475	594	224	182	168	200	148	73
Interpersonal	430	251	163	69	58	23	5	0

Figure 5.6 Distribution of DM domains across registers

The preference of both rhetorical and sequential DMs over ideational DMs is particularly surprising in classroom lessons, which can be expected to behave more like expository and objective texts. The preference for ideational DMs is however confirmed in the other two "factual" settings, namely political speech and news broadcast, where they show an equal or slightly superior frequency than sequential DMs.

The three patterns discussed so far (sequential in spontaneous discourse; rhetorical in argumentative discourse; ideational in factual discourse) are illustrated in Examples (13)–(15).

(13) I think she actually likes it but (0.727) she has a sense of proportion hold on here's a napkin oops (0.280) *by the way* did I mention my dustbin's been blown over in my back garden again (EN-conv-04)

(14) and this also gives a rather cool perspective on Bristol *because* many of the people living and working in Bristol (0.350) are creative designers (EN-intf-05)

(15) we have done best (0.960) when we've seen the community not as a static entity to be resisted and contained (0.840) but as an active process which we can shape often decisively (0.790) *provided* we allow ourselves to be fully engaged in it (0.680) with confidence (EN-poli-01)

The topic-shift expressed in (13) by "by the way" is representative of the frequent changes of subject during impromptu conversation where topic is not pre-established

nor constrained: here an element of context (a napkin probably falling on the floor) triggers a shift from discussing a female referent ("she") to a dustbin. In (14), the speaker is trying to advertise the dynamism of the city of Bristol to the interviewer and justifies his evaluation ("cool") by an argument about art and creation introduced by "because". The politician in (15) is laying out facts and presenting a goal ("we have done best") as a logical and hypothetical result of the condition introduced by "provided". Political speeches as well as news broadcasts are also relatively profuse with sequential DMs (cf. their similar frequency with ideational DMs in Figure 5.6), mostly in additive, topic-shift and enumerating functions. However, on the whole, the hypothesis from Denke (2009) on the higher frequency of the sequential domain in monologues is not confirmed, as shown in Table 5.8.

Table 5.8 Relative frequency of domains (ptw) by number of speakers

	Sequential	Rhetorical	Ideational	Interpersonal	Total
monologue	10.18	8.07	10.27	1.22	29.74
dialogue	19.88	20.64	14.07	8.71	63.30
multilogue	8.69	9.31	8.69	2.48	29.17

We see that sequential DMs are as frequent as ideational DMs in monologues and do not occur relatively more in monologues than in dialogues either (on the contrary, they are half as frequent). This is probably due to the inclusion of functions related to turn exchange in the sequential domain, which are by nature related to dialogues, as well as the very basic *addition* function which is not restricted to any particular register. Moreover, this table shows that the distribution of domains in monologues is relatively equal among the top three domains (from 8 to 10 DMs ptw in the sequential, rhetorical and ideational domains). A similar balance is found in multilogues (around 9 DMs ptw) and dialogues, but only between the sequential and rhetorical domains for the latter (around 20 DMs ptw).

The interpersonal domain, which has scarcely been discussed so far, appears with a low frequency across all registers, especially those related to formal features (monologues, prepared, non-interactive). By definition, interpersonal DMs are connected to dialogue, which is reflected in Table 5.8. In fact, 84% of all interpersonal DMs occur either in conversations, face-to-face interviews or phone calls. Their frequency in other registers is low, even null in formal settings. The interpersonal domain stands apart from the others by this highly uneven balance between registers, more so than any other domain, especially in comparison with the sequential domain, as graphically represented in Figures 5.7 and 5.8.

In these graphs, we see that sequential DMs consistently take up about 30% of all DMs in each register, whereas the situation is much more irregular for the

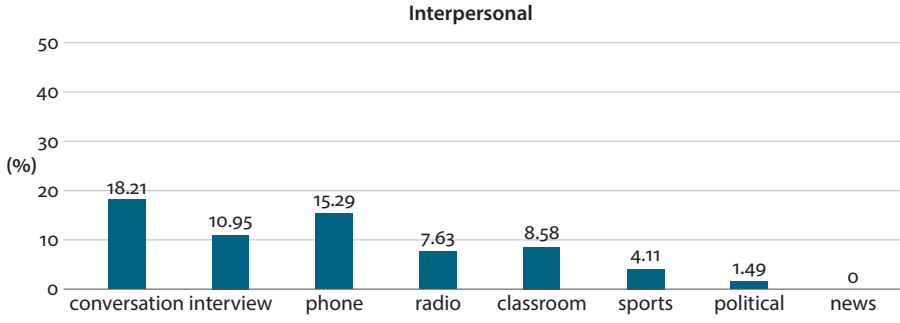


Figure 5.7 Proportions of interpersonal DMs in each register

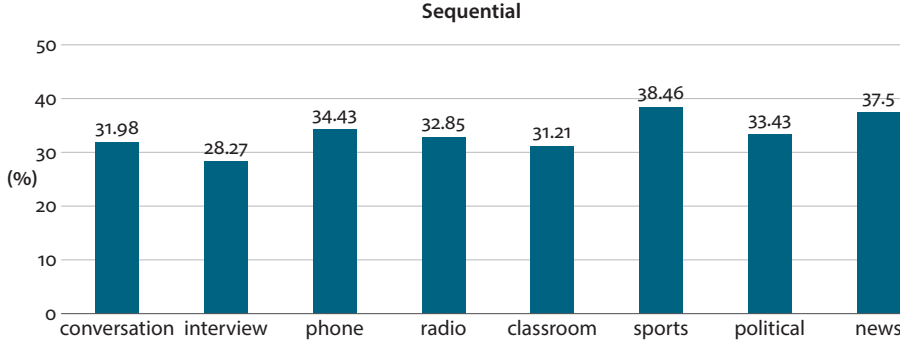


Figure 5.8 Proportions of sequential DMs in each register

interpersonal domain. Nonetheless, in conversation, speakers attend to interpersonal functions of discourse almost as much as they connect their speech with ideational relations (18% vs. 20%, respectively). In other words, inter-subjectivity appears on a par with objectivity in this very natural and casual situation of language: conversational partners are not so much concerned with facts (in comparison with other registers) as they are attentive to the hearer's needs and the communicative success of the exchange.

A final remark on the distribution of domains across registers addresses the hypothesis of the higher functional variety of intermediary settings such as interviews, as opposed to more extreme (i.e. very formal and very informal) contexts which are expected to be more restricted to ideational and interpersonal DMs, respectively. At domain level, we saw that political speeches and news broadcasts show (almost) no interpersonal DMs, while the other contexts include occurrences of all four domains. Apart from this restriction, no monopoly can be observed in one register or

another. This is not to say that there is no internal variation: for instance, ideational DMs do take up a larger proportion in political speeches and news, while, as we just saw, interpersonal DMs are more frequent in conversations. However, no domain takes up the majority of all DMs in any register and intermediary settings such as interviews are not particularly more diversified or balanced than more extreme contexts. In this respect, semi-prepared settings are more similar to spontaneous than to prepared interactions, as can be seen in the three pie charts in Figure 5.9 (the variation by degree of interactivity is highly similar).

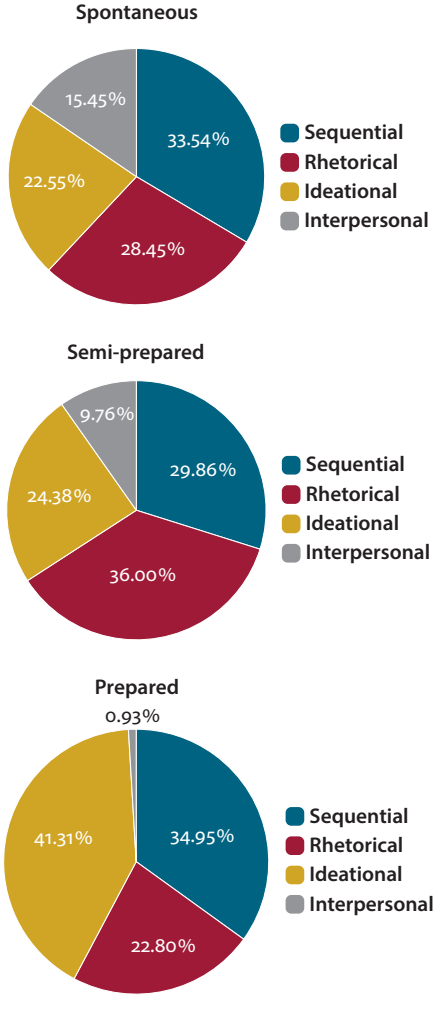


Figure 5.9 Balance of domains in the three degrees of preparation

While semi-prepared settings do appear intermediate between spontaneous and prepared settings, especially when looking at the decreasing proportion of interpersonal DMs, we see that spontaneous discourse is actually more balanced between the four domains, thus disproving the hypothesis. At domain level, such an analysis is limited: a more fine-grained account of the pragmatic diversity of registers will be provided in Section 5.3.2.3 at function level by carrying out an analysis of the DM function ratio (number of different function types by total number of DMs) by register.

5.3.1.3 *Domain-specific DMs*

Patterns of domain variation can be further refined by looking for any domain-specific POS-tags or particular expressions, which only or mostly correspond to one of the four domains. Such formal associations, if observed in the data, can serve as robust cues for the automatic disambiguation of DM domains (see also Bolly et al. 2015, 2017 for a similar ambition applied to DM identification), or at least as reliable criteria for the annotator. All nine possible POS-tags of the DM category found in *DisFrEn* are ranked by overall frequency and cross-tabulated with the four domains and two languages in Table 5.9.

Table 5.9 Cross-tabulation of domains and part-of-speech tags in English and French

	Sequential		**Rhetorical**		**Ideational**		**Interpersonal**	
	EN	**FR**	**EN**	**FR**	**EN**	**FR**	**EN**	**FR**
coord. conj	835	810	442	391	426	328	1	7
adverb	352	210	432	493	206	162	44	18
subord. conj	0	2	161	194	510	401	0	0
interjection	21	213	0	30	0	9	47	366
verbal phr.	47	12	146	65	0	0	201	109
pronoun	0	80	0	28	0	1	0	168
prep. phr.	5	14	55	118	2	19	0	0
adjective	7	70	0	11	0	0	22	9
noun phr.	2	0	83	1	0	0	7	0

A number of interesting observations can be drawn from this table. First, we see that adverbs (e.g. *so*, *well, now, actually*, French *donc* 'so', *alors* 'well', *enfin* 'I mean') appear to be the most polyfunctional syntactic class of the category, with a substantial frequency in each domain, as opposed to all the other values which are restricted to two or three domains. In light of this finding, adverbs can be considered as the most representative syntactic class of DMs, as opposed to the often mentioned coordinating conjunctions (e.g. Lee 2002) and to what overall frequency would suggest. Coordinating conjunctions are only very rarely used to express interpersonal

meanings such as *monitoring* or *disagreeing*: of the 8 occurrences of interpersonal conjunctions, 7 are in French, including six *mais* 'but' as in Example (16).

(16) il faut tout négocier avec eux tu vois euh pff (1.210) c'est fatigant tu vois on prend leurs bics pour le TU ben euh (0.900) quoi *mais* on n'a pas dit qu'on voulait bien gnagna tu vois
we have to negotiate everything with them you know uh pff it's annoying you know we take their pens for the meeting well uh what mais *'but' we did not say we agreed blabla you know* (FR-conv-05)

In this example, the speaker is reporting someone else's words ("quoi mais on n'a pas dit qu'on voulait bien") in a conflicting situation where the reported speaker is not willing to lend his pens: he (supposedly) introduces his objection with an interjection of surprise ("quoi") followed by a disagreeing "mais". Such cases are quite rare and their interpersonal interpretation might be questioned since traces of the contrastive meaning of *mais* are still present.[24]

Another observation concerns subordinating conjunctions, which are the third most frequent POS-tag overall while only occurring in the ideational and rhetorical domains (with two exceptions). The lack of subordinating conjunctions in the sequential and interpersonal domains is compensated by their highly frequent ideational use (44% of all ideational DMs), where they are more frequent than coordinating conjunctions. This pattern includes DMs such as *because, if, when* or *while* and their French equivalents. In other words, although this POS-tag ranks very high in frequency on the whole DM category, it seems particularly restricted in terms of domain, which lessens the contribution of frequency information alone without further qualitative (here, functional) filters.

Three POS-tags stand out as particularly associated with the interpersonal domain, namely interjections, verb phrases and pronouns. They are the only categories which are most frequent as interpersonal DMs and, once combined, they take up 89% of all interpersonal DMs. Although interjections tend to frequently occur as sequential DMs (e.g. French *ben* 'well') and verb phrases as rhetorical DMs (e.g. *I mean*) as well, the strong association between the interpersonal domain and the three abovementioned POS-tags could be safely considered as a reliable pattern and cue for sense disambiguation (a more complex multivariate model integrating positional variables will confirm this finding in Section 5.4). Examples (17)–(19) illustrate these interpersonal patterns.

24. This is one of the reasons why Crible & Degand (in press) propose to re-structure the functional taxonomy and annotate these cases as "interpersonal contrast".

(17) moi il me gonfle comme tous les écrivains mais Céline aussi *hein* tout n'est pas (0.239) du génie (0.102) absolu personne
he bores me like every writer but Céline as well hein *'right' not everything is absolute genius no one* (FR-intr-03)

(18) yeah I'm just phoning up and doing that thing I was talking to you about *you know* (0.300) recording (EN-phon-05)

(19) si tu veux il y avait des personnages mais qui étaient pas animés *quoi* hein c'était tout euh euh figés
if you will there were characters but who were not animated quoi *'you know' right it was all uh uh fixed* (FR-conv-05)

A last pattern of domain-specific POS-tags is that of prepositional phrases (e.g. *in fact, for example*) and noun phrases (e.g. *sort of*), which are almost exclusively used as rhetorical DMs in 81% and 90% of all their occurrences, respectively (both languages combined). Again, such patterns could prove useful in predictive and statistical perspectives such as automatic classification (see Section 5.4).

In sum, analyses at domain level reveal clear tendencies of variation across languages, registers and specific DM types. The polyfunctionality of the DM category is confirmed by the functional diversity of each register, especially intermediary and informal ones, even at this rather coarse-grained level of analysis (as opposed to more specific function values).

5.3.2 Single functions

5.3.2.1 *Functions across languages*

The more fine-grained functional variable deals with the thirty function values which are categorized in the four domains discussed above (as a reminder, the function *cause* is always ideational, while *motivation* is always rhetorical, for instance). At this level of analysis, no particular hypotheses were drawn from the literature beyond the investigation of any relevant contrast between languages and registers. In addition, I will replicate the mapping of variables as carried out in the previous sections, to test whether some functions are associated to specific registers or DM expressions. Given the large number of values, the full table of all functions with their frequency by language and DMs expressing them will not be discussed here but is provided in Appendix 3. Only the ten most frequent functions are reported in Table 5.10.

Not surprisingly, the most frequent function in both languages is *addition*, typically expressed by the basic conjunctions *and* / *et*: every thousand words, eight DMs are used to merely connect two utterances together with no additional meaning

Table 5.10 Ten most frequent functions and their relative frequency by language

English		French		Total	
Function	**ptw**	**Function**	**ptw**	**Function**	**ptw**
addition	7.86	addition	7.42	addition	7.66
specification	3.99	monitoring	6.40	monitoring	4.44
consequence	3.21	opposition	3.84	specification	3.87
temporal	3.14	specification	3.73	opposition	3.38
conclusion	3.10	conclusion	3.21	conclusion	3.15
opposition	2.97	temporal	3.04	temporal	3.09
monitoring	2.73	consequence	2.67	consequence	2.96
opening	2.48	punctuation	2.62	opening	2.5
concession	2.44	opening	2.52	concession	2.28
condition	1.61	topic-shift	2.24	punctuation	1.76

other than inter-sentential coordination. Apart from *addition*, only *conclusion* occupies the same rank (5th most frequent) in English and in French. Most other functions in this top ten are shared between the two languages but in different ranks. The main crosslinguistic difference concerns the *monitoring* function, which ranks 2nd in French against only 7th in English with a highly significant gap in frequency ($LL = 123.32$, $p < 0.001$). *Monitoring* is mostly expressed by *you know* in English (180/236) and *hein* 'right', *quoi* 'you know' and *tu vois* 'you see' in French (256, 112 and 53/482, respectively).

Language-specific functions which do not enter the top 10 in the other language are *concession* and *condition* in English (respectively ranked 11th and 16th in French), *punctuation* and *topic-shift* in French (ranked 14th and 16th in English). This comparison is reflected in the higher proportion of ideational functions in English than in French (28% vs. 21%; $z = 7.459$, $p < 0.001$), while the crosslinguistic difference in sequential functions is not significant (31% vs. 33%; $z = -1.195$, $p > 0.05$).

5.3.2.2 *Functions across registers*

The picture becomes more complex with register information. When comparing the top five functions in each subcorpus, a number of interesting observations emerge, which are summarized in the following (see Appendix 4 for the distribution of these functions by register).

Addition is always the most frequent function except in English and French phone calls (*opening*), English interviews (*specification*) and French conversations (*monitoring*). *Monitoring* is highly affected by register (in the top five of most informal and intermediary registers, least frequent in political speeches and news

broadcasts). The *opening boundary* function (i.e. turn-taking) only makes the top five in conversations (English only) and phone calls, which reflects the interactivity and rapid exchange of turns in these settings. Ideational functions such as *temporal*, *consequence* or *concession* appear in the top five of intermediary and formal registers but not in casual conversations. The *approximation* function is completely absent from broadcast monologues (news broadcasts, political speeches and sports commentaries), which might relate to the professionalism of these settings and the need to appear confident. News broadcast is the only register where *topic-shift* ranks among the most frequent functions (5th and 4th in English and French), which can be interpreted as a result of the artificial nature of this type of language where topics are usually explicitly changed.

5.3.2.3 *Functional diversity*

Another source of contrast between registers might be their varying functional diversity. An analysis of DM function ratio can reveal whether high frequency of DMs is necessarily associated with high number of function types in a particular register: the higher the ratio, the greater the diversity. Such a score would be strongly affected by the overall frequency of DMs in each register, given that more DMs give more occasions to express a wider panel of functions. Therefore, a more comparable method to identify functional diversity (or, on the contrary, monopoly) is to count how many different functions it takes to reach half of all DMs in each register, in other words, to use the cumulative frequency of function types. This data is shown in Figure 5.10.

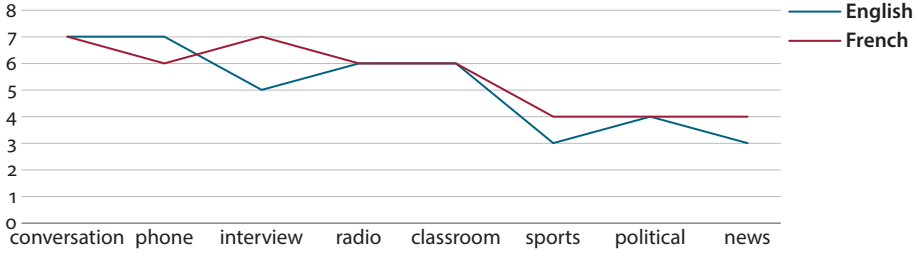


Figure 5.10 Number of function types making up 50% of DMs by register and language

This graph should be read as follows: in conversations, more than 50% of all DMs are distributed across seven function types in English and French. The registers are ranked by decreasing frequency of DMs from left to right, which also roughly corresponds to increasing formality. We see that it takes fewer and fewer different function types to amount to half of all DMs, from seven to three, with a large drop occurring in sports commentaries. Although the differences are small, they suggest

a decrease in functional variety in more formal, broadcast and monologic registers, as previously shown by Castellà (2004).

A second method of counting is inspired by the so-called "standardized type-token ratio": the ratio neutralizes differences in corpus size (here, differences in DM frequency by register). I adapted it by computing the ratio of function types on random samples of 50 DMs in each register and language (e.g. 15 function types for 50 DMs in French conversations). The results can be seen in Table 5.11. They tend to confirm our previous observations: sports commentaries, political speeches and news broadcasts stand apart with a lower DM function ratio than all other registers, especially in English as far as sports and politics are concerned, although the differences are quite small.

Table 5.11 Standardized DM function ratio by language and register

	English	**French**
conversation	0.34	0.3
phone	0.4	0.34
interview	0.32	0.42
radio	0.4	0.42
classroom	0.34	0.34
sports	0.26	0.34
political	0.24	0.3
news	0.28	0.26

Perhaps more surprisingly, conversations (where the relative frequency of DMs is the highest, especially in French) do not show the highest ratio on this random sample but rather appear intermediate (less so in English) between registers such as radio interviews on the one hand and news broadcasts on the other. This tentative result might be seen as partial confirmation of the higher restriction of registers at either extreme of the formality scale: although disproven at domain level, this hypothesis is at least not incompatible with the ratios found in Table 5.11, which place conversations at an intermediary level of functional diversity against the more varied range of radio interviews, for instance.

Many more analyses could be carried out on all or some of the thirty functions annotated in *DisFrEn*, answering different research questions investigating particular DMs or functions. In the present descriptive perspective, the results were deliberately limited to significant trends of variation between the two languages and eight registers and observations of functional diversity. Functional considerations at such a fine level of granularity will be taken up in Chapter 7 in relation to hypotheses of fluency and in Chapter 8 focusing on the functions of DMs within repairs.

5.3.3 Double domains and functions

Once combined, the four domains of the DM category amount to 14 possible values, including single domains (e.g. ideational), repeated domains (ideational-ideational) or combined domains (ideational-sequential). Such a high number of categories makes any statistical modeling difficult to handle quantitatively, which is why double domains are treated separately in this section, where they will be discussed with information from the function level as well. Given the large number and low frequency of double tags (either domains or functions), quantitative analyses are very limited. Only a few observations will therefore be discussed in the following, with no reference to cognitive hypotheses of fluency. Nevertheless, double domains and functions might provide further insights into the multifunctionality of the DM category. Their distribution is reported in Table 5.12.

Table 5.12 Distribution of double domains per language

	English	French	Total
RHE-SEQ	97	82	179
INT-SEQ	40	18	58
INT-RHE	13	26	39
RHE-RHE	16	11	27
IDE-SEQ	15	6	21
IDE-IDE	5	6	11
SEQ-SEQ	4	3	7
IDE-RHE	3	3	6
IDE-INT	1	0	1
INT-INT	1	0	1
Total	195	155	350

We see that, out of the eight possible combinations, half of all occurrences are rhetorical-sequential combinations ("RHE-SEQ"). RHE-SEQ cases cover 36 different combinations at function level. The most frequent of these combinations is illustrated in Example (20), where *so* expresses both a conclusion and a topic-resuming function.

(20) because of the history here there's a lot of people that know the machines know the original DUKWs that they're based on (0.500) […] because they were originally (0.260) uh the Americans actually (0.130) constructed them here in Plymouth yeah they constr- constructed a huge amount of them here (0.300) actually at Qu- Queen Anne's battery (0.340) which is now a marina which is also where our slipway is so the slipway we're using was used (0.310) uh by the original machines […] (0.380) *so* there's a lot of history (0.330) with Plymouth and the original machines (EN-intf-02)

The utterance introduced by "so" in (20) is related to its previous context both in a rhetorical ("I can say that there is a lot of history because…") and a sequential way ("to come back to my original statement, there is a lot of history in Plymouth"). Apart from this pattern, which accounts for 27 cases, the majority of double functions are *hapax legomena* or very rare cases, even within the relatively frequent RHE-SEQ domain (e.g. two occurrences of enumeration-opposition). The relatively high frequency of this combination, although covering many distinct functions, can be interpreted in multiple ways, either as a result of the very high and similar frequency of these two domains in general, as a sign of the conceptual proximity of sequential and rhetorical functions or, on the contrary, of their difference and complementarity: speakers tend to simultaneously attend to both of these domains (i.e. express their subjectivity and structure discourse) to maximize the connectedness of their speech.

As opposed to the analyses of single domains, where clear patterns of variation and association were identified, the low frequency of double domains does not allow for such interpretations, even with a more qualitative approach to the data. Looking at DM expressions, 68 different types were assigned a double tag, against the total of 218 different DMs in the corpus, a ratio which is particularly high given the low overall frequency of double tags. The most frequent double-tagged DMs roughly follow the general ranking of frequency (*but, so, well*, French *mais, donc*) – with the notable absence of *and* / *et* in this ranking – and this level of multifunctionality does not seem to be restricted to particular (speech-specific or other) DMs. In addition, no major restriction of register was found in the data, as can be seen in Table 5.13.

Table 5.13 Distribution of double tags and overall proportion by register

	DMs	**%**
conversation	113	4.57
interview	76	3.21
phone	73	6.41
classroom	29	4.11
sports	17	2.95
political	16	4.56
radio	16	1.74
news	10	4.76
Total	350	4

We see that, in absolute frequency, double tags occur in each register in roughly the same ranking order as the overall frequency of DMs. The proportion of double tags against all DMs is low in all registers, ranging from 1.74% (in radio interviews) to 6.41% (in phone calls).

In total, 105 different types of combinations at function level were annotated in *DisFrEn*. This very high ratio (105/350) does not guarantee strong replicability during the annotation, since the analyst cannot rely on the observation of recurrent patterns of use. The high variability and low frequency of double tags refrains me from pursuing their analysis any further. It may well be the case that the phenomenon of double-tagging has some formal or cognitive basis. For instance, it could be related to ambiguous DMs (expressions with a weak core meaning which cannot be disambiguated with one single tag only, e.g. French *quoi* 'right') or to co-occurrence with DMs or other fluencemes, in which case the multifunctionality of double-tagged DMs encompasses the pragmatic meanings of the elements they cluster with. However, such interpretations would require more data and a more reliable annotation procedure.

One way to ease the treatment of double tags is simply to remove the option from the annotation scheme, by suggesting systematic biases towards one of the two domains under consideration. This would however overlook the multifunctionality of some DMs and potentially skew the data. However, the present state of this research does not allow further analysis on a par with single tags, which is why the remainder of this study will focus on the 8,393 single-tagged DMs whenever functional variables are concerned.

5.4 Integrating syntax and pragmatics

The independence of the positional and functional annotations allows us to draw a number of conclusions regarding the mapping and integration of these variables. Previous research, as well as the very definition of the functional categories, suggest a number of hypotheses in this regard:

- the higher discourse scope of sequential DMs is expected to be reflected in a strong preference for initiality;
- final position has been identified as a typical *locus* for hearer-orientation and interpersonal DMs (Traugott 2007; Degand 2014);
- the rare cases of medial position potentially attract illocutionary (rhetorical) comments on the ongoing utterance.

In this section, I will try to verify these expectations through multivariate statistical models combining syntactic and pragmatic variables, focusing mostly on macro-syntactic position and functional domains. First, basic frequency information seems to confirm a number of these hypotheses, as can be seen in Figure 5.11.

We see that about 87% of sequential DMs occur in pre-field ("Pre"), i.e. initial non-integrated position, with only a few anecdotal cases in integrated slots ("Left"

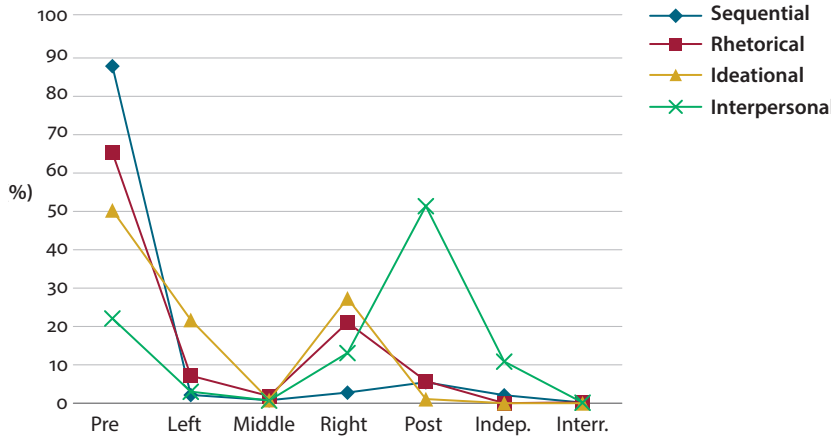


Figure 5.11 Proportions of macro-syntactic slots in each domain

and "Right") and some in post-field ("Post") mostly corresponding to the *closing* function. No other domain is associated to pre-field position in such a proportion (64% in rhetorical, 50% in ideational and 22% in interpersonal). Furthermore, the sequential domain is the only one showing no substantial frequency in the right-integrated field, which indicates its rejection of utterance-internal, syntactically embedded contexts. Such a finding confirms the higher discourse scope of this domain, which deals with turn exchange and topic structure, and not more local relations of content.

The rhetorical and ideational domains do not appear as particularly different based on this graph, apart from a higher proportion of left-integrated occurrences of the latter. Otherwise, they both favor the pre-field position, which is related to non-integrated conjunctions such as *and* or *but*, followed by the right-integrated field which was previously linked to subordinating conjunctions, to introduce either objective or subjective discourse relations. Frequency information does not confirm the expected attraction of rhetorical functions to medial position (here, middle field).

The interpersonal domain is, as hypothesized, strongly associated with final, non-integrated position ("Post") in half of all its occurrences, as opposed to all the other domains where this slot is very rare. A substantial proportion (22%) of interpersonal DMs also occur in pre-field position, although to a much lesser extent than the other domains. Zooming in on these initial interpersonal DMs, we see that they are twice as frequent in English as in French (33% vs. 14%) and mostly correspond to *you know*, *écoutez* 'listen' or *vous savez* 'you know'. These hearer-addressed expressions, built on verbs of knowing or hearing, may have inherited their initial

position from their origins as imperatives (*écoutez*) or "complement-taking mental predicate" (Van Bogaert 2011). The crosslinguistic difference might be explained by the high frequency of French *quoi* and *hein*, which are typically final. In fact, the interpersonal domain is the only one showing major discrepancies between the two languages (more "Pre" and "Right" in English, much more "Post" in French). A final characteristic of the interpersonal domain is its substantial proportion (10%) of independent position, which always corresponds to *monitoring* DMs (e.g. *right*, *okay, hein* 'right').

Examples (21)–(24) illustrate the most frequent pattern for each domain:

(21) we have had (0.310) quite a number of problems with communication (0.750) one of the things we do have we have a service where we have interpreters who (0.420) will come and uh (0.300) translate for us (0.350) *and* another one which has been I found very useful is using the internet (EN-intf-03)

(22) we can take babies from (0.390) the tiniest babies to (0.190) the big (0.360) chunky ones (0.300) uh *so* it's very variable in in (0.230) uh (0.490) what we have to do which (0.200) keeps us interested I think (EN-intf-03)

(23) that accent is spread out into the (0.270) uh (0.390) the parts of North Wales that are very near to the Wirral (0.450) uh *but* the Cheshire side is still very much a Cheshire accent (EN-intf-03)

(24) it must be very frightening to you if you don't know (0.480) can't understand it (0.280) *you know* (0.790) and actually a lot of the time mums just want to know (EN-intf-03)

The significance of these results is statistically confirmed in the following extended association plot (Figure 5.12) showing the strength of association between the two variables. Each rectangle represents the Pearson residuals, that is, the difference between observed and expected frequencies for each category. The width of the rectangle is proportional to the square root of the expected frequency, while the height of the rectangle is proportional to the standardized residuals. The color of the rectangles indicates a positive (blue) or negative (red) association (grey means no significant difference).[25] Extended association plots go beyond mere frequency and show which patterns are significantly more or less frequent, relatively to their competitors.

25. All extended association plots in this research were computed with the assoc function (Zeileis et al. 2007) from the {vcd} package (version 1.3–2, Meyer et al. 2014).

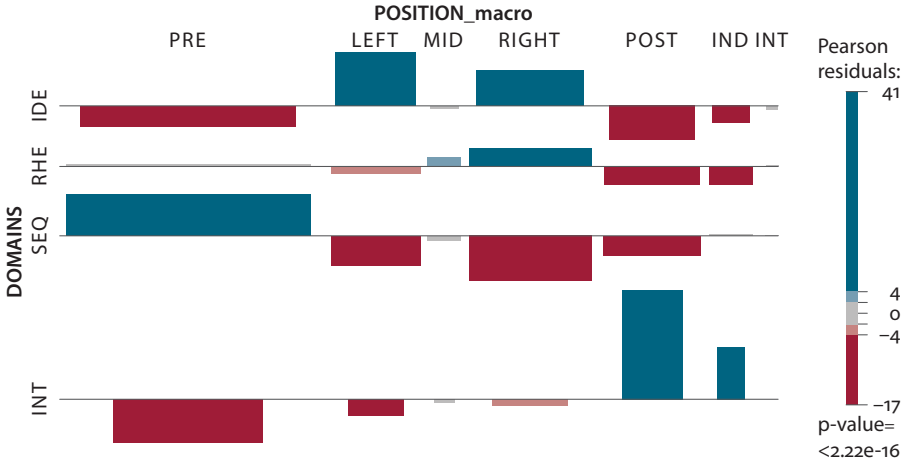


Figure 5.12 Extended association plot of domains and macro-position

Starting with significantly positive associations, this plot confirms (1) the attraction of sequential DMs in pre-field position, (2) the use of interpersonal DMs in post-field and independent positions and (3) the high frequency of ideational and rhetorical functions in right-integrated positions. In addition, this graph now allows us to confirm the hypothesis of medial rhetorical DMs, which was not corroborated by the basic frequency information of the previous figure. However, the association between medial position and sense-altering functions, suggested in Section 5.2.4.4 above, cannot be verified: the *approximation* function indeed appears as the most frequent one in middle-field position, but it is closely followed by more typical discourse relations from both ideational and rhetorical domains (e.g. *specification*, *consequence*). We also see that the attraction of ideational DMs to right-integrated positions extends to left-integrated contexts: objective discourse relations seem intrinsically related to syntax, which is why they are excluded from some DM definitions and taxonomies (e.g. González 2005: 57; Lewis 2006b: 55).

As for negative associations, we see that the pre-field position is dispreferred by both ideational and interpersonal DMs (relatively to the other two domains), in spite of the facts that (1) ideational DMs occur primarily utterance-initially (as in Example (23) above) and (2) some interpersonal DMs (especially in English) can occur initially as well. We can also notice that, besides the pre-field position, all other slots of sequential DMs are either negatively associated or not significantly different from the other domains, making the pre-field position its true distinctive feature. Lastly, rhetorical DMs seem more balanced than the other domains, with no significant monopoly over one particular slot. On the basis of this plot, and in line with the definition of the domains, I would like to suggest the following formal-functional patterns:

- *discourse-relational* functions, either objective or subjective (ideational and rhetorical), show a relative preference for (right-)integrated contexts and a dispreference – or at least absence of significance – for peripheral (pre- and post-field) positions;
- *discourse-structuring* (sequential) functions are strongly (and relatively) associated with the initial position;
- *hearer-oriented* (interpersonal) functions have a relative monopoly on final and independent positions.

Apart from the grouping of ideational and rhetorical functions, these patterns are not fundamentally different from the original definitions of the domains themselves, but the addition of positional information offers some empirical validation to the theoretical categories used in this research, as vouched by the independence of the variables and in line with the programme of corpus-driven cognitive linguistics (e.g. Glynn 2010).

Turning from a descriptive to a more predictive perspective, multivariate statistical models can be used to incorporate multiple factors and evaluate their respective influence on the observed outcome: the more exhaustive and predictive the factors, the more accurate the model. One such method is called Classification And Regression Tree (CART) and it works as a learning algorithm trying to predict the outcome (here, the DM domains) on the basis of the observed data. Statistically significant patterns are classified in different "leaves" on the tree and should be read as follows: the highest nodes in the tree are the most powerful to distinguish the outcomes; values on top of nodes are associated to the branches to their left, leaving the right branch to the remaining unmentioned values. Classification trees are usually reported after "pruning", which is a more conservative method maintaining only the nodes with a high predictive power, thus reducing overfitting (i.e. the model over-generalizes from the data) and improving predictive accuracy. Figure 5.13 displays the pruned classification tree for domains as the outcome and the following factors as independent variables: part-of-speech ("POS"), micro-position ("POSITION_micro"), macro-position, position in the turn ("POSITION_turn") and language. The aim of this analysis is to find out which of these (combinations of) formal features can help predict the domain (ideational "IDE", rhetorical "RHE", sequential "SEQ" or interpersonal "INT") in which the DMs are used.

We see that POS is the most predictive variable impacting the choice of domain, with a first significant divide opposing coordinating conjunctions ("CC"), noun phrases ("NN"), prepositional phrases ("PP"), adverbs ("RB") and subordinating conjunctions ("SC") on the one hand to the remaining four (verb phrases "VP", interjections "UH", adjectives "JJ" and pronouns "WP") on the other.

Position only comes up in a second step and it appears that distinctions at the micro-syntactic and turn level are more significant than macro-syntax, which was

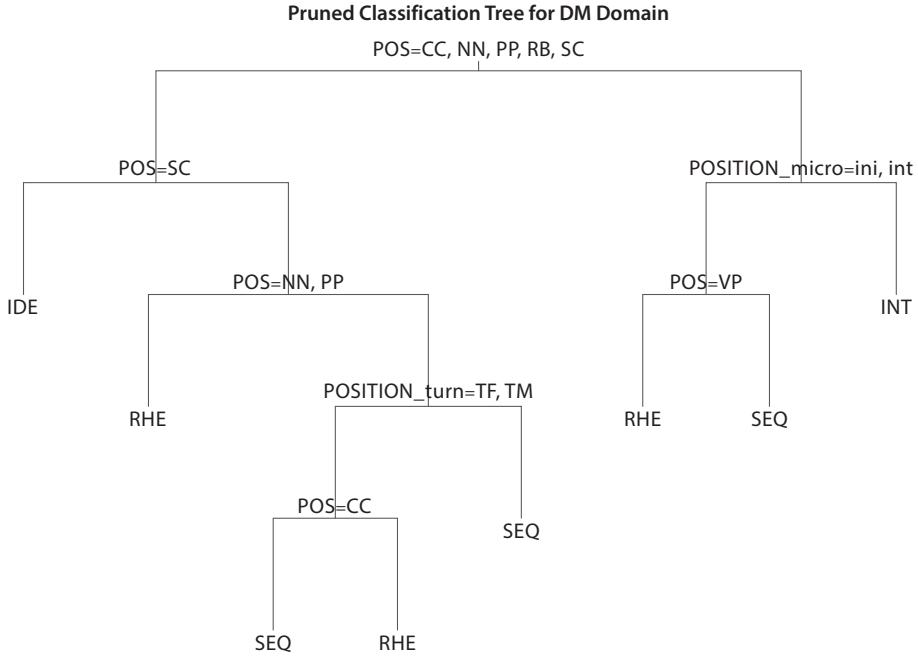


Figure 5.13 Pruned classification tree of domains

discussed so far in this section. This result does however not contradict the significance of the patterns observed above, given the different purpose of each quantitative method (describe vs. predict). We can also note that language differences do not appear significant enough to enter the pruned classification tree. Overall, this graph confirms that the polyfunctionality of DMs is not random but rather formally grounded. I summarize the main patterns in the following:

- *Interpersonal* DMs can be fairly reliably predicted as the combination of the four POS-tags on the right branch of the tree (viz. adjectives, interjections, verbal phrases and pronouns) in non-initial micro-syntactic position (e.g. *you know*).
- *Ideational* DMs are also strongly recognizable by their association to subordinating conjunctions (e.g. *although*).
- *Rhetorical* DMs tend to be expressed either by verb phrases in non-initial micro-syntactic position (e.g. *I mean*), adverbs in non-turn-initial position (e.g. *actually*) or noun phrases and prepositional phrases (e.g. *sort of, in fact*).
- *Sequential* DMs are also spread across three patterns, namely coordinating conjunctions (CC) in turn-medial and turn-final position (e.g. closing *but*), adverbs and CCs in turn-initial position (e.g. *well*, *and*), adjectives, interjections or pronouns in clause-initial position (e.g. French *ben* 'well').

To conclude, a number of form-function patterns have been identified through the mapping of positional and functional variables in increasingly complex statistical models (frequency graph, extended association plot and classification tree). These patterns allow us to associate functions of language in general, and of DMs in particular, to specific slots in the speech string where they are most typical. At this DM-based level of analysis, no further cognitive interpretation of these patterns will be proposed, since they will be refined and potentially questioned by considerations of the fluencemes in their co-text (Chapter 7). Before turning to the combination of DMs with other fluencemes in the typology, one feature of the DM category remains to be discussed, namely the tendency of DMs to directly co-occur with each other.

5.5 Co-occurrence of DMs

The phenomenon of DM co-occurrence is interesting because it is pervasive, especially in spoken discourse, as attested by the many studies in this modality (e.g. Waltereit 2007; Pons Bordería 2008; Cuenca & Marín 2009; Dostie 2013), and relates to the positional flexibility and polyfunctionality of DMs. Co-occurrence is presently understood as formal and immediate contiguity, regardless of syntactic segmentation. It can be assumed that elements occurring recurrently together in the speech string have some sort of connection and (semantic, functional) similarity, following the cognitive-linguistic assumption that people constantly categorize their environment and use this ability to model language (Lakoff 1987).[26] More specifically, I expect the most frequent patterns of co-occurring DMs to express similar or complementary functions and to take scope over the same discourse segments, as a result of their fixation through high frequency in use. By contrast, combinations of DMs occurring in different positions (e.g. final-initial) or expressing different functions and scope (e.g. local objective relation with global discourse structuring) should be less frequent.

This analysis aims to support the general hypothesis underlying this research, according to which recurrent clusters of linguistic elements are cognitively meaningful in that they facilitate production and comprehension. In this perspective, co-occurring DMs should be especially frequent in spontaneous registers, where they may constitute a planning or stalling strategy. Multiple factors influencing

26. Glynn (2010: 8), in particular, discusses the link between co-occurrence and the mental and linguistic processes of categorization as defined and used in cognitive corpus linguistics: "we can say that frequency of co-occurrence, which is fundamental to corpus research, is a quantitative operationalisation of the basic theories of Cognitive Linguistics – entrenchment and categorisation."

DM co-occurrence, both from metadata and DM annotations, will be progressively modeled in an all-encompassing statistical model.

The combination or co-occurrence of DMs is annotated in two ways in *DisFrEn*: very frequent combinations which are well-established in the linguistic community (i.e. non-idiosyncratic) were extracted from a pilot study resulting in a closed list of six "complex" DMs (*and then*, French *et puis, mais bon, bon ben, eh ben, ou sinon*); all other adjacent DMs were simply coded as co-occurring or not, along with their position in the co-occurring string of DMs (first in the string, in-between several DMs, last in the string). "Complex" DMs are presently considered to function as one fixed unit and are therefore not considered as co-occurring.

5.5.1 Co-occurrence across languages and registers

In *DisFrEn*, a total of 1,742 DMs were coded as part of a co-occurring string, which amounts to 20% of all DMs in the corpus: one in five DMs does not occur alone, which is a sufficient rate to confirm our hypothesis of the high frequency of this phenomenon. Counting by number of clusters instead of individual DMs, we find 837 tokens of DM strings, covering 388 different types of combinations, including 254 *hapax legomena*. The combinations with N > 1 are reported in Table 5.14 and ranked by frequency.

We see that the bulk of combinations are relatively rare, with only seven clusters equal or above 10 occurrences, including only two in English (*and so*, *and if*). It also appears that many of these combinations include the basic conjunction *and* / *et*. One tentative interpretation of this finding suggests that the basic and often underspecified meaning of *and* / *et* favors its combination with more explicit DMs such as *so* / *donc* (typically consecutive), *if* or *alors* 'then' (typically conditional).

Crosslinguistically, 155 of these types (including $N = 1$) are English against 233 in French, a difference which roughly corresponds to the gap in relative frequency of co-occurring DMs in the two languages: 16% of all English DMs vs. 24% of all French DMs. To test the significance of this contrastive effect, as well as that of register variation, a mixed-effects logistic regression was computed to predict the co-occurrence of DMs with language and register as input factors.[27]

Mixed-effects logistic regression (also called generalized linear mixed model or binomial logit regression) is a statistical method which is used to model a binary outcome (here, co-occurring or not) from the input of both fixed and random effects, that is, effects which apply to the full population in the sample and effects that

27. Mixed-effect models (both linear and logistic) were computed with the {lme4} package (Bates et al. 2014) on R-studio.

Table 5.14 Combinations of DMs (N > 1) by decreasing frequency

Occ.	English DM clusters	French DM clusters
48		et alors
37		et donc
26	and so	
25		quoi mais
17	and if	
10		mais alors
10		et puis alors
9	but I mean	
8	well if	et comme
7	but if; well I mean	
6	and therefore; because if	enfin je veux dire; et quand; et si; hein mais; mais si; quoi tu vois
5	but when; so when; you know and; you know because	ben écoutez; bon ben écoutez; enfin tu vois; et tout ça quoi; hein donc; mais enfin; mais quand; quoi donc; quoi parce que
4	and when; but then; so if; well you know; you know I mean	alors si; donc quand; parce que quand; quoi enfin; quoi hein; quoi et; tu vois et; voilà donc
3	actually when; and actually; and as; and I mean; and indeed; and in fact; because when; but in fact; but you know; okay so; right well; then when; well actually	alors donc; bon donc; donc en fait; enfin voilà quoi; et en conséquence; et en fait; hein parce; que; mais parce que; parce que sinon
2	actually sort of; and because; and even if; and once; and so on so; and yet; and you know; but anyway; but yes; for instance if; I mean because; I mean when; now if; now then; right so; so I mean; so now; so you know; though and; yeah I mean; yeah so; yeah well; you know when;	ben voilà; donc voilà; enfin hein; enfin voilà; et alors quand; et dès que; et par exemple; et pourtant; et tout ça enfin; et tout ça et; et tout ça quoi tu vois; et tout ça tu vois; et tout donc; et voilà; etcetera alors; etcetera et; etcetera hein; etcetera puis; hein alors hein et; hein etcetera; hein quand; hein si; mais donc; mais enfin bon; mais enfin si; ou si; par exemple quand; parce qu'en fait; parce que bon; parce que bon ben; parce que donc; parce que en fait; parce que si; quoi mais je veux dire; quoi puis; tiens au fond; tu vois enfin; voilà et

are subject-specific, respectively. In other words, mixed models make it possible to account for frequency differences between texts or participants (e.g. a text where co-occurring DMs are very rare against one where they are very frequent), thus enhancing the validity and generalizability of the model to other populations (e.g. Barth & Kapatsinski 2018).

Here, the final model includes language and register as fixed effects (their interaction was tested but not significant) and individual transcripts as random effect, in order to neutralize the weight of DMs produced by the same speakers. The significant effects are reported below:

- French significantly increases the chances of DMs to co-occur by 57% compared to English;
- all broadcast registers (radio interviews, news broadcasts, political speeches and sports commentaries) significantly decrease the chances of co-occurring DMs compared to classroom lessons, while more interactive settings (conversations, phone calls) are not significant (again, compared to classroom lessons).

In other words, the regression confirms the larger tendency of French DMs to co-occur and suggests a divide between broadcast and non-broadcast registers. The contrastive difference can only be related to language-specific preferences and is not surprising in light of previous studies on Romance spoken languages (e.g. Cuenca & Marín 2009 on Catalan and Spanish) showing the high frequency of the phenomenon. The impact of register, on the other hand, is more challenging to interpret beyond a potential effect of formality and presence of a public audience: speakers might refrain from combining several DMs and instead select expressions pragmatically sufficient to convey their intended meaning. No further conclusion can be reliably drawn at this stage.

5.5.2 Co-occurrence across positions

Given the tendency towards initiality of the DM category on the whole, and the special role of unit boundaries for speech planning and processing, DM co-occurrence is hypothesized to favor the initial position. We can see in Table 5.15 that this tendency is only confirmed in raw frequency but not in terms of proportions, since final position shows the highest share (26.26%) of co-occurring tokens over all final DMs.

Table 5.15 Number and proportion of co-occurring DMs across micro-syntactic positions

	Non co-occ.	Co-occ.	Total	% co-occ. by position	% co-occ.
initial	5625	1431	7056	20.30%	82.24%
medial	477	38	515	7.38%	2.18%
final	719	256	975	26.26%	14.71%
indep.	168	15	183	8.20%	0.86%
Total	**6989**	**1740**	**8729**	**19.93%**	**100%**

It appears that, while the bulk of co-occurring DMs are clause-initial (82.24%) as expected, they only amount to one fifth of all initial DMs (20.3%), against one fourth in clause-final position. Two configurations are possible in final position: several DMs cluster towards the end of an utterance (e.g. *quoi tu vois*); an utterance ends with a DM and the following one starts with another (e.g. *quoi mais*). However, this finding is not replicated at turn level, where only 8% of all turn-final DMs are co-occurring (against 14% of turn-initial).

Most cluster types are specific to one position only, although a very restricted number (nine) can occur in both initial and final position or, even more rarely, initial and medial or medial and final. For instance, *enfin tu vois* is clause-initial in Example (25) and clause-medial in (26).

(25) je vois ma grand mère elle elle a mal partout (2.330) mais elle a sa t- *enfin tu vois* elle est encore juste quoi
I see my grandmother she she hurts everywhere but she has her h- enfin tu vois '*well you see*' *she is still sane* (FR-conv-05)

(26) oui mais c'est un drôle de petit homme *enfin tu vois* qui fonctionne dans tous les …
yes but he's a funny little man enfin tu vois *'well you know' who works in all the …* (FR-conv-04)

Zooming in on clause-final clusters, it appears that the French DM *quoi* 'you know' is very often involved in a co-occurring string (30% of all *quoi* are co-occurring, against 10% of *and*, for instance), especially in a cluster with *mais* 'but' (4th most frequent cluster overall, cf. Table 5.14). The prominent place of *quoi* among co-occurring DMs might be a sign of its underspecification or ambiguity. Beeching (2007: 148) describes this DM as "virtually desemanticized", which might explain why speakers tend to combine it with other DMs to reinforce its pragmatic and inferential meaning, an interpretation which I already suggested when dealing with double-tagged DMs (cf. Section 5.3.3).

Other frequently co-occurring DMs in final position include *you know*, *hein* 'right' and *et tout ça* 'and all that'. Together with *quoi*, these speech-specific expressions often combine with more universal DMs such as conjunctions, as in Examples (27) and (28).

(27) il est très courtois faut faut pas faut pas (0.500) trop lui demander *quoi mais* euh je veux dire euh ça c'est ça dépend un peu *quoi mais* il s'est quand même vachement calmé
he is very courteous you you can't you can't ask too much quoi mais *'you know but' uh I mean uh it it's it depends a bit* quoi mais '*you know but' he did calm down a lot* (FR-conv-05)

(28) the cinema in a way is like (0.513) children's bedtime stories you kn- *you know and* it always seemed that way to me (0.190) just simple really (EN-intr-04)

Two occurrences of the "quoi mais" cluster appear in (27): each time, the speaker ends a rather generic, common-knowledge utterance ("faut pas trop lui demander", "ça dépend un peu") with a "quoi" and starts again with the contrastive conjunction "mais" to revise the previous statement. In (28), similarly, the speaker calls for the hearer's cooperation on her comparison of cinema with bedtime stories with the help of "you know" and goes on developing her statement with "and".

It is interesting to note that very different DMs such as these regularly combine in the speech string, an observation which constitutes a first limitation to the hypothesis of the similarity between co-occurring DMs. It might rather be the case that clustered DMs are more complementary than redundant, thus bridging the gap between speech-specific DMs on the one hand and more universal DMs on the other.

5.5.3 Integrated statistical model of co-occurrence

The results discussed so far seem to point to a multiplicity of factors influencing the tendency of DMs to co-occur, both from the general context (language, register) and linguistic behavior of the DM (position, semantics). I will now take up the previous regression model (with language and register as factors) and integrate more variables which I hypothesize to impact the tendency of DMs to co-occur. This full model includes, as input factors: language, register, POS, domains (including double tags), whether or not the DM was assigned a functional double tag, micro-syntactic position, position in the turn, as well as the variation within individual transcripts as random effects (cf. above).

As for language and register, the full model reports the same effects as the previous one (cf. Section 5.5.1), namely a preference for French and non-broadcast contexts. The other fixed effects of the final model are the following:

- *POS* (reference level: coordinating conjunction): all POS-tags except interjections are significantly more prone to co-occurrence than the reference level.
- *Domains* (reference level: ideational): the interpersonal and sequential domains (and combinations thereof) are significantly more prone to co-occurrence than the reference level.
- *Micro-position* (reference level: initial): independent and medial micro-positions are significantly less prone to co-occurrence than the reference level.
- *Position in the turn* (reference level: turn-initial): turn-medial positions are significantly more prone to co-occurrence than the reference level, while turn-final DMs are less prone to co-occurrence.

This regression confirms previous frequency results regarding the mismatch between final position in the clause and in the turn. In addition, it uncovers an effect of functional domains, which distinguishes interpersonal and sequential DMs on the one hand from ideational and rhetorical DMs on the other. This divide seems to point to a difference in semantic content, strength or (under)specification. In fact, it is particularly interesting to note that the interpersonal and sequential domains, which are, by the nature of the functions they include, more specific to speech than the other two more universal domains, are often involved in DM co-occurrence, which I interpret as evidence of the higher attraction of this phenomenon to the spoken modality (as argued in Crible & Cuenca 2017). There is, however, no effect of single vs. double tagging, against what was suggested in the analysis of double domains in Section 5.3.3.

Overall, co-occurrence of DMs is not random but seems to favor certain types of DMs in certain contexts, which points to a discourse-functional motivation behind their use. Statistical regressions are useful to decipher the relevant factors in a phenomenon affected by great variation such as DMs. However, a more fine-grained view of different types of co-occurrence (as the one in Cuenca and Marín 2009) would require qualitative methods of investigation to integrate corpus annotations in a more interpretative model (see Crible 2017b).

5.6 Summary

This chapter developed and discussed the major corpus-based findings regarding the behavior and variation of DMs, including only the variables annotated at DM level. Crosslinguistically, besides a higher frequency in French of DMs in general, and of utterance-final interpersonal DMs in particular, the two languages do not appear to differ in major ways. One possible explanation for this similarity is language contact, given the historical influence of French on English and the current overwhelming presence of English, although this factor would require additional evidence and is beyond the scope of this research.

Register, however, greatly impacts the distribution of DMs, which are favored by spontaneous dialogues, as expected. In particular, we saw that formal and informal registers are not only distinguished by frequency of occurrence but also, more interestingly, by the types and uses of DMs they seem to favor (e.g. more ideational, syntactically integrated DMs in formal settings). A number of language-specific and register-specific patterns were also identified, such as pronoun-based DMs in final position in conversational French (*quoi* 'you know') or noun-based medial DMs in English (*sort of*).

The bottom-up approach to corpus data allowed us to confirm the centrality of some DM features usually mentioned as criterial in the literature (e.g. initiality,

discourse-structuring role and tendency to co-occur) and to identify the proportions and conditions under which DMs diverge from their typical portrait. For instance, while the majority of DMs in *DisFrEn* come from the grammatical class of conjunctions, thus confirming their typical association in many definitions, we saw that the polyfunctionality of the DM category is in fact best represented by adverbs (second most frequent POS-tag), which are not restricted to any functional domain.

Thanks to the independence and flexible granularity of the variables, descriptive univariate patterns were refined by integrating more and more features both formal and functional. The following configurations are particularly noteworthy as they were identified across various levels of the analysis: coordinating conjunctions in pre-field position marking discourse structure; subordinating conjunctions in both left- and right-integrated position signaling discourse relations; adverbs in medial position expressing speakers' meta-comments and interjections as independent units serving interactional (speech-segmenting, interpersonal) purposes.

5.7 Interim discussion: The potential of bottom-up research

One potential caveat to this crosslinguistic portrait of DMs in English and French lies in the limitations of the representativeness of the corpus, notably regarding the different dates when the recordings were collected. In particular, the ICE-GB corpus, which constitutes the majority of the English transcripts in *DisFrEn*, dates back to 1990–1991, whereas some of the French corpora are more recent (e.g. LOCAS-F, C-Humour, cf. Section 4.1.1). This difference in the data might hinder the comparability between the two languages under scrutiny and potentially introduce a bias in the interpretation of the results.

The date of corpus collection is particularly relevant in DM research since these expressions have been shown to be strongly affected by diachronic change (e.g. Waltereit & Detges 2007 on French vs. Spanish *bien* 'right'; Hansen 2008 on French phasal adverbs). Such a limitation possibly overlooks emerging uses of DMs in English or French which could explain the differences observed in this chapter. For instance, the low frequency of final DMs in the English data is surprising in light of recent works on final *but* (Mulder & Thompson 2008; Izutsu & Izutsu 2014) and other expressions (Haselow 2012 on final *then*, *though*, *anyway*, *actually* and *even*). Controlling for such external factors would definitely provide an interesting avenue for further research.

This chapter was resolutely quantitative and mainly descriptive, combining univariate and multivariate analyses with statistical tools of increasing complexity. While formal considerations alone remain rather limited to a partial frequency-based portrait of the DM category, the integration of syntax and pragmatics in multivariate

models proved more innovative and relevant to the investigation of form-function patterns undertaken in this usage-based research. Such a flexibility in the analysis is thought to test and explore the potential of frequency-based corpus studies which can be more than purely descriptive but also theoretically relevant. This role of frequency, which is assumed to be central in all levels of language according to the usage-based framework, is here considered on equal grounds with other factors (categorical variables and metadata) as potentially telling of underlying cognitive processes.

Beyond its descriptive purpose, this chapter illustrates the advantages (and shortcomings) of relatively large datasets – even though *DisFrEn* is small with respect to most written corpora – and their exploration through statistical methods and flexible levels of analysis. Such a bottom-up approach to categorical phenomena (here, DMs) allows the analyst to (1) maintain a level of objectivity regarding the results, avoiding the circularity of "finding what one is looking for" and (2) to select, on the basis of this bottom-up method, the most relevant variables and levels of analysis or, as Gries (2011: 238) puts it, "the degree of granularity that provides the most insightful results". In the complex and highly variable field of discourse, the analyst cannot be sure beforehand what particular variables will answer their research question(s), which suggests two recommendations: to cover a wide array of variables and account for their combination at different degrees of granularity (e.g. position by domain, POS by function, etc.); to keep an open mind towards the data. Gries (2011: 254) summarizes the situation as follows:

> The distinctions one brings to the data as an analyst *a priori* need not at all coincide with the largest differences in the data, those that are actually reflected in the data, or those that are most noteworthy or theoretically revealing.

He argues that such an attitude is highly compatible with the usage-based model, which is grounded on the combination of frequency, formal and functional patterns.

In the next chapter, discourse markers will be studied in combination with other members of the disfluency typology, in order to describe this broader category of linguistic devices from the point of view of their combination and variation. Discourse marker features, as investigated in Chapter 5, will be integrated in the study of fluenceme sequences in Chapter 7.

CHAPTER 6

Disfluency in interviews

One of the main lines of investigation of this research is to situate DMs within the typology of fluencemes, thus addressing a gap in the literature given the irregularity with which previous studies have included DMs in corpus-based annotations of fluency (cf. Section 3.4). Such an integrated view of fluencemes is necessary to test the first general hypothesis concerning their syntagmatic behavior, namely that fluencemes tend to occur more frequently in clusters than in isolation. This tendency is expected to be largely due to the pervasiveness of unfilled pauses, as well as the high frequency of DMs in dialogues.

In order to assert such a general conclusion, the analysis needs to go beyond "textual frequency" (i.e. the frequency of a given structure in the corpus; fluenceme-by-fluenceme approach) and aim at "conceptual frequency" (i.e. the frequency of a structure with respect to all its competitors in the category; paradigmatic approach), taking up Hoffmann's (2004) distinction defined in Section 2.5.4. Only then can we model the inter-relations between each fluenceme type, identify recurrent patterns of combination and interpret these clusters in light of their features and distribution. The hypothesis of fluency-as-frequency (cf. Section 2.5.4) lies at the core of the following analyses, looking for evidence in support of its cognitive validity as well as its limitations.

6.1 Data

The analyses in this section will make use of the subcorpora of face-to-face and radio interviews, where all occurrences of fluencemes have been identified regardless of their type or position, as opposed to the remainder of *DisFrEn* where only fluencemes clustered with a DM have been annotated. As a result, the following analyses are limited in terms of register variation: face-to-face and radio interviews only differ in their degrees of elicitation (semi-elicited vs. natural) and broadcasting (non-broadcast vs. broadcast, respectively). The latter was found to have a significant effect on the co-occurrence of DMs (cf. Section 5.5.1).

Apart from register, the literature review did not suggest any crosslinguistic expectation regarding the distribution of fluencemes in English and French, apart from some quantitative differences uncovered by Grosjean & Deschamps

(1975) – although comparability between corpora and annotation schemes is never fully achievable. The following sections will therefore focus on identifying both language-specific and shared patterns across different types (or macro-labels, cf. Section 4.3.5) of sequences, striving to test the hypothesis on fluenceme clustering and providing tentative interpretations of (dis)fluency.

6.2 Fluenceme rates in English and French

6.2.1 Number of tags

As explained in Section 4.3.5, the content of the sequences can be queried with varying degrees of granularity at sequence level. A global idea of the rate of fluencemes is provided by counting each fluenceme tag in the corpus, that is each word tagged as (part of) a fluenceme (e.g. a repetition of a 10-word segment will count as 20 tags). Such a counting unit returns the proportion of the data (in number of words) which is involved in or covered by any (dis)fluent marker from the typology. As a result, "fluent" uses of fluencemes as well as the *reparans* part of a fluenceme (e.g. the second "I" in "I I think") will also be counted, thus potentially overestimating the rate. While a more conservative measure will be provided below, this first measure of frequency remains interesting in that it accounts for the actual length of fluenceme sequences. The results will now be presented and discussed.

Excluding unfilled pauses, 10,477 words were assigned one (or more) tag(s) from the typology (4,645 in English, 5,832 in French), which amounts to 20.04% of all words in interviews overall (17.98% in English, 22% in French).[28] In other words, one in every five words is (part of) a fluenceme, which points to the pervasiveness of the phenomenon in spoken language. This rate is higher than what most studies report in the literature, such as Bortfeld et al. (2001) who report a rate of 5.97 disfluencies per hundred words in their corpus of conversational English. A number of explanations can be proposed to account for this difference. Methodologically, the scope of the annotations is wider than in most previous works, with the inclusion of typically "fluent" devices in the typology such as modified repetitions as well as the broad coverage of DMs. The present 20% rate therefore covers potentially fluent

28. Since the present rates are given per number of words in the corpus (e.g. 20 words out of 100 are fluencemes) and unfilled pauses are not included in the word count, it would be erroneous to compute the rate of unfilled pauses per total number of words in the corpus. This issue, however, does not concern the relative frequency of unfilled pauses (how many pauses occur in the span of 1,000 words) which is provided in Table 6.1 below.

and disfluent devices alike, similar structures which are ambivalent (e.g. stuttering repetition vs. enumerating repetition), discourse markers signaling local cohesive relations and others more related to the interactive and spontaneous nature of speech. By contrast, Bortfeld et al.'s (2001) rate only includes repetitions, "restarts" (truncations and false-starts) and "fillers" (e.g. *uh*, *ah*). Methodological differences therefore certainly play a decisive role in the reported rate of fluenceme tags.

Another explanation is empirical and suggests an effect of data type. The present interview data can be expected to include more fluencemes than more casual (conversation, as in Bortfeld et al. 2001) or formal (political speech) registers, following the hypothesis on intermediary settings: the heigtened degree of speaker's attention towards their own production, coupled with the low degree of preparation, may explain why these registers have previously been found to contain more disfluencies (Broen & Siegel 1972). The effect of register is, however, challenging to assess reliably in the present study since the two registers where all fluencemes have been annotated, namely face-to-face and radio interviews, are quite similar and cannot serve to test hypotheses on the role of preparation or interactivity, for instance. Moreover, comparing fluenceme rates across registers from several corpora annotated with different typologies and procedures runs into the issue of inter-operability, which relates back to the methodological differences noted above.

6.2.2 Number of tokens

Fluencemes in the corpus can also be counted per number of fluenceme tokens, which makes it possible, in particular, to situate fluencemes with respect to each other regardless of their internal structure or respective size (e.g. a repetition of a 10-word segment will count as one occurrence of repetition, a word tagged as both a DM and a repetition will count as one occurrence of each, etc.). The relative frequency of all fluencemes in face-to-face ("ftf") and radio interviews is reported in Table 6.1.

We see that, in both languages and registers, the top two fluenceme types are the same, namely unfilled pauses ("UP") and discourse markers ("DM"). This result is not surprising given the particularly high ambivalence of these two simple fluencemes, which can range from quite disruptive uses to more segmenting or hearer-oriented functions. By contrast, we see that fluencemes which are described as typical disfluencies, such as false-starts ("FS") or explicit editing terms ("ET") are much less frequent in the studied register. It should be noted that the assumption of functional ambivalence also applies to these fluencemes, so that, in principle, not all occurrences of FS and ET are necessarily disfluent.

Table 6.1 Relative frequency (per thousand words) of fluenceme tokens in each subcorpus

	English		French		Total	
	ftf	radio	ftf	radio	ftf	radio
Unfilled pause (UP)	110.88	78.31	87.62	64.52	98.92	71.56
Discourse marker (DM)	62.86	54.60	71.88	52.16	67.50	53.41
Filled pause (FP)	30.96	13.56	22.83	19.84	26.78	16.64
Identical rep. (RI)	11.84	16.98	14.96	17.94	13.45	17.45
Truncation (TR)	5.34	5.02	4.77	6.06	4.87	5.53
Modified rep. (RM)	4.98	3.53	5.82	6.18	5.58	4.83
False-start (FS)	3.87	3.19	7.43	6.18	5.30	4.65
Propositional sub. (SP)	3.05	1.94	2.94	2.26	3.39	2.09
Morphosynt. sub. (SM)	0.76	0.34	2.94	2.85	1.88	1.57
Editing term (ET)	0.18	0.11	0.94	0.24	0.56	0.17
Total	234.71	177.59	222.13	178.23	228.25	177.90

The interview data does not make it possible to draw strong conclusions on register variation beyond the effect of broadcasting. Nevertheless, a number of observations can be made on the basis of Table 6.1 regarding differences in distribution:

- unfilled pauses are the only fluenceme consistently more frequent in English than in French across the two interview settings;
- DMs and identical repetitions ("RI") are significantly more frequent in French than English face-to-face interviews ($LL = 6.38$, $p < 0.01$);
- filled pauses ("FP") are more frequent in English face-to-face interviews (compared to French ones) but less frequent in English radio interviews than in French ones;
- all other differences are much smaller.

Mixed-effect logistic regressions have been computed for each fluenceme (including language, register and individual random effects when they improved the model). The main significant effects corroborate the frequency findings from Table 6.1 and can be summarized as follows: more DMs, RIs, RMs (modified repetitions) and FSs in French; more UPs in English; more FPs in face-to-face interviews. In other words, English and French seem to favor different types of fluencemes. As a result, the higher frequency of DMs in French discussed in the previous chapter cannot be extended to all other fluencemes in the typology, especially because of the substantial weight of unfilled pauses in English.

Overall, the total number of fluenceme tokens is not significantly different between the two languages once the two subregisters are combined (5561 vs. 5508;

$LL = 3.14$, $p > 0.05$), although the frequency of fluencemes in face-to-face interviews is significantly higher in English than in French ($LL = 6.07$, $p < 0.05$). The higher frequency of UPs in English stands out as the major crosslinguistic difference in this table, while all other differences are much smaller, with the notable exceptions of FPs, DMs and, to a lesser extent, RIs mentioned above.

While comparison of fluenceme rates to corpora using different annotation schemes is prohibited by the differences in scope and definitions, the present findings can be directly mirrored with the frequency results reported in a comparative study (Crible et al. forthc.) where the same typology was applied to data in native French, native and learner English and Belgian French Sign Language by four different annotators. The authors found that relative frequencies of individual fluencemes are highly similar across corpora, especially between the native languages and for fluencemes such as unfilled pauses ("UP"), modified repetitions ("RM") or false-starts ("FS"). The ranking is also similar across the various spoken languages, with pauses (unfilled, then filled) and identical repetitions on top. It is particularly interesting to note that these fluencemes which, along with DMs, hold a prominent place in the typology, all correspond to what Ginzburg et al. (2014) call "forward-looking disfluencies", that is, structures which do not modify already-uttered speech but announce or signal the incoming completion of the on-going utterance.[29] In other words, fluencemes are omnipresent in speech production, covering about one fifth of the sound signal, and such momentary interruptions of the smooth unfolding of speech mostly attend to the upcoming rather than previous material.

6.2.3 Radio vs. face-to-face interviews

Regarding the effect of broadcasting on fluenceme frequency, we can observe an overall difference with more fluencemes in non-broadcast (face-to-face) interviews in both languages. However, this difference does not affect all fluencemes equally. Unfilled pauses and DMs are more frequent in face-to-face interviews in the two languages. Filled pauses are twice as frequent in English non-broadcast as in broadcast interviews ($LL = 68.08$, $p < 0.001$), while this difference is not significant for French. On the other hand, the effect is reversed for identical repetitions (more frequent in radio than face-to-face, $LL = 12.19$, $p < 0.001$). Lastly, the difference is not significant for truncations ("TR"). RIs stand out as the only fluenceme type showing

29. Cf. also Levelt's (1983) "covert repairs".

a major preference for the broadcast context, especially in English (LL = 18.12, $p < 0.001$). This result might suggest a specific "radio style" whereby speakers tend to repeat themselves either for rhetorical or stylistic effects.[30]

The generally higher frequency of fluencemes in the face-to-face setting may first be interpreted as the result of a potentially lower degree of preparation in non-broadcast interviews, where the interviewee does not necessarily know in advance all the questions which he or she will have to answer, as opposed to the generally "rehearsed" setting of radio shows (whether or not the interview was rehearsed is not available in the metadata). A second explanation involving the role of topic familiarity could be proposed but is harder to assess since this variable was not controlled for in the metadata. The interviews in *DisFrEn* cover quite distinct types of topics: sociolinguistic interview in some of the French face-to-face texts, personal experience in some others; questions about one's profession in the English face-to-face interviews; questions about the artist's current work (comedy show, new book, etc.) in all English and French radio interviews. In any case, a previous study on the relation between fluency and topic familiarity, conducted by Merlo & Mansur (2004), showed that it is not the frequency of disfluencies but rather the types of disfluencies which are affected by differences in familiar vs. unfamiliar topic.

A third potential explanation for the observed quantitative difference is the different degree of professionalism between the speakers in the two settings. All interviewees in the broadcast interviews are artists or public-speaking professionals (e.g. humorists, novelists), hence potentially more comfortable than the range of speakers from a variety of backgrounds and professions in the non-broadcast interviews (e.g. nursing home manager, nurse, CEO, factory worker). These interpretative leads cannot be tested any further but illustrate the benefits of detailed text and speaker metadata as a research avenue for future studies.

A last observation at the fluenceme level focuses on modified repetitions ("RM"), which can be expected to occur more frequently in broadcast registers because of the different possible uses of this fluenceme. More specifically, modified repetitions represent any repeated material which includes some modification of form and/or content, and therefore cover very different phenomena such as enumerations built on a repeated syntactic anchor or actual corrective reformulations. Given this ambivalence, modified repetitions might be expected to be used relatively more often in the professional setting of radio interviews, where the speakers are trained to speak publicly and even creatively, as already mentioned

30. This interpretation evokes Léon's (1993) and Simon et al.'s (2010) notion of *phonostyle*, which is defined as a speaking style mostly based on prosodic features and characterizing a speaker, social group or specific setting.

above. Radio interviewees might thus resort to this fluenceme type for rhetorical purposes besides more "disfluent" uses. In the interview data, however, this hypothesis is neither confirmed in English, where the frequencies are reversed, nor in French, where the higher frequency of RMs in radio interviews is not statistically significant ($LL = 0.12$, $p > 0.05$). Yet, a qualitative exploration of the occurrences in each subcorpus uncovers typical patterns of use for this fluenceme which are used differently across broadcast and non-broadcast interviews, as illustrated in the following examples respectively:

(1) une des choses qui m'avaient retenue qui m'avaient bouleversée (0.633) en lisant les quelques biographies de de Hendrix qui existaient (0.482) c'est *qu'il était d'une timidité extrême dans la vie (0.822) et qu'il était d'une audace extrême sur scène*
one thing that caught my eye that moved me when I read the few existing biographies of of Hendrix is that he was extremely shy in life and he was extremely bold on stage (FR-intr-04)

(2) <VAL_3> on ne peut pas dire que on parle sans accent ou sinon vous *ne sauriez pas parler*
<VAL_2> tout à fait
<VAL_3> *ne pourriez pas parler* plutôt
<*VAL_3*> *we cannot say that we speak without an accent otherwise you* would not be able to speak
<*VAL_2*> *exactly*
<VAL_3> could not speak *rather* (FR-intf-01)

We can see in Example (1) that the modified repetition of "qu'il était d'une… extrême…" serves an enumerating, even contrastive purpose which reflects the literary skills of the speaker (an autobiographer). By contrast, in Example (2), the speaker (a CEO) substitutes one modal verb ("sauriez") by another more standard one ("pourriez"). The postponed editing term "plutôt" ('rather') further corroborates this reading as a lexical error in need of correction. Further comparison of registers regarding the "fluent" vs. "disfluent" uses of modified repetitions would require quantification of these differences through a systematic categorization of RM types, in order to uncover the multi-faceted nature of this fluenceme (cf. the approach to repairs in Chapter 8).

6.3 Clustering tendencies

6.3.1 Isolation vs. combination

It appears that analyses at fluenceme level are limited to basic information of rates since one fluenceme type can cover multiple uses, given their intrinsic ambivalence. I have previously argued (cf. Section 2.5.2) that fluencemes, and as a general rule any linguistic item, should be studied in their local context of occurrence in order to account for their combinatory patterns. Indeed, the very first hypothesis of this research is that fluencemes are more often clustered than isolated, as already observed in previous fluency research, thus confirming the tendency in spoken communication to "pack together" similar elements.

A basic way to test this hypothesis is to look at sequence length in number of fluenceme tokens, measuring the proportions of sequences of more than one fluenceme. In the data, 57.66% of the 6,315 annotated sequences are single, isolated fluencemes (55.64% in face-to-face and 62.42% in radio interviews). This result does not allow us to confirm the hypothesis on fluenceme clustering, although not by much. The proportions of different sequence lengths by subcorpus are reported in Table 6.2. We see that the bulk of sequences in the data include up to three fluenceme tokens (together more than 90% of all sequences, including isolated fluencemes), while sequences of six or more fluencemes are anecdotal, up to a maximum value of 15. This decrease is strikingly similar across registers and languages, which points to a stable tendency of (very) short sequences.

Table 6.2 Sequence length (in number of fluenceme tokens) by register and language

	Face-to-face interviews				**Radio interviews**			
	EN	**%**	**FR**	**%**	**EN**	**%**	**FR**	**%**
1	1239	54.85	1231	56.47	703	67.73	468	55.85
2	608	26.91	539	24.72	212	20.42	213	25.42
3	222	9.83	189	8.67	82	7.90	82	9.79
4	110	4.87	99	4.54	22	2.12	41	4.89
5	39	1.73	57	2.61	12	1.16	16	1.91
6	21	0.93	23	1.06	3	0.29	9	1.07
7	7	0.31	13	0.60	2	0.19	4	0.48
8	7	0.31	15	0.69	0	0.00	0	0.00
9	2	0.09	10	0.46	0	0.00	2	0.24
10	3	0.13	3	0.14	0	0.00	1	0.12
11	0	0.00	0	0.00	1	0.10	1	0.12
12	1	0.04	0	0.00	0	0.00	0	0.00
13	0	0.00	1	0.05	0	0.00	1	0.12
15	0	0.00	0	0.00	1	0.10	0	0.00
Total	**2259**	**100.00**	**2180**	**100.00**	**1038**	**100.00**	**838**	**100.00**

Yet, the results of a linear mixed-effect regression show a significantly higher likelihood of longer sequences in French and shorter sequences in radio interviews, in a model with language and register as fixed effects, individual transcripts as random effect and no significant interaction of factors. These significant effects, however, only account for a small percentage of the variance in the data (conditional $r^2 = 0.04$), which means that additional factors are responsible for the variation in sequence length besides language and broadcasting. The observed differences between each subcorpus remain valid and in line with previously obtained results.

6.3.2 Most frequent clusters

The specific clusters which lie behind this numeric information will now be presented, as well as which fluencemes are most attracted to one another, in order to test a number of hypotheses and claims laid out in Chapter 2. I will start with the most specific level of granularity, namely the actual instances of fluenceme clusters and leave more abstract categories or macro-labels for tentative interpretations of relative fluency in the next section. In the interview data, 577 different types of clusters were found, of which only nine show more than 100 occurrences. They are reported in Table 6.3.

Table 6.3 Relative frequency of sequences (N > 100) ptw by language and register

	English			French		
	ftf	**radio**	**total**	**ftf**	**radio**	**total**
UP	45.85	45.48	45.73	35.69	26.02	32.62
DM	17.88	22.23	19.36	19.95	17.59	19.20
UP + DM	14.31	9.80	12.78	11.47	7.01	10.05
FP	4.46	3.19	4.03	5.27	4.52	5.03
RI	2.87	7.52	4.45	4.27	4.52	4.35
UP + FP	8.85	1.48	6.35	2.11	1.66	1.97
DM + UP	2.35	2.05	2.25	2.38	2.14	2.31
RI + UP	2.05	1.71	1.94	2.49	2.61	2.53
FP + UP	2.23	1.37	1.94	2.11	2.73	2.31

These nine patterns of sequences all include the same four types of fluencemes, namely unfilled pauses ("UP"), discourse markers ("DM"), filled pauses ("FP") and identical repetitions ("RI"), which correspond to the most frequent fluencemes overall when counting by individual tokens instead of sequences. We see that this top nine includes both isolated and clustered uses of these four fluencemes, starting with UP and DM (in isolation and then in combination as UP + DM) and followed by combinations of UP with the other three fluencemes, sometimes in each order (UP + DM and DM + UP, UP + FP and FP + UP).

In sum, the results in Table 6.3 point to the important role of unfilled pauses in clustering: highly frequent clusters always include an unfilled pause, either as the first or second fluenceme in the sequence. The pervasiveness of unfilled pauses reflects their high functional ambivalence, from purely physiological respiratory reasons to segmentation and planning purposes. The detailed configurations and contexts of DM + pause clusters (DM with UP and/or FP) are the focus of the study by Crible et al. (2017).

Regarding register and language effects, we notice once more that not all fluenceme sequences are affected equally. For instance, isolated UPs show the exact same relative frequency in the two settings in English, while the difference is more clearly marked and significant in French ($LL = 17.14$, $p < 0.001$). Most sequences are either not significantly different between the broadcast and non-broadcast situations or favor the latter, except for isolated DMs and RIs in English radio interviews.

Crosslinguistically, the only major difference consists in the higher frequency of sequences containing a UP (UP, UP + DM, UP + FP) in English than in French. In particular, the UP + FP cluster stands out from the other less frequent sequences with a substantial frequency in English face-to-face interviews (8.85 sequences ptw), which impacts the overall ranking of these sequences in the two languages (4th and 9th). In sum, language and register variation do not affect the same fluenceme sequences and not always to the same effect (e.g. strong effect of broadcasting on isolated RIs in English but no contrastive difference once both settings are combined).

6.3.3 DMs in clusters

Table 6.3 also provides an answer to the hypothesis gathered from Boula de Mareüil et al. (2013), who found that DMs more often precede than follow other disfluencies. Based on the top-nine clusters reported here, we see that this is not the case in the interview data, where the UP + DM cluster is much more frequent than its reverse order DM + UP. If we extend the results to all sequences in interviews containing at least one DM and one other fluenceme, it appears that DMs are the first element in only 426 sequences, leaving a great majority of 1,188 sequences (73.61%) where DMs occur in the middle or at the end of the cluster. Boula de Mareüil et al.'s (2013) finding is therefore not confirmed by the present results.

Another observation on DMs in clusters takes up Beliao & Lacheret's (2013) findings on the relative independence of "prosodic disfluencies" with respect to DMs. They found that the proportion of clustered DMs is higher than that of clustered disfluencies, that is, pauses attract DMs more than DMs attract pauses. In the interview data, it can be observed that 57% (3,618 clusters) of all sequences

(not only restricted to pauses) do not contain a DM, while 41% (1,083 clusters) of sequences containing one or several DMs do not include any other fluenceme type. In other words, as in Beliao & Lacheret (2013), sequences of fluencemes are more "independent" from DMs than vice versa since the majority of DMs cluster with other fluenceme types. However, considering that DMs are but one out of nine types of fluencemes, their presence in 43% of all sequences in interviews is quite substantial and argues for their prominent place in the typology.

In fact, these results nuance our previous rejection of the hypothesis regarding the general clustering tendency of fluencemes. For DMs alone, we do observe a higher frequency (59%) of clustered vs. isolated contexts, against only 48% for unfilled pauses, for instance. Overall, the very high frequency of isolated UPs, observed in Table 6.3, is in part responsible for (1) the general ranking of sequences, (2) the proportion of isolated and clustered fluencemes and (3) the difference of relative "independence" between DMs and other fluencemes.

6.4 Fluency as frequency

The paradigmatic annotation of fluencemes in interviews provides first insights into the fluency-as-frequency hypothesis: does the combination of fluencemes give any clue regarding the relative fluency of the sequences at different degrees of abstraction? Based on the usage-based assumption that high frequency of use contributes to cognitive entrenchment, I expect rare sequences to be more marked and potentially more disfluent than very frequent fluencemes, which should be more accessible and less intrusive for production and comprehension. Given the great variability of fluenceme sequences (cf. 577 different types of clusters), the analyses in this section will resort to various ways of summarizing the content of sequences and try to identify which macro-label(s) better fit(s) the fluency-as-frequency hypothesis.

6.4.1 Frequency and structural complexity

In this section, sequences are grouped in 10 categories based on the structural "complexity" of the fluencemes they include. At this degree of abstraction, sequences are distinguished based on (1) the structure of the fluenceme(s) (simple or compound), (2) the number of fluencemes (one or multiple) and (3) whether simple fluencemes co-occur with compound ones and in what position (within, peripheral, both). As explained in Section 4.3, the first distinction (simple vs. compound) is based on the definition of each fluenceme and is provided by

Crible et al.'s (2016) typology and annotation guidelines. Simple fluencemes comprise pauses, DMs, editing terms, false-starts and incomplete truncations, and roughly correspond to Levelt's (1983) "covert repairs" and Ginzburg et al.'s (2014) "forward-looking disfluencies". Compound fluencemes cover repetitions, substitutions and completed truncations.

At a general conceptual level, the occurrence of compound fluencemes can be expected to be more disruptive in the utterance and to signal the presence of linguistic material in need of repairing, especially in clusters with additional simple fluencemes. Embedded or peripheral pauses and DMs can be interpreted as signals of an upcoming or ongoing disfluency such as a reformulation. This generalization, however, does not account for the ambivalence of fluencemes such as modified repetitions, which can be involved in either "fluent" enumerations or "disfluent" corrections. The objective of this section is therefore not to draw firm conclusions on the relative (dis)fluency of sequences solely based on their content, but rather to test the extent to which the combination of objective cues (sequence structure, sequence length, frequency) maps a more fine-grained examination of specific sequences in the corpus, zooming in from broad structural categories to annotation labels and to actual examples. Table 6.4 reports the relative frequencies of the 10 types of internal structures of sequences extracted from the interview data.

Table 6.4 Relative frequency (ptw) of sequence structures in each subcorpus

	English			French		
	ftf	**radio**	**total**	**ftf**	**radio**	**total**
Simple (one)	68.60	71.24	69.50	62.79	50.02	58.73
Simple (multiple)	44.80	24.39	37.87	36.47	24.48	32.65
Compound (one)	4.28	9.57	6.08	5.43	6.30	5.71
Compound (one) + within	4.05	2.62	3.56	5.27	5.47	5.33
Compound (one) + periph.	3.17	5.02	3.79	3.21	4.28	3.55
Compound (one) + both	2.70	1.94	2.44	2.38	2.73	2.49
Compound (mult.) + both	1.88	1.60	1.78	2.22	1.90	2.12
Compound (mult.) + within	1.06	0.57	0.89	1.27	1.19	1.25
Compound (mult.) + periph.	1.17	0.68	1.01	0.89	1.54	1.10
Compound (mult.)	0.76	0.68	0.74	0.89	1.66	1.13

The frequency differences reported here take up the same language and register effects observed in the previous sections and will therefore not be commented upon any further. Moreover, the ranking is fairly stable across languages and registers and shows only minor differences for the rare values. The overall frequency-based ranking of sequence structures is therefore the following: simple fluencemes (one,

then multiple), one compound fluenceme (alone, then clustered with simple fluencemes) and multiple compound fluencemes clustered with simple ones. The occurrence of multiple compound fluencemes without simple fluencemes is very rare (the least frequent category), which points to the signaling role of simple fluencemes in structurally dense sequences.

This table allows us to establish a convincing association between increasing complexity and decreasing frequency. There is a steady decrease in frequency from unique simple fluencemes to sequences with more numerous and more complex fluencemes. Differences in frequency cease to be significant amongst very complex sequences. In other words, this table seems to confirm a link between type (simple vs. compound) and number (single vs. multiple) of fluencemes on the one hand and frequency on the other, which provides some evidence in favor of the fluency-as-frequency view. This result is in line with Candéa (2000: 442), who also found a negative correlation between frequency and "*degré de rupture*" (degree of interruption).

Zooming in on the rarest structures, it appears, however, that complex sequences do not necessarily correspond to what might intuitively be considered major disruptions in the utterance. The last category in the table covers a small variety of clusters (49 occurrences) which reflect the recurrent attraction of some compound fluencemes in the typology, in particular modified repetitions with completed truncations (RM + TR, 8/49 cases), propositional substitutions (RM + SP, 6/49) or morphosyntactic substitutions (RM + SM, 4/49), combined in rather short and non-disruptive contexts as in Examples (3) and (4).

(3) oui oui *ils sa- ils savaient* pas utiliser un ordinateur
yes yes they di- they did not *know how to use a computer* (FR-intf-06)

(4) they want stories of humanity where you see people on stage *in all their with all their* flaws and contradictions (EN-intr-08)

The truncation of "savaient" in (3) and the substitution of "in" by "with" in (4) involve partial repetition of anteposed ("ils") or postposed material ("all their"). Repeating available linguistic material has been experimentally shown to be generally associated with positive fluency strategies and high-skill speakers (e.g. Ejzenberg 2000; Götz 2013). Moreover, the interruption or stagnation of the ongoing utterance lasts two and three syllables, respectively, which indicates a small disruption – if any – on the perception of the unfolding utterance, although experimental research would be necessary to confirm this. It seems that the rarest sequences in the data are not necessarily the most disfluent, in comparison with mixed sequences containing both compound and simple fluencemes, as in the following examples:

(5) il raconte une histoire *euh (0.436) euh qui est la mienne euh qui est la mienne euh (0.444) euh disons (0.193)* entre ma naissance puisqu'il y a un poème sur la naissance
it tells a story uh which is mine uh which is mine uh uh let's say *between my birth because there is a poem on birth* (FR-intr-02)

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

In both of these examples, there is only one compound fluenceme, namely an identical repetition ("qui est la mienne", "have") which is clustered with a rather high number of simple fluencemes, mostly discourse markers, filled and unfilled pauses, in embedded and peripheral positions. These numerous signals add to the stagnating effect created by the repetition, either a repetition of several words in (5) or a word repeated several times in (6).

It appears from qualitative examination of such examples in the data that it is this combination which is a more robust indicator of major disruptions in the utterance, a suggestion which is corroborated by Candéa's (2000) experimental study where she found that the presence of a pause next to a disfluency (i.e. a hesitation marker, repetition or self-repair) increases the participants' perception of the disfluency. This analysis thus brings forward an important nuance to the fluency-as-frequency hypothesis: simple fluencemes, although most frequent as isolated sequences in the data, tend to occur in rather disfluent contexts as well once combined with one or several compound fluencemes, whereas compound fluencemes on their own (i.e. without simple fluencemes) do not appear to be particularly disruptive despite their very low frequency. The apparent association between frequency and structural complexity suggested by Table 6.4 with rather broad categories is undermined by zooming into specific annotation labels (e.g. RM + TR) and actual instantiations of sequences.

6.4.2 Frequency and sequence length

One way to possibly reconcile frequency with fluency is to add sequence length as a filter to the internal structure of sequences. Once again, we can vary the macro-labels by grouping sequences according to the main distinctions brought forward by Table 6.4, namely complexity and number of fluencemes. Figure 6.1 thus represents three coarse-grained groups of sequences, either simple (both isolated or clustered), compound (one compound fluenceme, potentially including additional simple fluencemes) and multiple compound (either clustered with simple fluencemes or not).

By contrast, in a more fine-grained perspective, Figure 6.2 reproduces each of the 10 levels from Table 6.4. In the two graphs, the colored areas correspond to the proportions of each type of sequence by sequence length measured in number of tagged words (for instance, 50% of 20-word sequences are taken up by compound fluencemes and another 50% by multiple compound fluencemes). The information from each graph will be compared in the following.

From the coarse-grained approach (Figure 6.1), we see that very short sequences represent the quasi-monopoly of simple fluencemes, which disappear after 10-word sequences. The small rise of simple-fluenceme sequences at this point (10-word sequences) is noteworthy and corresponds to three occurrences from the French interviews, as in Example (7).

(7) oui mais ça c'est la peur du débutant *mais bon ben il faut bon ben et puis alors s'*il y avait quelque chose qui n'allait pas euh
yes but that is beginner's fear but well you have to well and then so if *there were something wrong uh* (FR-intf-06)

The sequence in this example includes no fewer than six DMs ("mais", "bon ben" twice, "et puis", "alors", "si") and a false-start after "faut". Although containing only two different fluenceme types and seven fluenceme tokens, the sequence length in number of words is quite excessive for clusters of simple fluencemes and is partly due to the "complex" DMs (i.e. fixed unit made up of two components). The punctuating DMs ("bon ben" 'well') and the false-start might be interpreted as signals of trouble on the part of the speaker trying to order or select what to say next. This type of pattern is consistent for long sequences of simple fluencemes and tends to show the relative disfluency of such contexts. On the other hand, some long sequences of compound fluencemes can occur in fluent contexts, as in Example (8).

(8) est-ce qu'il y a *des régions où l'on parle le mieux le français et des régions où l'on parle moins bien*
are there regions where people speak French better and regions where people speak less well (FR-intf-02)

This sequence is 16-word long and only contains two fluenceme types and tokens, namely a modified repetition ("des régions où l'on parle") and a propositional substitution ("le mieux" by "moins bien"). Apart from the types of fluencemes included, this sequence is quite similar to the one in Example (7) in terms of length. However, in terms of fluency, we no longer see an effect of stagnation or interruption, which is instead replaced by an elaborate interrogative structure built on a repetition for a contrastive construction opposing "*mieux*" 'better' to "*moins bien*" 'less well'. Here, the length of the sequence reflects a strategic recycling of already uttered material for a stylistic, discourse-functional effect which is positive for both the speaker (since it does not require additional processing costs) and the hearer.

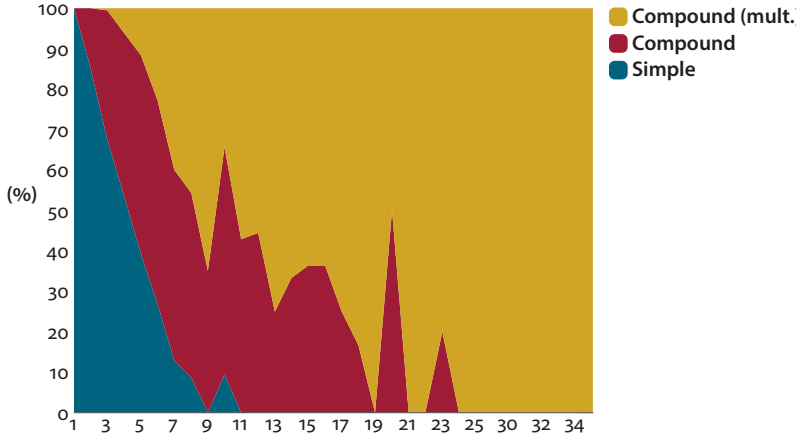


Figure 6.1 Proportions of sequence type (coarse-grained) by sequence length

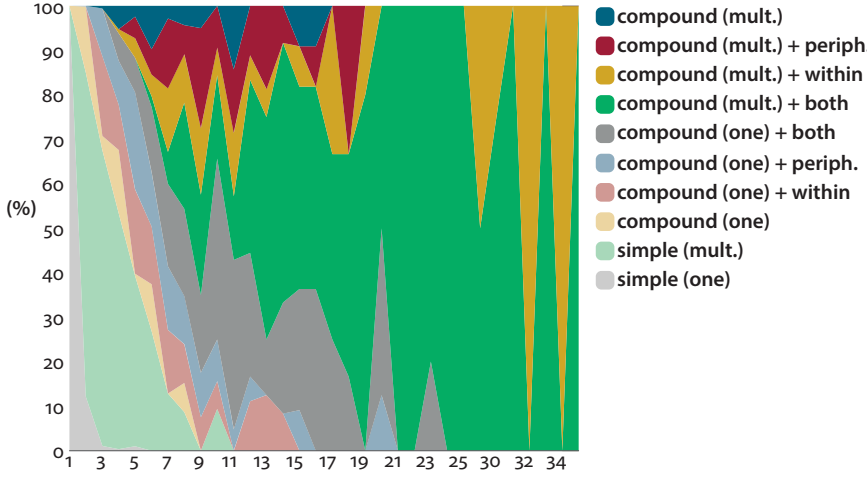


Figure 6.2 Proportions of sequence type (fine-grained) by sequence length

To sum up so far, length alone is not a reliable indicator of relative (dis)fluency, nor is sequence type alone. At the coarse-grained level of Figure 6.1, we see that long sequences of compound fluencemes are not necessarily problematic, as in Example (8) above, whereas long sequences of simple fluencemes tend towards disfluency, as in Example (7). In other words, this degree of granularity does not seem fine enough to map the variety of contexts each sequence type covers, which motivates the use of a more fine-grained analytical grid as provided by Figure 6.2,

where interesting patterns emerge. We see that the rarest sequences (multiple compound; multiple compound with peripheral fluencemes) are not the longest but rather range from short to medium size (cf. Examples (3) and (4) above), while very long sequences (around 30 words) correspond to occurrences of multiple compound with embedded and peripheral simple fluencemes (dark yellow and green), as in Example (9).

(9) *well I used to think that and I used to think that she should have had more courage and that she should have actually (0.433) gone on teaching or gone on doing* something with her mind (EN-intr-05)

In this example, three repetitions are intertwined ("I used to think that", "that she should have", "gone on"), sometimes with partially substituted material ("had" by "gone", "teaching" by "doing") and simple fluencemes such as DMs ("well", "and", "actually", "or") and an unfilled pause, amounting to 11 fluenceme tokens and 30 tagged words. Again, this extract does not appear particularly problematic since each repetition moves the discourse forward either by enumerating or alternating different contents expressed through the same formal structure. By contrast, Figure 6.2 shows a high proportion (50%) of long sequences (20 words) with only one compound fluenceme (and simple fluencemes in different positions) which are rather disruptive to the utterance linearity. These cases very often involve a parenthetical insertion, which signals a problem of message ordering, as in Example (10).

(10) <VAL_5> et qui reçoit cette revue (FR-intf-03)
<VAL_6> ah tout qui en fait la demande *et puis alors euh (0.480) euh on lui offre la revue ça paraît trimestriellement (0.580) on lui offre la revue* pendant un an
<VAL_5> and who receives this magazine
<VAL_6> ah everyone who asks and then uh (0.480) uh we offer them the magazine it is published every trimester (0.580) we offer them the magazine *for a year*

The speaker <VAL_6> is explaining how he runs his small journalistic business by providing different information (clients, frequency of publication, method of payment) in a certain order which he then finds inappropriate as attested by the repetition ("*on lui offre la revue*") and the insertion of "*ça paraît trimestriellement*". This corresponds to what Levelt (1983) terms an issue of linearization, that is when speakers edit the order of the contents they want to express so that they better fit the intended message. In this case, <VAL_6> feels the need to specify the frequency of publication before taking up the method of payment, which results in a repetition and several simple fluencemes. Examples such as (10) seem to indicate

that disfluency, or at least disruption of the flow, is more related to the presence of simple fluencemes and related phenomena (such as parenthetical insertions) in combination with compound fluencemes, rather than several compound fluencemes together.

Overall, the situation represented in Figure 6.2 reflects a complex interplay of factors, namely sequence length, fluenceme type and frequency. Medium-size sequences show the greatest variety of sequence types, with the special role of sequences containing one compound fluenceme with embedded and peripheral simple fluencemes (cf. Example (10)), while very short and very long sequences are more restricted in terms of frequency (very frequent and very rare, respectively) and fluenceme types. However, once confronted to actual instantiations from the corpus, the patterns do not necessarily map our expectations of (dis)fluency. Very long sequences are not the rarest in the data nor are they systematically disfluent; medium-size sequences are rather frequent and disfluent.

To conclude, the fluency-as-frequency hypothesis cannot be fully confirmed at this stage. What we can assert is that there seems to be an effect of length in a complex relation with frequency, which requires a more qualitative analysis of examples to make generalizations based on fine-grained observations. Another element missing from the present analysis is register variation and, in particular, the effect of planning (degree of preparation) available to the speakers. More conclusions could be drawn from comparing sequence patterns across different contexts which are cognitively more or less demanding, as opposed to the present interview data which only opposes broadcast and non-broadcast dialogues. In the next chapter, the same endeavor will be pursued with the integration of register variation as an additional clue to the (dis)fluency of sequences, focusing on clusters including at least one DM.

6.5 Summary

Paradigmatic annotation of fluencemes in the subcorpus of interviews established an overall fluenceme rate of 20% (in number of words tagged as fluencemes), which is considerably higher than what previous corpus studies reported – although it is largely explained by the intrinsic ambivalence of the fluencemes included in this study as well as by the prominent weight of unfilled pauses and DMs, which are usually excluded or highly restricted in most typologies. In the data, fluencemes associated to "covert repair" (Levelt 1983) or "forward-looking disfluencies" (Ginzburg et al. 2014) were found to be considerably more frequent than less ambivalent markers such as false-starts or interruptions. Major effects of variation include the higher frequency of unfilled pauses in English and of identical repetitions in radio interviews, which I connected to a potential "radio style".

Against my hypothesis, fluencemes appear more often isolated than clustered on the whole, which is again explained by the weight of pauses and DMs. The most frequent pattern consists of one word for one fluenceme token and type. Sequences of six tokens and more are very rare in the data, while the most extreme cases reach eight types, 15 tokens and 43 words in one sequence. Nevertheless, the hypothesis on clustering is confirmed for DMs, which appear more frequently in combination with other fluencemes than not.

The more complex the internal structure of a sequence, the less frequent it is in the corpus, yet close analysis of examples does not confirm that low frequency equates with disfluency. In fact, it is rather the combination of compound and simple fluencemes which is a reliable indicator of fluency, especially in medium-size sequences.

CHAPTER 7

The (dis)fluency of discourse markers

In this chapter, the patterns identified in Chapters 5 and 6 will be merged and refined by the integration of DM features with sequence types. This chapter thus complements Chapter 6 by extending the data to all subcorpora in *DisFrEn* and by integrating DM-level variables with sequence-level variables, focusing in particular on the overarching goal to rank different functional uses of DMs on a register-sensitive scale of (dis)fluency.

7.1 Sequence types across registers

This section will test the hypothesis according to which spontaneous discourse leads to more frequent and more varied fluencemes than planned speech, and intermediary registers are expected to be more similar to spontaneous dialogues than to formal monologues. Rates and types of sequences will be systematically compared across the eight settings in *DisFrEn*, combining metadata and frequency to further refine our understanding of the link between corpus frequency and fluency. Sequences which are specific to informal situations are hypothesized to be typically disfluent, while sequences shared across all registers are expected to be more ambivalent.

Exploratory investigation of any crosslinguistic difference in this respect will be carried out without any specific hypothesis. This section follows the same approach as the previous ones, attempting to test cognitive hypotheses with a combination of quantitative statistical analyses and qualitative functional interpretation of examples, at different levels of abstraction, here focusing on DM-based sequences.

In *DisFrEn*, 7,244 sequences containing at least one DM have been annotated across all registers and languages. Table 7.1 reports their relative distribution in each subcorpus. We see that DM-based sequences are more frequent in French, especially in conversations and phone calls where the gap with English is very large (these are the only significant crosslinguistic differences in this table). Apart from face-to-face interviews, where DM-based sequences are the most frequent in English (as opposed to conversations in French), the ranking of registers is the same in the two languages, following that of DMs discussed in Chapter 5.

Table 7.1 Relative frequency of DM-based sequences ptw in *DisFrEn*

	English	French	Total
conversation	46.11	66.83	56.46
phone	52.73	62.66	56.81
interview	53.30	55.87	54.62
radio	47.87	42.78	45.38
classroom	43.93	39.48	42.67
sports	37.76	36.47	37.20
political	21.04	18.79	19.97
news	13.62	16.35	14.96
Total	42.26	47.71	44.80

Looking at register variation, there is a sharp decrease in frequency of sequences in political speech and news broadcast, while the other registers are not so neatly contrasted, especially in English with rates averaging 47 sequences ptw in the remaining six registers. French, however, is more affected by register variation with two subcorpora above 60 sequences ptw, which reflects the general distribution of DMs. Overall, DM-based sequences do appear more frequent in spontaneous and intermediary registers than in the formal settings of news and political speech, as expected.

To identify the specific clusters of DMs and fluencemes behind this frequency table, the following analysis investigates fluenceme sequences according to three macro-labels, ranked by increasing order of granularity:

- cluster (3 types) which specifies, for each annotated DM, whether it occurs alone, with other DM(s) or in a cluster with other fluencemes;
- sequence category (6 types), which is a hierarchical DM-based system;
- internal structure (10 types), which was the basis of the analysis in Section 6.4.

A number of multivariate models have been computed at each of these degrees of granularity. I will report the main findings of these analyses, using either register or situational features as metadata factors depending on the research question and hypothesis.

7.1.1 "Cluster"

Starting with the most coarse-grained degree of abstraction, we find that about 60% of all 8,743 DMs in *DisFrEn* are clustered with other fluencemes (excluding co-occurrence with DMs only), a proportion which is higher in political speeches (86%). Figure 7.1 reports on a conditional inference tree (a type of decision tree

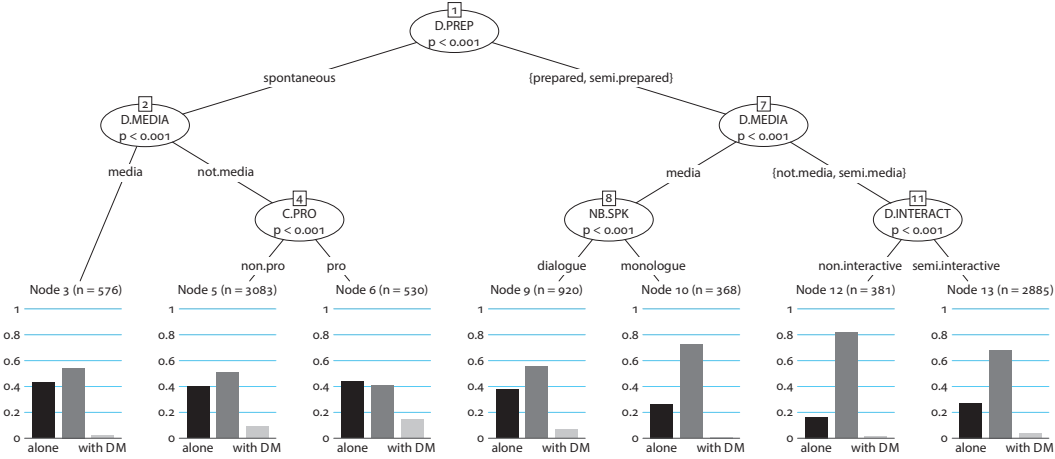


Figure 7.1 Conditional inference tree for isolated, clustered and co-occurring DMs

based on significance tests) with situational features as input factors instead of register labels. Each column in the barplots corresponds to one of the three levels (alone in black, clustered in dark grey, co-occurring with DMs in light grey, respectively) and each node of the tree corresponds to a significant divide between the situational features. The abreviations in the graph are: "D.PREP" for degree of preparation, "D.MEDIA" for degree of broadcasting, "C.PRO" for (non-)professional category, "NB.SPK" for number of speakers and "D.INTERACT" for degree of interactivity.

It appears that, as expected, the degree of preparation is the most influential variable (on top of the tree), with spontaneous contexts on the one hand and (semi-)prepared contexts on the other. In the latter, clustered DMs are always much more frequent than isolated or co-occurring DMs, while the difference between clustered and isolated uses is much smaller in spontaneous discourse and even reversed (slightly more isolated than clustered DMs) in non-broadcast professional settings, which only corresponds to the French subcorpus of phone calls.

While already pointing to some attraction between factors, this level of analysis is not informative enough since it does not provide the specific type of fluencemes with which DMs cluster and does not allow us to identify register-specific patterns.

7.1.2 "Sequence category"

Turning to the second degree of granularity, the clusters can be distinguished according to the fluencemes they contain. Three categories exclusively include simple fluencemes, namely DMs alone (type "D"), DMs and pauses (type "P"), DMs, pauses and interruptions (type "F"). Two types correspond to compound fluencemes, either repetitions (type "R") or a combination of repetitions and substitutions (type "S"), which can also include the contents of "D" or "P". Finally, the mixed type "Z" includes both interruptions ("F") and compound fluencemes ("R" and/or "S"). Figure 7.2 reports on their association to each register through a conditional inference tree.

The first divide reveals two major groups of registers. Firstly, conversations, phone calls and radio interviews share a similar preference for sequences containing exclusively DMs ("D"). Secondly, the other five registers correspond to intermediary and formal contexts favoring "P" sequences (DMs and pauses) although to various extents: almost exclusively in political speeches (cf. the 86% of clustered DMs mentioned above), almost no difference with "D" in sports, a steady gap in interviews, news and classroom lessons. In this respect, our hypothesis that intermediary registers such as interviews or classroom lessons would behave more like informal contexts as far as (dis)fluency is concerned is not confirmed at this level of analysis.

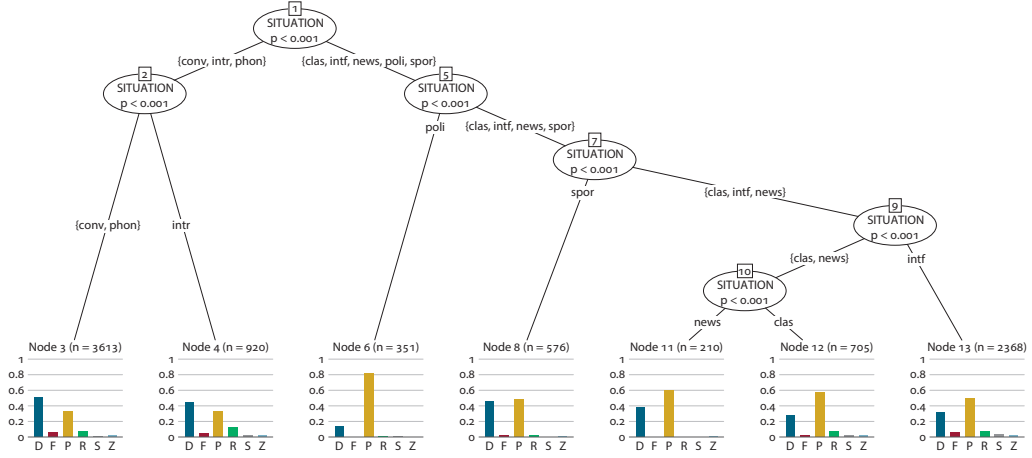


Figure 7.2 Conditional inference tree for sequence category by register

Lastly, we see in this figure that, apart from D- and P-sequences, the other types are very rare, especially in political, sports and news discourse. Rare sequences seem to be responsible for the categorization of radio interviews under the same branch as conversations and phone calls, which all share a substantial proportion of "R" (repetitions) and "F" (false-starts and truncations). This result provides a first confirmation of the diversity hypothesis (more different types of sequences in informal registers), although we see that D- and P-sequences are overwhelmingly frequent across all registers.

These restrictions of sequence categories by register are represented in an extended association plot in Figure 7.3, where differences in proportions are shaded according to the statistical significance of Pearson residuals. A number of observations are confirmed by this graph. Firstly, D-sequences (DMs only) are attracted to the informal settings of conversations and phone calls (blue boxes, more observed than expected), whereas classroom lessons, face-to-face interviews and political speeches seem negatively associated to them (red boxes, fewer observed than expected). The attraction of isolated DMs ("D") to interactive contexts converges with the previous observation of turn-initial and turn-final DMs, which also favor these settings (cf. Figure 5.4). It could well be that many isolated DMs occur in these specific slots in the turns which are naturally less prone to co-occur with pauses.

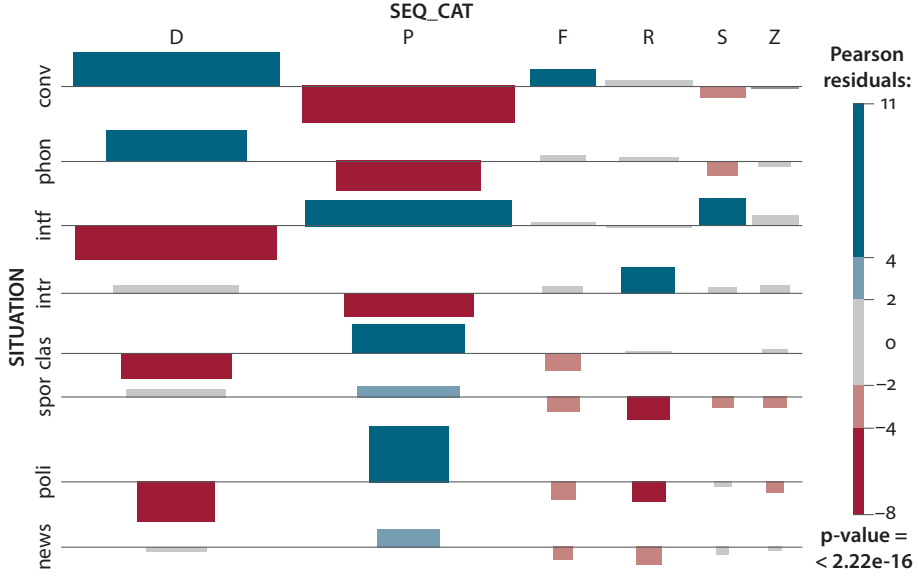


Figure 7.3 Extended association plot of sequence categories by register

By contrast, clusters of DMs and pauses ("P") show the exact opposite pattern, with a strong attraction to classroom lessons, face-to-face interviews and political speech (also, to a lesser extent, news and sports; light blue boxes) and a significant absence from conversations and phone calls (as well as radio interviews, albeit less significantly).

Turning to less frequent categories, we see interesting associations between sequence types and registers, such as the significant frequency of F-sequences (false-starts and truncations) in conversations, R-sequences (repetitions) in radio interviews (previously identified as a potential "radio style", cf. Section 6.2.3) or S-sequences (substitutions) in face-to-face interviews. These three categories of sequences are rare, structurally more complex and potentially more disruptive (less ambivalent) than DMs and pauses. Their significant attraction to informal and intermediary registers (and absence from highly prepared and formal contexts) provides some evidence of both the fluency-as-frequency hypothesis and the hypothesis regarding the (dis)fluency of register-specific sequences.

Nevertheless, closer qualitative interpretation of authentic examples only partially confirms this conclusion. The restriction of F-sequences to conversations and their resulting degree of disfluency seem to fit the data well enough, with most examples attesting the disruptiveness of the false-start and truncation fluencemes, as in (1).

(1) 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)

This sequence includes both a truncation ("wo-") and a false-start at the second "I" after which the speaker restarts with a DM "after". The conceptual definition of false-starts and truncations, their interruption of the ongoing structure, their low frequency and restriction to informal conversations all converge in pointing to a rather disfluent category of fluencemes. However, this is not necessarily the case: according to the fluency-as-frequency hypothesis, the more frequently a particular pattern occurs in a corpus, the more accessible it becomes. A corollary to this hypothesis is that non-typical sequences can be expected to be more disruptive in registers where they are rare than in settings where they are more frequent and therefore less marked. For instance, we saw in the extended association plot (Figure 7.3) that S-sequences are strongly associated to interviews, negatively associated to phone calls and neutral with respect to political speeches, three patterns which are illustrated with the examples below.

(2) and you know *is it going to look like dad is it going to look like mum* (EN-intf-03)

(3) *well I wasn't driving well I was driving* partly on the road but also on through open country (EN-phon-02)

(4) il n'y a pas de dialogue social sans respect de l'autre (0.814) mais il n'y a pas de vrai dialogue social (0.400) sans (0.158) culture de la responsabilité
there is no social dialogue without respect for each other (0.814) but there is no true social dialogue (0.400) without (0.158) a culture of responsibility
(FR-poli-01)

In (2), the propositional substitution ("dad" by "mum") illustrates a strategic use of this fluenceme for an enumeration, which is typical of face-to-face interviews where S-sequences tend to frequently occur. By contrast, the example in (3) comes from the phone calls subcorpus with which S-sequences are negatively associated, and we see that the speaker is correcting himself, thus confirming that relatively rare sequence types are potentially more marked and disfluent. In the political speech of (4), the whole extract constitutes the sequence (modified repetition with propositional substitution of "respect de l'autre" by "culture de la responsabilité") in a stylistic effect of emphatic enumeration. The strategic use of a S-sequence in (4) is compatible with the association of this sequence type to political speeches, where they are not significantly more or less frequent than in other registers. These examples tend to show that it is not only the low raw frequency of a particular structure but mostly its low frequency relative to other registers, and its restriction to (or absence from) some settings which might be a better indicator of its relative fluency (here, relatively more disfluent in phone calls).

7.1.3 "Internal structure"

The (dis)fluency of register-specific sequences might also be an effect of the relative diversity vs. restriction of different registers in terms of different sequence types, which were hypothesized to be more varied in spontaneous than planned speech. When considering the 10 types of internal structure, news broadcasts appear to be the most restricted setting with occurrences in only five structural possibilities and with very anecdotal frequencies in the patterns of compound fluencemes. As a result, the same structure occurring in news broadcasts and in another register which is less restricted in sequence types should show a difference in markedness, especially if this structure is significantly more frequent in the second register. This is the case for the following two Z-sequences (i.e. interruptions mixed with repetitions and/or substitutions) which instantiate the clustering of multiple compound fluencemes with peripheral and embedded simple ones:

(5) cent trente pigeons (0.310) sont aujourd'hui guéris *et on commen- on a commencé (0.560) ils ont commencé* à être relâchés ils sont tout à fait sains
a hundred and thirty pigeons (0.310) are now healed and we star- we have started (0.560) they have started *to be released they are perfectly healthy* (FR-news-07)

(6) *les filles je couraient quand même un peu après les garçons ou les garçons couraient après les filles (0.970) bon* la toute première fois que j'ai vu mon mari
the girls I ran a little after the boys or the boys ran after the girls (0.970) well *the very first time I saw my husband* (FR-intf-04)

In the context of a news broadcast, the truncation and substitutions in (5) are highly unusual and detrimental to the expected standards of journalistic speech. In interviews, however, recycling strategies including additional simple fluencemes such as false-starts ("*je*") or DMs ("*ou*", "*bon*") are much more frequent and might not appear to be strongly marked or disruptive, although perceptive ratings would be necessary to assert such a conclusion. Overall, such mixed sequences can be expected to be particularly marked in broadcast formal registers, which are restricted in their diversity of sequence types and where mixed sequences are significantly less frequent than in other registers. In fact, the restriction of rare sequences to informal and diverse registers provides additional evidence for their relative disfluency, as opposed to sequences of DMs and pauses which are highly frequent across all registers, thus attesting to their functional ambivalence on the fluency-disfluency scale.

Coming back to the hypothesized special place of settings with an intermediary degree of preparation, an analysis of sequence complexity can provide additional evidence signaling an enhanced attention of speakers towards their speech, thus leading to more disfluent discourse (Broen & Siegel 1972). Most statistical modelling techniques such as classification trees or random forests fail to account for rare sequences beyond the great majority of types D and P. As a result, the comparison between potentially fluent sequences (D, P) and potentially disfluent ones (the other four) cannot be modeled beyond the information already provided by the extended association plot in Figure 7.3. Similarly, at a more fine-grained level of analysis such as the 10 types of internal structure, only the most frequent clusters are included in the models (i.e. one simple fluenceme and multiple simple fluencemes), which is why I merged several structure types in two groups based on the results of Section 6.4. Mixed sequences (involving both simple and compound fluencemes) are compared to single-type sequences (only simple or only compound fluencemes), based on the previous finding that it is the combination of both types which is linked to disruptive and disfluent uses.

A binomial logistic regression was computed on this data, with situational features as input factors. The effect of intermediary levels is confirmed with a

significant increase of mixed sequences in semi-prepared compared to spontaneous settings (the model selection process did not include language or degree of interactivity in the final model). As already suggested by Broen & Siegel (1972) and Halliday (1987), hesitations and disfluencies, here operationalized in the form of mixed sequences of fluencemes, thus tend to occur more frequently in intermediary registers with a heightened attention towards one's speech than in informal dialogues where speakers do not monitor their speech too closely, and than in planned discourse where the cognitive demands on the speaker are lower. To sum up, the high frequency and significant attraction of mixed sequences (both simple and compound fluencemes combined) to intermediary registers could be interpreted as a sign of the relative disfluency of these settings, in line with cognitive hypotheses in the literature and with results from the paradigmatic annotations (Section 6.4) of this research.

7.1.4 Sequence-specific DMs

Finally, we can replicate the analysis of diversity of sequence type by DM expressions in order to try and identify tokens which are typically fluent (i.e. specific to sequence types associated with formal registers) or typically disfluent (i.e. specific to mixed patterns identified as potentially disruptive). I will comment on a selection of DMs, excluding *hapax legomena*. *When* (129 occ. in *DisFrEn*) shows no occurrence in Z-sequences and very rarely occurs in S- and F-sequences, which would argue for its fluency (speakers produce *when* in otherwise non-problematic contexts). *Then* (94 occ.) is restricted to D- and P-sequences except for two occurrences in F-sequences (i.e. with false-starts and/or truncations), which is a sign of its fluent segmentation role. *For example* (16 occ.), *for* (6 occ.), *meanwhile* (6 occ.), *yet* (4 occ.) or French *tandis que* 'while' (10 occ.) only occur in sequences of simple fluencemes (isolated or clustered), which reflects their discourse-structuring role, typically connecting two segments in a specification, causal, temporal, concessive or contrastive relation, respectively. French *disons* 'let's say' (9 occ.) occurs mostly in R-, Z- or F-sequences (only one in P), which could reflect its semantics of encoding lexical access trouble and point to a relative disfluency.

These particular configurations were manually and qualitatively identified, so that the conclusions may not be generalizable. Nonetheless, there seems to be a coherent link between the semantics of some DMs and their restriction in sequence types. This line of investigation will be further pursued with the integration of functional annotations in Sections 7.2 and 7.3 and of qualitative repair categories in Section 8.4.

To conclude this section, I have identified some patterns of co-variation between sequence types and registers which point to a divide between formal registers on the one hand, where sequences are rare and mostly restricted to simple fluencemes, and informal and intermediary registers on the other showing a greater diversity of sequences. The fluency-as-frequency hypothesis has been refined with register variation, especially showing the difference in markedness between uses of the same sequence type across registers where it is more or less typical.

7.2 Sequence types across DM features

7.2.1 Disfluency and functional domain

In this section, I will try and test whether the association of functional domain by sequence type can be a clue to the fluency of the DMs expressing this domain. In particular, I expect sequential DMs to be highly attracted to pauses given their discourse-structuring and segmentation role. The extended association plot in Figure 7.4 reports on the mapping between sequence type and functional domains. Only single-tagged DMs are included in this analysis (N = 8,393).

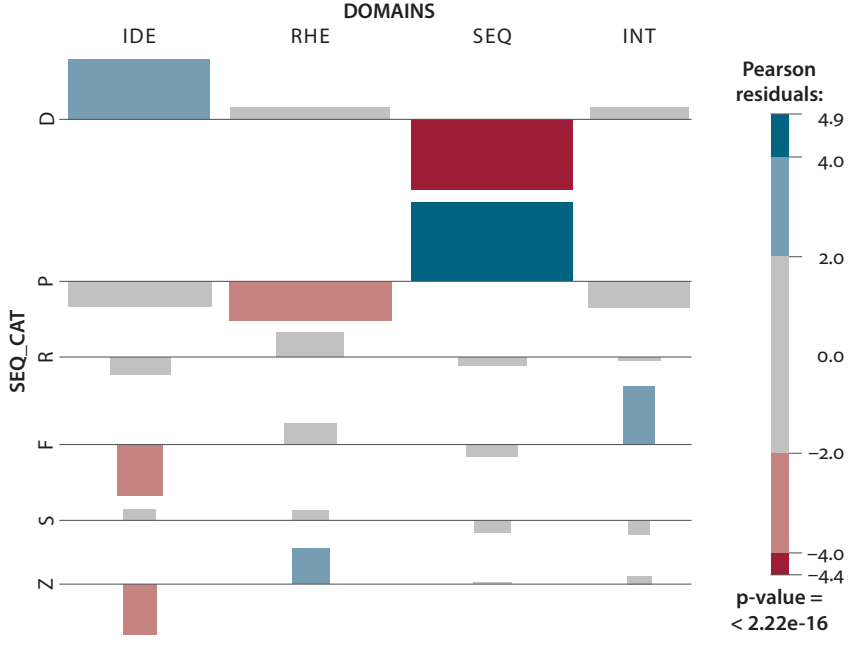


Figure 7.4 Extended association plot of functional domains by sequence type

We see that each domain has one favorite sequence type and one dispreferred category (or two at the most). The hypothesis for sequential ("SEQ") DMs is confirmed with many more observed than expected clusters with pauses (49% vs. 40% in the other domains) and, conversely, significantly fewer isolated uses than in the other domains, which corroborates their high-level structuring role. Given the ambivalence and pervasiveness of pauses, including in formal registers, this strong association can be seen as a sign of fluency connected to the information packaging and planning role of sequential functions such as *addition*, *topic-shift* or *turn-taking*. In the same line of reasoning, the frequent occurrence of sequential DMs at the onset of turns (19% of sequential DMs are turn-initial, vs. 5% on average for the other three domains) supports this generalized fluent interpretation of sequential DMs as operators of the discourse organization across topics, turns and utterances.

This graph also suggests a relatively high degree of fluency for ideational ("IDE") DMs, which are quite frequently well-integrated in the utterance (isolated, D-sequences) and negatively drawn to sequences of interruptions (F) and mixed fluencemes (Z). By contrast, the rhetorical ("RHE") and interpersonal ("INT") domains are each associated to one of these typically disfluent types of sequences, namely Z-sequences (interruptions with repetitions and/or substitutions) for rhetorical and F-sequences (false-starts or truncations) for interpersonal DMs. The examples below illustrate the most typical – although not necessarily most frequent – pattern for each of the four domains.

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

(8) 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)

(9) 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)

(10) 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)

The ideational *concession* in (7) is inter-sentential (i.e. coordinating), yet the connected utterances are not separated other than by the DM "but" (D-sequence). The rhetorical *reformulation* in (8) occurs in a rather fragmented segment where the DM "I mean" starts over after a false-start on "really" and leads to a repetition of "I never had a career" with a long embedded parenthetical insertion, a pause and a partial repetition with modification including a truncation ("I n- didn't even have

a career") (Z-sequence). The *topic-shift* in (9) is prosodically independent (unfilled pauses at both sides) and marks a major discourse boundary between two points of the academic lecture. Lastly in (10), the speaker invites the hearer to follow her reasoning after a false-start, thus creating common ground and maintaining or *monitoring* the communicative success of the exchange. This use of interpersonal DMs in the context of interruptions relates to the *ellipsis* function (also belonging to the interpersonal domain) typically expressed by DMs such as *and so on*, whereby the speaker assumes that the hearer can infer the rest of the enumeration or, as in (10), the rest of the interrupted utterance, thus relying on the participant's cooperation to compensate their own incompletion – whether this incompletion is voluntary or not.

To sum up so far on the associations of form and function and their proposed interpretation as more or less fluent, we can suggest the following scale by decreasing order of fluency (Figure 7.5). Sequential DMs occupy the fluent end of the scale with their attraction to clusters with pauses. Ideational DMs are also ranked as fairly fluent on the basis of their dissociation from typically disfluent sequences such as F- (interruptions) or Z-sequences (mixed).

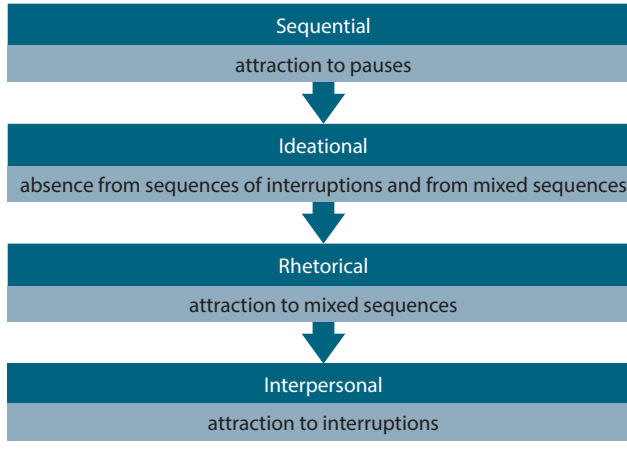


Figure 7.5 DM domains on the scale of (dis)fluency

Lower on the scale, rhetorical DMs (i.e. DMs with a rhetorical function) tend towards disfluency, as attested by their attraction to Z-sequences. Lastly, at the bottom of the scale, interpersonal DMs seem to be attracted to interruptions, thus suggesting a relatively low degree of fluency. This scale is obviously relative and generalized, albeit based on the combination of objective cues (corpus frequency, formal features) and qualitative interpretation of examples.

7.2.2 Disfluency, domain and position

We can refine this cognitive-functional scale of (dis)fluency by taking syntactic position into account. Each major slot in the utterance can be related to expectations of (dis)fluency. Initial position should be the preferential slot for clusters of sequential DMs and pauses given their segmenting function and inter-sentential scope. Medial position should be linked to rather disruptive sequences interrupting the unfolding of the utterance or modifying its contents and/or illocutionary force (cf. Section 5.2.4.3).[31] Final position can be expected to attract interruptions and signals of trouble detection since, according to Levelt (1983), speakers' attention towards their own speech is enhanced towards the end of utterances. As a reminder from the previous chapter, final position was found to be strongly associated with interpersonal functions, initial (pre-field) position with sequential functions and medial (middle-field) position with rhetorical functions. By integrating domains, positions and sequence types, we can therefore confirm or not the previous interpretations of (dis)fluency.

For readability purposes, Table 7.2 only shows the mapping of sequence types by micro-syntactic positions, focusing on the three most frequent slots. The distribution of domains within each sequence type will also be included in the discussion of these results. Starting with the initial position, we see that it is consistently the most frequent slot to feature fluencemes across sequence types, which is explained by the initiality of DMs on the whole. Turning to less typical positions, it appears that P-sequences show the lowest proportion of medial DMs and are the only sequence type below the cross-type average of 5.88%. All other sequence types show a very similar proportion of medial DMs, which can therefore not be used as a relative indicator of disfluency.

Table 7.2 Proportions of micro-syntactic positions by sequence type

		initial	medial	final	Total
Pauses	(P)	86.62%	3.58%	9.80%	3521
DMs	(D)	78.86%	7.77%	13.37%	3372
Repetitions	(R)	84.79%	6.80%	8.41%	618
Interruptions	(F)	83.19%	6.55%	10.26%	351
Mixed	(Z)	81.22%	8.29%	10.50%	181
Substitutions	(S)	85.63%	8.62%	5.75%	174
Total %		83.01%	5.88%	11.11%	100%
Total occ.		6821	483	913	8217

31. Unlike in written English and French, where medial position is a typical feature (Altenberg 2006; Dupont 2015), it is very rare in spoken language (5.75% in *DisFrEn*), hence the assumption of the intrusiveness of medial DMs.

Within each sequence type, the interpersonal and rhetorical domains always take up the highest proportions of medial positions. Interpersonal DMs are most frequent in the medial position of sequences with repetitions (R), as in Example (11), while rhetorical DMs are mostly medial in sequences with DMs only (D), as in (12).

(11) it's just *you know* the the the qualities that spring to mind (EN-conv-03)

(12) there was this rock in the path (0.527) and uhm (0.740) and I *sort of* assumed I could go over it (EN-phon-02)

Example (11) illustrates the recurrent use of interpersonal *you know* for planning or stalling purposes when it is combined with repetitions (often longer than a single reiteration as here). The pattern of isolated and medial rhetorical DMs is very often represented by occurrences of *kind of* or *sort of* as in (12). Therefore, so far, our expectations for the initial and medial positions are confirmed.

Regarding the final position, two thirds of the interpersonal DMs are clause-final in sequences with DMs only and pauses. However, interpersonal DMs occur equally in initial and final position in more disfluent sequences (up to 50%), especially in sequences with interruptions as in (13) and (14).

(13) she said nothing on God's earth would make me (0.820) *you know* with her present job she's sort of uhm (0.650) having (0.810) high job expectations (EN-conv-08)

(14) il m'a frappée l'autre m'a bat- *hein* je m'étais disputée et je lui ai raconté
he struck me the other hi- hein *'right' I had an argument and I told him* (FR-intf-04)

In (13), "you know" is utterance-initial, following a false-start on "me" and an unfilled pause. In (14), "*hein*" ('right') is utterance-final and follows a truncation ("bat-" for "*battue*").[32] In light of these parallel examples, it appears that the effect of interpersonal DMs in F-sequences does not fundamentally differ depending on the initial vs. final position. The hypothesis that final interpersonal DMs signal disfluency is thus not confirmed.

Overall, we can conclude from Table 7.2 that proportions of positions alone do not allow us to distinguish sequence types, apart from a binary divide opposing P-sequences to all others combined, based on the lower proportion of medial DMs in P-sequences. In addition, it is not possible to confirm the potential disfluency of interpersonal and rhetorical DMs through a mapping of sequence type and position. In F-sequences (interruptions), which were identified as the least ambivalent

32. The prosodic contour of the DM is often necessary to distinguish final from initial position of these DMs which are more flexible and not entitled to a specific syntactic position, as opposed to conjunctions, for instance.

(most disfluent) of all sequence types, the two potentially disfluent domains take up smaller proportions of medial occurrences than in all other sequence types. In other words, these three potential sources of disfluency (F-sequences, medial position, rhetorical or interpersonal domain) do not converge. Similarly, the typical final position of interpersonal DMs is the least frequently represented in disfluent sequences of interruptions and most frequent in the unmarked sequences of pauses and DMs only, which tends to disprove the association of disfluency to clause-final interpersonal DMs.

In sum, these results indicate a complex interplay of three factors at a rather coarse-grained level of analysis, which shows that the proposed two-way scale of (dis)fluency represented in Figures 7.4 and 7.5 does not hold against the inclusion of an additional variable, let alone against qualitative analysis of a variety of examples. The only statistically valid associations which can be drawn from these three variables are represented in the multiple correspondence analysis (MCA) graph in Figure 7.6.

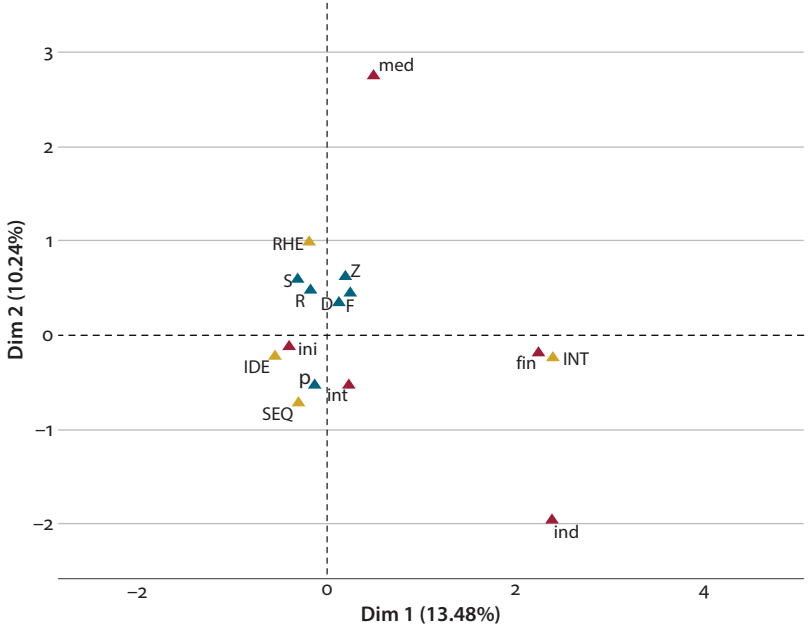


Figure 7.6 Multiple correspondence analysis of domains, position and sequence type

We see that no strong three-way association can be found. The ideational domain is attracted to the initial position. The sequential domain confirms its attraction to clusters with pauses ("P"). The rhetorical domain seems to frequently co-vary with sequences of substitutions ("S") and repetitions ("R"), while the interpersonal

domain is very close to the final ("fin") position. Another interesting grouping of variables is located in the top-right quadrant, where we see that the medial ("med") position shows some connection with sequences of isolated DMs ("D"), interruptions ("F") and mixed fluencemes ("Z"). This result tends to confirm our hypothesis regarding the disruptiveness of DMs occurring utterance-medially, although it is not associated to any particular domain. In other words, potentially disfluent sequences cannot be reliably distinguished according to the function and position of the DM.

7.2.3 Synthesis of variables

A more encompassing view of factors impacting the distribution of DMs in more or less fluent contexts is provided by the multiple logistic regression computed for each domain and including as independent variables not only sequence type and position but also language, register and sequence length (internal structure of sequences was removed from the final model after stepwise selection). It returns the significant effects summarized in Table 7.3.

Table 7.3 Significant effects for the multiple logistic regressions by domain

Domain	Increase in likelihood	Decrease in likelihood
Ideational	initial and medial position; news, political and sports registers	P, F and Z-sequences; French; radio interviews; longer sequence length
Rhetorical	Z-sequences; initial and medial position	independent position; all registers except interviews
Sequential	P-sequences; independent and initial position; French; phone calls and sports; longer sequence length	medial position
Interpersonal	P, R and F-sequences; conversations and phone calls	initial, medial and independent positions; political and sports registers

Overall, not all variables are relevant to all domains, although register preferences, position and sequence types consistently appear either as positive or negative (or both) effects for each of the four functional categories. By contrast, language is not always involved to explain differences in the data, which is consistent with the results discussed so far in this book, where few major crosslinguistic effects have been found.

These regression models help us understand the restrictions and favorable conditions that trigger the production of one type of DM over the others. Nevertheless, to draw conclusions on the relative (dis)fluency of domains based

on these significant associations of variables would over-generalize and overlook the high variation within domains and within sequence types, as illustrated many times in this chapter and the previous one. For instance, we have seen numerous examples of substitutions and long sequences which did not meet the expectation of disfluency but instead were very elaborate structures of enumeration or parallelisms.

In sum, it is not certain whether strong associations between variables can be systematically and directly interpreted in terms of fluency or disfluency, even when multiple sources of evidence converge (e.g. a medium-size Z-sequence of simple and compound fluencemes with a medial rhetorical DM in a spontaneous conversation), firstly because of the intrinsic ambivalence and variation within each variable (e.g. medium-size sequences are not always disfluent) and secondly because, according to the fluency-as-frequency hypothesis, the high frequency of these potentially disfluent sequences vouches for their high cognitive accessibility and entrenchment, which would mitigate their negative effects on production and perception.

Even with careful example-based analysis of each possible combination of variables, the precise evaluation of the fluency and disfluency of DMs and sequences in the corpus would remain invariably speculative without external perceptive validation. The findings from this section therefore need to remain general and prospective, suggesting coarse-grained – yet statistically significant – trends and opening up avenues for further investigation. Replicating the statistical analyses in this section to more fine-grained variables, such as functions (instead of domains) and the internal structure of sequences, as carried out in the next section, can reduce the variation within these relevant variables of the fluency scale.

7.3 Potentially Disfluent Functions

In Chapter 3, I posited the existence of a set of "Potentially Disfluent Functions" or PDFs which are conceptually related to fluency and disfluency, namely *reformulation, punctuation* and *monitoring*. *Reformulation* covers both paraphrases for clarification or other purposes and corrective reformulations (related to substitutions in terms of fluencemes). 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. PDFs are expected to frequently occur in rather disfluent sequences, that is patterns identified in the previous sections as associated to disruptive, non-ambivalent contexts of use.

7.3.1 PDFs across registers

PDFs (restricted to single tags) take up 1,250 DMs in total in *DisFrEn*, that is 14.3% of the data. *Monitoring* and *punctuation* are particularly frequent as they appear among the 10 most frequent functions overall (only in French for *punctuation*). This general observation of high frequency is a potential sign of the greater functional ambivalence of these two PDFs compared to *reformulation*. In line with the approach taken in this study, register variation is considered as a first approximate indicator of (dis)fluency insofar as frequent occurrences of a particular function or DM in formal registers vouch for its strategic or at least unmarked use, while restriction to informal spontaneous dialogues points to disfluency.

Table 7.4 reports the relative distribution of the three PDFs across registers and languages. We see that, overall, the frequencies of PDFs follow the general distribution of DMs across registers (cf. Section 5.1), from spontaneous dialogues to intermediary and formal settings. Their high frequency in conversations and phone calls is mainly due to the French data, where they are well represented (cf. 14 *monitoring* DMs ptw in French conversations). This major crosslinguistic gap in conversations corresponds to the large number of *quoi, hein* and *tu vois* which were previously identified as very frequent interpersonal DMs in French. Bearing this role of French *monitoring* DMs in mind, language variation will no longer be discussed here.

Table 7.4 Relative frequency (ptw) of PDFs per language and register

	Monitoring		**Punctuation**		**Reformulation**		**Total**
	EN	**FR**	**EN**	**FR**	**EN**	**FR**	
conversation	3.78	13.54	2.12	4.19	1.77	4.36	14.87
phone	4.31	8.40	2.26	6.19	2.46	3.10	12.58
interview	4.28	6.93	0.59	2.44	0.76	1.88	8.52
classroom	3.71	3.49	0.95	3.49	1.80	1.34	7.00
radio	2.17	3.56	0.57	1.54	1.03	0.95	4.89
sports	0.12	3.19	0.49	1.43	0.49	0.64	2.89
political	0	0.13	0	0.26	0.12	0	0.24
news	0	0	0	0.15	0.14	0	0.14
Total	**2.73**	**6.4**	**1.01**	**2.62**	**1.16**	**1.97**	**7.73**
Total occ.	**236**	**482**	**87**	**197**	**100**	**148**	**1250**

Regarding the hypothesis on the register variation of PDFs, this table allows us to confirm their quasi-absence from very formal broadcast settings (political speeches and news broadcasts) and, to a lesser extent, from other broadcast registers such as sports commentaries and radio interviews with the exception of *monitoring* DMs.

PDFs thus seem to favor more spontaneous and interactive settings of conversation, which is consistent with their "potentially disfluent" interpretation.

In order to evaluate whether the distribution of PDFs across registers is more restricted to informal dialogues than the other functions in the taxonomy, we can compare the proportions in which they occur in the different registers to those of non-PDFs, that is all the other functions combined. Figure 7.7 represents the extended association plot run on this data.

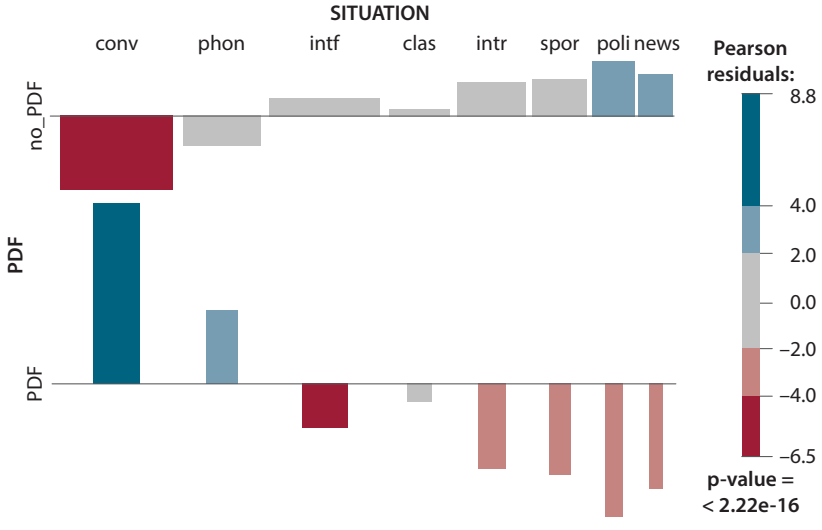


Figure 7.7 Extended association plot of PDFs and non-PDFs across registers

The analysis returns the following significant differences:

- PDFs are strongly and positively associated to conversations (42% PDFs vs. 26% non-PDFs) and, to a lesser extent, to phone calls (17% vs. 12%);
- PDFs are not significantly different from non-PDFs in classroom lessons (7% vs. 8%) and negatively associated to all the other registers (e.g. 0.32% vs. 4.63% in political).

In sum, the disfluency of PDFs seems to be confirmed at a general level by their distribution in registers and, in particular, the fact that they are more frequent in spontaneous dialogues (conversations and phone calls).

7.3.2 PDFs and sequence types

Another potential cue to the disfluency of PDFs can be found in their clustering tendency, following the results of previous works (e.g. Candéa 2000; Brennan & Schober 2001) showing that combinations of disfluencies are more reliable signals of hesitations than isolated occurrences. In the data, it appears that PDFs and non-PDFs show the same preferences and ranking, with a majority of clustered contexts, followed by isolated and co-occurring (with DMs only) cases. However, the proportion of these patterns is significantly different depending on whether the DM expresses a PDF or not: 66.72% of PDFs occur in a sequence with fluencemes against 57.69% for the other functions ($z = -6.006$, $p < 0.001$) and another 8.4% co-occur with DMs only (against 6.18% for the other functions, $z = -2.949$, $p < 0.01$), leaving a smaller proportion of isolated PDFs than all other functions combined. Closer investigation of the types of fluencemes in these clusters is necessary to draw reliable interpretations of (dis)fluency, but the significant differences identified at this level, added to the above-mentioned effects of register, so far point to a coherent classification of these PDFs as rather disfluent.

In the majority of their occurrences, PDFs combine with other fluencemes than DMs. Their "potential disfluency" leads us to hypothesize frequent clusters in sequence types previously identified as less ambivalent, i.e. absent from formal registers, structurally complex, longer and less frequent. This hypothesis is confirmed by the following extended association plot (Figure 7.8), where we see

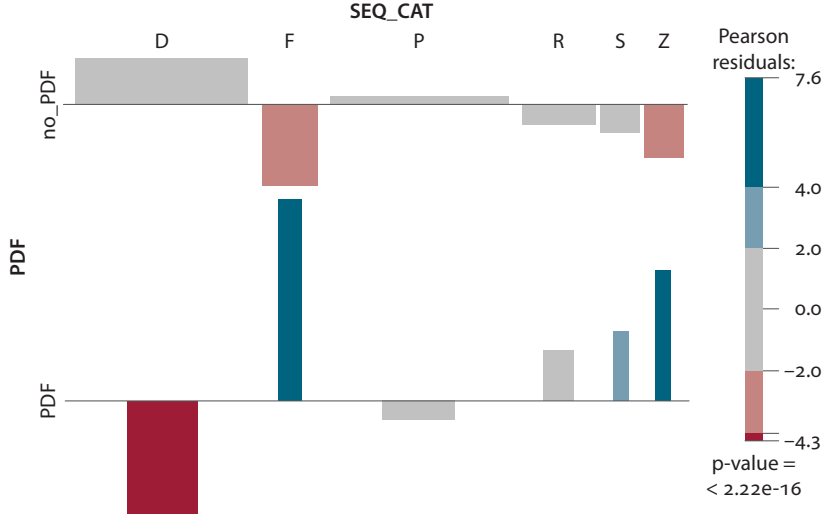


Figure 7.8 Extended association plot of PDFs and non-PDFs across sequence types

significant positive residuals (i.e. more observed than expected frequencies) for PDFs in sequences with interruptions ("F", false-starts and truncations), substitutions ("S") and combinations thereof ("Z", including interruptions with repetitions and/or substitutions).

By contrast and as mentioned before, DM-only sequences ("D") are significantly less frequent for PDFs. It is particularly interesting to note that F- and Z-sequences are positively associated with PDFs and negatively associated to other functions, which points to the particular connection between these sequence types and the functions of *monitoring*, *punctuating* and *reformulation*.

Zooming in on these three functions, it appears that, while all three show larger proportions of F- and Z-sequences than all other functions combined, the proportion for the *reformulation* function is particularly high: 17% of "F" and 9% of "Z", against 7% and 3% on average for the other two functions (4% and 2% for non-PDFs). Sequences of repetitions (R), however, take up a larger proportion in *punctuation*, which is consistent with the "time-buying" role of this function, although not in a significantly different proportion from non-PDFs.

The following examples illustrate the most frequent pattern for each sequence type with positive residuals (blue boxes in Figure 7.8).

(15) il avait donné les configurations qu'il a qu'il a qu'il avait qu'il avait choisies pour *enfin* la manière dont il configurait ses routeurs pour faire ça
he had given the settings that he that he that he had that he had chosen for enfin '*well' the way he set up his routers to do that* (FR-conv-01)

(16) <ICE_9> and what is she doing these days where is she working
<ICE_10> for an interior design c- *well* not design uhm (1.427) furnish (0.420) company (EN-conv-02)

(17) the (0.347) p- tradition in painting is very much for (0.730) the artist (0.213) to reveal himself *or* the artist to reveal his own attitude (EN-intr-04)

In (15), the reformulating DM "enfin" follows a false-start on "pour" and leads to a new phrasing of "les configurations" by "la manière dont il configurait" (F-sequence). In (16), <ICE_10> substitutes "interior design" by "furnish" after the truncation of "company" and various pauses (Z-sequence). Lastly, in (17), there is a substitution of "himself" by "his own attitude" with a modified repetition of "the artist to reveal": here, the reformulation brought about by "or" is not as clearly corrective as in (16) but rather seems to specify the referent in the first segment which is not completely erased by the second one (see the approach in Chapter 8 to account for such distinctions).

7.3.3 PDFs and sequence structure

The finer classification of sequences by their internal structure might shed some additional light into the complexity and disruptiveness of sequences containing PDFs. Besides the smaller proportion of isolated DMs already discussed, PDFs mostly differ from non-PDFs by their larger proportion of (1) mixed sequences of multiple compound and simple fluencemes both embedded and peripheral (especially with *reformulation*, Example (18)), (2) single compound fluencemes with simple fluencemes in both positions (for all three functions, Example (19)), (3) single compound fluencemes with peripheral simple fluencemes (especially with *punctuation*, Example (20)) and (4) single compound fluencemes with embedded simple fluencemes (especially with *reformulation*, Example (21)).

(18) 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)

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

(20) 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)

(21) <ICE_32> a rebate is what
<student> is it when they send the money back
<ICE_32> yes but I mean in what sense *how I mean how how how* do you define it in economic terms (EN-clas-02)

With these examples, we see how the three PDFs are related to disfluency each in their own way, either by introducing a nuance or correction (18), calling for cooperation and help during lexical access trouble (19), stalling for planning and maintaining the floor during a very long pause (20) or rephrasing with a different syntactic construction (21). The rarity of these types of structures in the data and their attraction to PDFs support the classification of these functions as a subset tending towards the disfluent end of the scale, in line with the fluency-as-frequency hypothesis.

The last variable in the equation is sequence length, which was repeatedly identified as a reliable indicator of disfluency, especially for medium-size sequences, in combination with structural complexity. Figure 7.9 represents the curve of sequence length, measured by number of fluenceme tokens, across proportions of PDFs and non-PDFs.

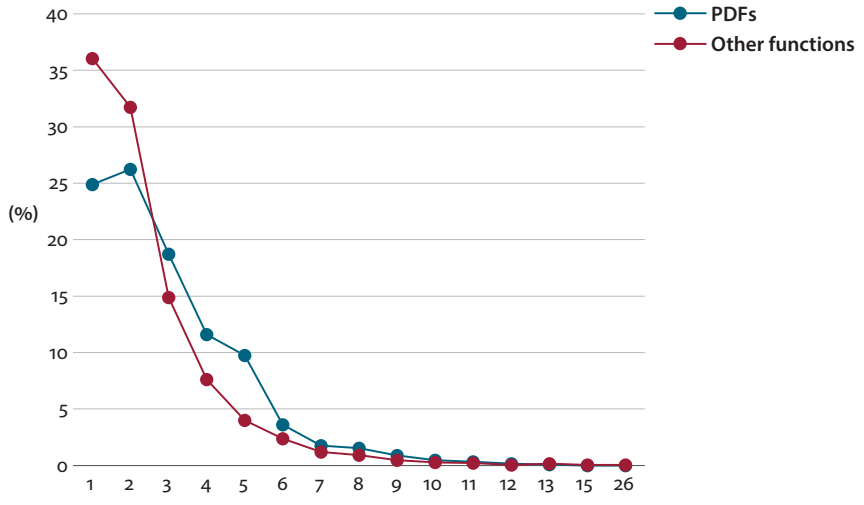


Figure 7.9 Length of sequences in fluenceme tokens in PDFs and non-PDFs

The difference for one-token sequences (here, one-DM sequences) can be explained by the preference of PDFs for clustered and co-occurring uses, compared to non-PDFs, which are more often isolated. More interestingly, we see that the blue curve tops the one for non-PDFs from three-token sequences until seven-token sequences, where the differences cease to be significant. In other words, sequences from three to seven fluenceme tokens are significantly more frequent in PDFs than in all other functions combined, which corresponds to the above-mentioned medium-size subset previously related to disfluent contexts. In particular, Pearson residuals (computed with an extended association plot) show that five-token sequences are typical of PDFs (positive for PDFs, negative for non-PDFs). Qualitative exploration of these 122 cases reveals that they often include two or more pauses and quite frequently other fluencemes such as false-starts or identical repetitions, as in (22).

(22) <ICE_4> it was the local one (0.180) it was not pasteurised milk
<ICE_3> yes (0.280) the Brie and the butter is superb
<ICE_4> and it was *very uh (0.260) well we we* often say that (0.800) farmhouse cheese some of the French farmhouse cheese in this country are smelly but this was (1.060) distinctly smelly (EN-conv-07)

This sequence contains a false-start at "very", a filled pause "uh", an unfilled pause, a DM "well" expressing *punctuation* and an identical repetition of "we". This example is particularly telling of the conceptual proximity between the *punctuation*

and *reformulation* functions within PDFs since both interpretations could be motivated in this excerpt. The absence of syntactic or semantic connection between the left and right contexts signals the beginning of a new start after an interruption (*punctuation*), although it could also be suggested that, at a very general level, the introduced segment is a reformulation of the previous aborted one. In any case, the association between PDFs and medium-size sequences, especially five-tokens sequences as in (22), provides yet another validation of their categorization as "potentially disfluent".

Once more, I will end this section by a summarizing multivariate model evaluating the weight of the different variables analyzed so far. A logistic regression predicting the function of DMs (PDFs or not) was computed with register, sequence type, sequence structure and sequence length as input factors ($C = 0.683$, $r^2 = 0.093$) and returned the following significant effects:

- PDFs are significantly attracted to the conversational register (compared to classroom lessons) and negatively associated with all other settings except face-to-face interviews (which are not significantly different between PDFs and non-PDFs);
- sequences of interruptions (F), mixed fluencemes (Z) and, to a lesser extent, substitutions (S) increase the chance of PDFs compared to isolated DMs;
- the longer the sequence (in number of tokens), the higher the probability of a PDF (marginally significant).

The internal structure of the sequences, although included in the final regression after stepwise model selection, did not return any significant effect. To conclude this section, the close investigation of the subset of "Potentially Disfluent Functions" allowed me to confirm their tendency towards disruptive and disfluent uses through cross-tabulation with annotations and metadata previously identified as less functionally ambivalent. In other words, all functions in the present DM taxonomy are not equal in terms of (dis)fluency and the converging evidence analyzed in this section makes it possible to validate the conceptual category of PDFs through multiple corpus-based variables. Although these promising results should not be over-generalized (not all occurrences of PDFs would necessarily be produced and perceived disfluently), they do illustrate the potential of corpus-based discourse analysis for fluency research, which should also benefit from additional methods of investigation.

7.4 Summary

This chapter tested a number of hypotheses put forward in Chapters 2 and 3 regarding the distribution of fluencemes, the relationship between discourse markers and other fluencemes in the typology, and the association between some DM characteristics and sequence types, always pursuing the usage-based programme of a frequency-based scale of (dis)fluency.

Focusing on DM-based sequences across all registers in *DisFrEn*, it appears that 60% of DMs are clustered with at least one other type of fluenceme, and that this proportion varies with the degree of preparation available to the speakers: much more clustered in (semi-)prepared registers vs. more balanced (even more isolated) in spontaneous settings, which is first evidence against the hypothesis on the similarity between intermediary (semi-prepared) and informal registers. Each sequence type shows a particular attraction to one context or another, namely DMs only in interactive settings (which may be related to the shorter size of turns), pauses in formal settings, interruptions in conversations, repetitions in radio interviews and substitutions in face-to-face interviews. Regarding substitutions, not all of them are disfluent and especially not in interviews, where they are relatively frequent, in concordance with the fluency-as-frequency hypothesis. The most disruptive uses of substitutions tend to be found in registers where this type of sequence is comparatively rare as in phone calls.

In terms of restrictions of sequence types by register, news broadcasts appear to show the least variation, which means that an atypical sequence will be more marked in this highly constrained register than in conversations where they are more usual. Similarly, sequence types restricted to informal settings are less ambivalent (more disfluent) than the pervasive pauses and DMs. Sequences combining compound and simple fluencemes were found to be significantly drawn to intermediary registers, which tends to confirm the special role of these settings, where speakers' attention towards their own speech is heightened. Some DM expressions were also found to be restricted to sequence types indicating their relative (dis)fluency, such as *when* (rather fluent) or French *disons* 'let's say' (rather disfluent).

A cognitive-functional scale of (dis)fluency was proposed based on the mapping of functional domains with sequence types, namely, by decreasing order of fluency: sequential (clusters with pauses for major segmentation functions), ideational (isolated, well-integrated DMs), rhetorical (special attraction to sequences of mixed fluencemes) and interpersonal (special attraction to the non-ambivalent fluencemes of false-starts and truncations). However, limitations to this scale were soon brought about by the inclusion of a third variable in the equation, viz. syntactic position, where no three-way interaction clearly emerged from the data. Expectations of domains, positions and sequence types were not met (such as the

hypothesized link between final interpersonal DMs and interruptions, or medial rhetorical DMs and substitutions or mixed fluencemes), suggesting that this scale should be refined at function level and with more detailed macro-labels for fluenceme sequences.

Zooming in on a subset of DM functions, the analysis of three "Potentially Disfluent Functions" (PDFs) revealed the strong association of *monitoring, punctuation* and *reformulation* to informal, interactive settings, as well as their higher tendency to cluster with fluencemes than the other functions in the taxonomy. Their association to sequences of interruptions and mixed fluencemes (F and Z) as well as to medium-size sequences (especially five-token long) all converge in confirming this top-down category of rather disfluent functions of DMs.

7.5 Interim discussion: The "silence" of corpora

So far, the analyses and results have illustrated the potential of corpus-based fluency research, and in particular the merits of paradigmatic annotation to describe the inter-relations between members of complex categories such as fluencemes. Such large-scale coverage of the investigated phenomena allows for powerful statistical modeling techniques which reveal the significant association (or repulsion) between different independent variables, thus vouching for the reliability of the conclusions.

The contribution of this research compared to the bulk of fluency studies was also to bring register variation to center stage with a wide panel of interaction settings, and to use this metadata information to interpret the observed patterns in light of cognitive and interactional hypotheses. One last remarkable feature of corpus-based linguistics is the ability to confirm top-down categories and theoretical hypotheses by converging evidence of different types (form and function, syntax and pragmatics, annotations and metadata) to a much larger extent than studies which do not rely on empirical authentic data, or even than experimental research which is usually restricted to one or two contrasted conditions.

In sum, the intensive (fine-grained) and extensive (paradigmatic) annotations in *DisFrEn* offer a strong basis for quantitative modeling of complex, highly variable categories such as DMs and fluencemes and provide empirical validation to abstract constructs still under debate in current research.

However, the previous chapters have also shown the limitations and drawbacks of a corpus-based approach to (dis)fluency. Firstly, corpus (and in particular statistical) analyses fail to account for the high variation in the data for such complex phenomena as DMs and fluencemes. There seems to be an irreconcilable gap between statistical patterns on the one hand and particular instances on the other,

so that findings are always limited to rather coarse-grained trends beyond which authentic examples cease to match the generic rule.

Secondly, corpora are "silent" in terms of perception and online interpretation, especially in the field of fluency research, where literature has amply shown that fluency ratings and judgments are not all "rational", i.e. based on observable formal features of the language (e.g. Ejzenberg 2000). A linguist's interpretation of the relative (dis)fluency of a particular example will not necessarily match the perception of the original speakers participating in that interaction, which is where experimental studies come into play and complement corpus-based research. The next chapter will address some of these limitations by providing a more direct access to fluency interpretations through a more qualitative, conversation-analytic approach to the same data.

CHAPTER 8

Discourse markers in repairs

In Chapter 7, I investigated the relationship between the functional behavior of DMs and the types of fluenceme sequences in which they occur in order to see whether this combination of functional and formal variables could refine our interpretation of (dis)fluency and bring us closer to a cognitive-functional scale of (dis)fluency. In the present chapter, I pursue the same endeavor with different empirical evidence, namely a qualitative categorization of particular sequences of fluencemes which identifies the cause of the repair, turning from the *how* to the *why* of (dis)fluent sequences. More specifically, in addition to the word-level tagging of fluencemes and fine-grained annotation of DMs, sequences of fluencemes are here classified functionally through a qualitative identification of the cause or motivation behind the repair (e.g. the *reparans* corrects an error in the *reparandum*). The following examples illustrate the scope of this chapter:

(1) and they in fact were responsible (0.570) or added to contributed to (0.220) to the abdication the uh abolition (0.400) of slavery (EN-intf-05)

(2) they all want to come and have a go and they all want to (0.247) chat and talk (EN-intf-02)

While both examples contain similar fluencemes (namely modified repetitions, pauses, discourse markers and propositional substitutions), they illustrate the functional ambivalence of fluencemes, from corrective to non-corrective, stagnating or progressing, disfluent or fluent: the verbs ("were responsible", "added") and the noun ("abdication") are replaced by more accurate terms ("contributed", "abolition") inserted within repeated words ("to" and "the") in the first case, while the repetition of the main structure in the second case adds new propositional content and moves the narration forward. Therefore, the objective of the present chapter is to refine the information available from the annotation of discourse markers with an additional layer of analysis, digging into the speakers' intentions.

The major influence behind the present chapter is Levelt's (1983) typology of repair, which makes a basic distinction between error-correction and appropriateness-adjustment. The analysis carried out in the following sections strives to relate DM uses to different repair types in English and French. In doing so, I hope to complement the tentative scale which has been sketched so far and

against which DMs could be "diagnosed", from very strategic and fluent to very disruptive and disfluent uses, thus converging evidence from the findings of the previous chapters.

8.1 Previous approaches to repair

Although this chapter is strongly rooted in Levelt's (1983) model of self-repair and monitoring, other authors have dealt with repair and reformulation from many different perspectives. This section provides a selective review of the most relevant works in the field, from which a number of research questions and hypotheses have been gathered. It will become apparent that, although they do not exactly cover the same phenomena, repair and reformulation are both very much related to (dis)fluency, especially in connection with DMs.

8.1.1 Reformulation and its markers: The French classics

Interest for reformulation sprung in French linguistics in the 1980s with three major contributions to the field: Charolles & Coltier (1986), Gülich & Kotschi (1987) and De Gaulmyn (1987).

8.1.1.1 *Charolles & Coltier*

Charolles & Coltier (1986) focused on paraphrastic reformulation in French written texts. They consider paraphrastic reformulations as a sign of the writer's skill and intention to attend to the reader's needs. Reformulations are defined as developments or expansions of a term by a new formulation to which it is equivalent, and are necessarily signaled by a marker such as *c'est-à-dire* ('that is to say'), *autrement dit* ('to put it differently') or *en d'autres termes* ('in other words'), expressions which qualify as DMs, although not labeled as such by the authors. Charolles & Coltier (1986) further distinguish three subtypes of paraphrastic reformulation, namely consecution, correction and denomination, which are expressed by partially specialized markers, as in the following (invented) examples from their paper (1986: 56–57):

(3) Le R.P.R., *autrement* dit J. Chirac, n'est pas contre la cohabitation.
The R.P.R., autrement dit *'to put it differently' J. Chirac, is not against cohabitation*.

(4) Le R.P.R., *c'est-à-dire* J. Chirac, n'est pas contre la cohabitation.
The R.P.R., c'est-à-dire '*that is to say' J. Chirac, is not against cohabitation*.

(5) Le R.P.R., *c'est-à-dire* le Rassemblement pour la République, n'est pas…
The R.P.R., c'est-à-dire *'that is to say' the Rassemblement pour la République, is not…*

According to their analysis, Example (3) is a case of consecution which could be replaced by *donc* 'so' and expresses an argumentative value; (4) is an example of corrective reformulation which could be marked by *enfin* 'well'; and (5) illustrates denomination, typically expressed by *ou* 'or'. The authors stress the fact that the paraphrastic relation between the elements connected by the marker is not an intrinsic property of these elements but the result of a deliberate discursive act by a cooperative writer, in order to ease the interpretation process. In this sense, their definition is entirely compatible with the fluent or "signal" account of fluencemes in general (cf. Clark & Fox Tree 2002) and the addressee-oriented function of DMs in particular (e.g. Hansen 2006).

8.1.1.2 *Gülich & Kotschi*

Gülich & Kotschi (1987) work on speech and focus on paraphrastic reformulation, of which they identify two main types: auto-reformulation and hetero-reformulation, targeting either one's own utterance or someone else's, respectively (cf. self- vs. other-repair in conversation-analytic terms, Schegloff et al. 1977). They further distinguish three subtypes which are quite different from those identified by Charolles & Coltier (1986): paraphrase (with semantic equivalence, either as an expansion, a reduction or a variation), correction (partial or total cancellation of a faulty utterance) and rephrasing (repetition of the syntactic and lexical structure).

In a later article (Gülich & Kotschi 1995), however, this complex picture is reduced to a major dichotomy, which will prove seminal in future works (see next section): expansion (either specification or explanation) vs. reduction (summary or denomination of a complex matter). These two types represent different moves or directions of the reformulation, as illustrated by the following examples:

(6) Tarbull would say the railroads are common carriers I mean they are obliged by their charters not to discriminate in this way (EN-clas-02)

(7) we live in a (0.500) small rural village on the edge of the Mendip hills (0.660) uh and we're about four miles from the sea (0.520) uh with the river Severn and th- the channel (0.530) leading into the atlantic (0.890) so uh it's a beautiful area (EN-intf-06)

In Example (6), the speaker explains what he means by "common carriers" with a longer phrasing introduced by "I mean", thus developing the first utterance, while in Example (7), a reverse move of reduction is introduced by "so" which summarizes the previous lengthy description into a simpler description "it's a beautiful

area". Gülich & Kotschi (1995) share with Charolles & Coltier (1986) the claim that paraphrastic reformulation is always signaled by dedicated markers, although they admit that prosody alone can take on this marking function (1987: 44).

8.1.1.3 *De Gaulmyn*

Finally, De Gaulmyn (1987) differs quite neatly from the two previous references by taking into account the particular features of spoken (unplanned) discourse. While basically taking up Gülich & Kotschi's (1987) taxonomy, De Gaulmyn (1987: 86) further distinguishes four subtypes of rephrasing which she terms "repetition": repetition (including modifications by partial addition or substraction), delayed restart, repetition of a truncation, and repetition of self-dictation. Some of these subtypes of repetition correspond to others in our typology of fluencemes: the first type corresponds to the annotation of insertions and deletions embedded in modified repetitions, as in *the house the big house*; repeated truncations such as *the g- g- girl* would also be accounted for by the annotation system with increasing numbers. However, only the first two subtypes correspond to reformulations (i.e. bringing forward a change in form or content), so that her typology would only be partially relevant to the present study, in addition to its lack of empirical validation.

These seminal typologies have been very influential in more recent works (e.g. Cuenca 2003; Ciabarri 2013), and some of the distinctions are still relevant for the present approach to fluencemes. However, the authors do not provide compelling evidence for the empirical validity of their sometimes subtle distinctions. Moreover, they all share a focus on paraphrastic reformulation, which may be too restrictive against the broad range of repair categories potentially expressed by fluencemes. Finally, while their interest for reformulative markers might seem promising for this chapter on DMs in repair, the presumably necessary presence of a marker in a reformulation is a bold claim which remains to be tested in authentic data, as I will attempt below.

8.1.2 Contrastive perspectives on reformulation markers

The next series of notable works on (markers of) reformulation is very much indebted to the classic references presented above, and consists mainly of contrastive approaches, with the exception of Ciabarri (2013) who compared modes of communication instead of languages. They all share with their predecessors a strong focus on the markers which can signal reformulation, as well as similar typologies regarding subcategories or functions of reformulation. However, with the emergence of corpus linguistics, most of these recent works make use of authentic data to support their claim, apart from Rossari (1990, 1994) who still belongs to the theoretical tradition of Charolles & Coltier (1986) and others.

8.1.2.1 *Rossari*

In her French-Italian project, Rossari (1990, 1994) built a model of "reformulation operations" drawing on Roulet et al.'s (1985) framework of interactive functions for pragmatic connectors. She makes a major distinction between paraphrastic and non paraphrastic uses and focuses on the latter, of which she further identifies four types: "*récapitulation*" (summarization), "*réexamen*" (reexamination), "distanciation" and "*renonciation*" (renunciation).

Like her predecessors, she adopts a marker-based approach to reformulation whereby the presence of a dedicated marker is necessary to identify a case of reformulation and its particular subtype. However, she nuances this criterion and restricts it to non paraphrastic reformulation. Paraphrastic uses, on the other hand, can be signaled by other (syntactic, lexical or prosodic) cues and indicate a general relation of equivalence or replacement similar to other mechanisms of repair (or "*reprise*" in French). In this sense, paraphrastic reformulation seems to be closer to the generic construct of (dis)fluency whereby an on-going utterance is interrupted, repeated, replaced and/or modified.

The core of her contribution lies in the contrastive study of selected French markers and their Italian equivalents, of which she compares a number of characteristics and uses. She concludes on the prevalence of pragmatic weight over morpho-semantic properties. Overall, Rossari's (1990, 1994) approach remains formal and prescriptive: the lack of empirical validation, in addition to the circular definition of reformulation by its markers, does not recommend the use of her categories in a bottom-up approach such as the present one.

8.1.2.2 *Murillo*

In the same line of research, Murillo (2016) proposes a theoretical account of reformulation markers grounded in the notion of polyphony, which was already prominent in Rossari (1994). The author compares the merits of Relevance Theory (Sperber & Wilson 1986), the Theory of Argumentation in Language (Anscombre & Ducrot 1983) and the *théorie scandinave de la polyphonie linguistique* or ScaPoLine (Nølke 2006) in their treatment of reformulation markers, for which she identifies a large number of functions divided in two groups:

- functions related to explicit content: identification of referents, specification, orientation, explanation, introduction of restrictions, correction;
- functions related to implicit content: definition of terms, denomination, conclusion, mathematical operation, and consequence.

We see that these functions are quite heterogeneous and fine-grained, with some surprising members (mathematical operation, for instance) and few details to reliably identify them. Her final model distinguishes two "patterns" of reformulation

markers with different degrees of polyphony, which are defined according to the number of "*locuteurs*" (speakers) and "*énonciateurs*" (enunciators) as well as the type of reported speech (indirect, quasi-indirect, direct, pseudo-direct).

By applying this complex and abstract analytical grid to a Spanish-English corpus, Murillo (2016) finds a higher polyphony of implicit content-related functions in general and of Spanish markers in particular. Although corpus-based, this proposal seems particularly abstract and not directly related to the concept of fluency, which makes it difficult to adapt to the aims of this chapter.

8.1.2.3 *Cuenca and Ciabarri*

The next group of contrastive references is more strongly attached to the field of DMs studies and discourse analysis, striving to situate reformulation in a comprehensive view of (meta)discourse functions such as contrastive relations or common-ground requests. Cuenca (2003), Cuenca & Bach (2007) and Ciabarri (2013) are convincing representatives of this approach. Cuenca (2003: 1071) defines reformulation as "a discourse function by which the speaker re-elaborates an idea in order to be more specific and 'facilitate the hearer's understanding of the original' (Blakemore 1993: 197), or in order to extend the information previously given", which reminds us of Charolles & Coltier's (1986) addressee-oriented definition. She starts by analyzing the forms of reformulation markers (simple vs. complex, different unit lengths, lexical-semantic groupings) in English, Spanish and Catalan.

In Cuenca & Bach (2007), she combines this formal analysis with a functional layer by taking up Gülich & Kotschi's (1995) dichotomy between expansion and reduction, to which she and Bach add "permutation" (i.e. "a change in the conclusions that can be derived from the first utterance", 2007: 165). The main findings are two-fold: from a contrastive point of view, English tends to prefer fixed and non-polysemous forms (as also shown by Fernandez-Polo 1999) while the two Romance languages use more complex and ambiguous markers; from a more language-internal perspective, specific forms seem to be associated with specific functions, thus relating syntax to discourse.

In a very similar study, Ciabarri (2013) contrasts spoken and written Italian reformulation markers across a functional typology which largely overlaps with previous proposals: to the classic expansion-reduction pair, she adds a third – debatable – group of "discursive" reformulation which includes request for common ground, topic reprise, generalisation and time-taking (applied in particular to the marker *cioè* 'that is').

The discourse-functional perspective of these works, while inspiring for the study of DMs in general and as pursued in this book, might be too focused on the types of markers themselves rather than the types of reformulations. In particular, Ciabarri's category of "discursive reformulations" does not seem to bear any

relationship to what reformulation generally stands for, but rather extends in a slightly incoherent way the typology in order to include all functions of *cioè*. For the purpose of this chapter, such a redundancy with the functions of DMs might be too circular to allow for the identification of patterns of reformulations, where repair types and DM functions need to remain independent variables.

8.1.2.4 *Auer & Pfänder*

The last reference in this cluster of contrastive works is Auer & Pfänder's (2007) qualitative analysis of "multiple retractions" in spoken French and German. The authors insist on the ambivalent use of this type of structure, which consists in "re-us[ing] a syntactic position which has already been filled" (2007: 59) with or without an "anchor", either to signal hesitation, turn-holding or list construction. Its relation to repair is made explicit: "Syntactically speaking, retraction is the basis of repair, but not all retractions do repair work, let alone correct a previous item. Retraction is also the basis of list construction, and it is used for numerous other, non-repair functions" (2007: 59). In other words, retraction is considered a syntactic affordance of French and German which can either be used fluently as a structuring device (to "create cohesion in complex descriptions or argumentations", 2007: 75) or disfluently as stagnating repetitions, an observation which is, in principle, generalizable to all fluencemes according to the hypothesis of functional ambivalence in the present approach. The following examples borrowed from their paper illustrate the two uses of retractions in fluent and disfluent uses:

(8) mais nous sommes des gens qui aimons la mer pour le paysage qu'elle nous offre pour tout ce qu'elle nous apporte en bruit en en odeur euh pour s'y baigner
but we are people who love the sea for the landscape it offers us for everything she gives us in sound in in smell uh for bathing in

(9) elle a trouvé du travail à la à la gare de à la gare de Charles de Marseille
she found a job at the at the station at the Charles the Marseille station

In Example (8), the multiple retraction starting with the anchor "pour" introduces three reasons why the speaker loves the sea, decomposing the attributes of the sea in several arguments in a highly structured way, although the full utterance is not completely planned as attested by the repetition "en en" and the filled pause "euh". In (9) however, the retraction of "à la" does not serve any structuring purpose but rather expresses lexical search, which is also evidenced by the syntactic incompletion of the retracted elements (progressive completion of the prepositional phrase).

Their results indicate that retraction is used quite similarly in the two languages except for an additional rhetorical function in French that does not appear as frequently in German, a stylistic difference which the authors explain by a higher sensitivity to norms and standards in French. Auer & Pfänder (2007) offer a rich

background which is inspiring for the following reasons: it targets spoken language; it manages to encompass very different functions (from local hesitation to global structuring) under a coherent object of study; forms and functions are seen as interacting yet independent; the absence of a marker or "anchor" is a structural possibility but their presence is meaningful; finally, it is more explicitly grounded in the field of repair and fluency studies (rather than DM studies), acknowledging the functional ambivalence of formally similar structures, from fluent to disfluent uses.

To conclude this review of classic and contrastive approaches to reformulation, it appears that the notion of reformulation is narrower than that of repair which is not so much focused on the semantics of discourse relations and DMs but is more structurally defined, and therefore more suited to be combined, in a later stage, with an independent, more discourse-functional level of analysis. Repair not only includes reformulation but also lists, repetitions and false-starts. However, not all functions of reformulation markers are included in repair and the overlap remains partial for cases of specification, for instance. All in all, the term *reformulation* remains too redundant and potentially confusing with some functions of DMs, while *repair* appears to be the best term to account for the full (dis)fluent potential of fluencemes, as far as this study is concerned.

8.1.3 From reformulation to repair: Levelt's (1983) typology of repair

As explained in Section 2.1, the notion of repair was largely developed by Levelt (1983, 1989) in his production-perception model of speech monitoring, both as the general phenomenon and as a structural component, along with the *reparandum* and the editing phase. Levelt's main assumption holds that there are some structural and systematic dependencies between the original utterance (or *reparandum*) and the new one (or *repair*), and that this transfer aims at helping the listener solve the "'continuation problem', i.e. how to relate the repair to the original utterance" (1983: 50).

Levelt (1983) argues that the source of the repair (i.e. whether it is phonetic, lexical, syntactic or more structural such as linearization of messages) has a strong impact on the form of the repair: the corrective action "is based on the character of the trouble, the still available parsing results (such as wording and constituent structure of the original utterance), and the estimated consequences for the listener" (1983: 50). Whether this hearer-orientation is empirically valid remains to be verified. Ciabarri (2013), for instance, suggested that speakers are more self-oriented than writers. Still, this strong statement is in line with our ambivalent definition of (dis)fluency, and the attention given to form-function correlates motivates my resort to Levelt's model and typology, which I will now describe in detail.

The first divide is between overt and covert repairs: the former are actual modifications of previously uttered linguistic material (at any linguistic level), whereas the latter may consist of just a hesitation or repetition without modifying anything and therefore leaving the target of the monitoring impossible to identify. I will follow Levelt and focus on overt repairs only, of which he distinguishes four categories:

Delay repairs (henceforth D-repairs) answer the question "do I want to say this now?" and correspond to linearization problems, where "the speaker may realize that another arrangement of messages would be easier or more effective" (1983: 51). In fluencemes terms, they mostly correspond to false-starts and insertions.

Appropriateness repairs (henceforth A-repairs) answer the question "do I want to say it this way?" and target adequacy with what was previously said, with social features of the interaction, with levels of precision, or other reasons. A-repairs are not errors *per se* but signals of a need for minor changes. Levelt (1983) identifies three subtypes of A-repairs:

- ambiguity in context (AA-repairs), which usually applies to deictics and referentially ambiguous items;
- terminology levelling (AL-repairs), which usually interchanges a generic term with a more specific equivalent, or vice versa;
- terms coherence (AC-repairs), where the repair aims at maintaining lexical or terminological consistency throughout a discourse. Levelt admits that this subtype is often complex to distinguish from AL-repairs and therefore suggests an in-between category, ALC-repairs. This subtype will not be included in the present study because of its ambiguity.

Error repairs (henceforth E-repairs) answer the question "am I making an error?" and can be divided into lexical errors (EL-repairs), syntactic errors (ES-repairs) and phonetic repairs (EF-repairs). It is unclear in Levelt (1983) whether he counts as occurrences of E-repairs cases where an error can be identified against a linguistic norm or standard but has not been identified and repaired by the speaker himself (e.g. uncorrected misarticulation). In order to remain consistent with the annotation of fluencemes, such unnoticed cases will not be part of my analysis.

R-repairs are originally defined as the "rest" category for complex cases which are "so completely confused that they defy any systematic categorization" (1983: 55). Since I strive to avoid such coding strategies in my own annotation procedure, I would like to suggest another definition for this category which draws on Levelt's own notion of transferring structural properties from one utterance to the other: *resonance* repairs, which correspond to structures which are partly repeated and partly modified in order to build a strong formal correspondence between their parts, either for the purpose of list construction, contrastive focus or other

rhetorical uses. R-repairs, as they are re-defined in this study, are therefore clearly fluent cases of repairs.

Levelt's (1983) model also includes other variables, rules and assumptions regarding the form of the repairs and the association between form and type of repair. Four major components of a repair are identified: the occasion for repair (i.e. the element which triggered the repair), the moment of interruption (i.e. the type of constituent boundary which is interrupted, from syllable to full utterance), the distance between the occasion and the interruption (originally measured in number of syllables) and the way of restarting the new utterance after the interruption.

One of his most famous (and criticized, e.g. Seyfeddinipur 2006) rules is the so-called Main Interruption Rule which states that speakers tend to "stop the flow of speech immediately upon detecting the occasion of repair" (1983: 56), regardless of linguistic structure and without necessarily completing on-going constituents. His own results lead him to nuance this rule and he admits that a stronger tendency might be to detect trouble towards the end of constituents where attention for monitoring is supposedly higher.

Furthermore, Levelt (1983) analyzes a number of expressions which typically occur between the original utterance and the repair, in the "editing phase". His goal, which I share, is to relate the use of specific editing terms to the source of the repair, focusing in particular on the filled pause *uh*.[33]

Levelt's (1983) model, and in particular his repair typology, has been directly replicated in a number of studies (e.g. Brédart 1991; Geluykens 1994; Fox et al. 1996; Kormos 2006) which will not be discussed any further here since they follow different, less related agendas (e.g. L2 studies). Other publications can be related to Levelt's framework in that they acknowledge the ambivalence and potential productivity of non-standard structures, such as Auer (2005), Ginzburg et al. (2014) and Du Bois (2014). These authors all share the idea that disfluencies are resources that truly belong to grammar and should therefore be viewed as regular discourse moves.

8.1.4 Research questions and hypotheses

In this chapter, I target the use of discourse markers in sequences of fluencemes which correspond to different repair types from Levelt's (1983) model. Each of his categories displays an intrinsic degree of fluency which I repeat here: E- and D-repairs are strong disruptions of the syntactic, lexical and/or phonetic structure

33. This use of "editing term" refers to Levelt's (1989) terminology: it concerns the (optional) elements occurring in the intermediary position between *reparandum* and *reparans*. In that sense, it differs from "explicit editing terms" which are defined in the fluenceme typology as "lexical expression[s] by which the speaker signals some production trouble" (cf. Section 4.3.1.3).

and occupy the disfluent end of the scale; A-repairs are moderate changes which signal a lack of appropriateness, thus intermediate on the scale; R-repairs (redefined presently as *resonance*) are creative uses of repetitions for structuring or rhetorical purposes and they stand therefore on the fluent end of the scale. This qualitative information was combined with the existing annotations of DM functions in the corpus, in order to answer the following questions:

Are DMs distributed evenly across the different repair types or not, in what position (periphery or editing phase) and with what function? Regarding this final aspect, I will pay particular attention to three functions which are conceptually related to repair, viz. *reformulation* (typically error-correction, also rephrasing), *specification* (precision, disambiguation) and *enumeration* (list construction). Given the polyfunctionality of DMs, I do not start from a list of lexemes but from a group of functions in order to remain consistent with the onomasiological approach adopted in this study, which stands in sharp contrast with the majority of works on DMs and reformulation in particular, as I have shown in the literature review above. This does not exclude the possibility that DMs expressing other functions can occur in the editing phase (i.e. between the *reparandum* and the repaired segment).

Are DMs and modified repetitions (RMs) redundant? I expect that RMs and DMs do not tend to co-occur frequently, since their signaling function would be redundant with each other: structural resonances should be sufficient to instruct the hearer on how to integrate the repair in the original utterance without the additional presence of a (reformulative or other) DM, and vice versa. This is partly in line with Heeman & Allen's (1999) findings which showed that DMs tend to be involved in fresh starts (D-repairs) but not in modification repairs (E- and A-repairs).

Do French and English differ in any way? Results from the contrastive papers reviewed above tend to suggest that Romance languages are more verbose and make use of more complex and more ambiguous markers than English (Cuenca 2003; Cuenca & Bach 2007). I therefore expect to find more types of DM lexemes in French than English repairs. Moreover, Auer & Pfänder (2007) found that French has a tendency to build parallel constructions with a rhetorical function, which should show in the data as more frequent R-repairs in French than in English.

8.2 Data and method

For this study, I used the subcorpus of face-to-face interviews in English and French from *DisFrEn* (17,000 and 18,000 words in each language, respectively). By carefully reading and listening to the audio-aligned transcription under the EXMARaLDA interface, I progressively extracted all repair sequences in their chronological order, following the criteria presented in the following.

8.2.1 Selection criteria

The scope of this analysis is rather broad: the general rationale is to include any structure or fluenceme which meets the definition of same-turn self-repair and could be analyzed within Levelt's (1983) typology of overt repairs. It therefore covers the following fluencemes: modified repetitions (RM), false-starts (FS), incomplete truncations (TR), propositional and morphosyntactic substitutions (SP, SM), lexical and parenthetical insertions (IL, IP). The other fluencemes in the typology (pauses, discourse markers, completed truncations, identical repetitions) might be termed covert repairs. In these cases, the cause of the repair is internalized and cannot be reliably identified. For this reason, these fluencemes were not included in the selection of overt repairs.

As opposed to the annotation of fluencemes which was primarily formal, the identification of repairs is more flexible, more qualitative and relies more strongly on semantic interpretation of content equivalence. I believe that this independence between the two analytical levels is beneficial for the analysis since it avoids circularity (see also Crible 2017c).

All repairs were identified manually through careful reading of the transcripts on the EXMARaLDA interface, making use of the audio when necessary. Access to prosody turned out to be particularly useful for the coding of repair type.

8.2.2 Repair category

I have already introduced above some of the revisions that were implemented to Levelt's (1983) original typology, namely regarding the selection of uncorrected errors and the re-definition of R-repairs as fluent resonances. Other repair categories required to be specified with more operational criteria in order to ease the coding process. After a first round of analysis on the interviews subcorpus, the final revised typology includes eight different types of repair which can be found in Table 8.1 below (more details and criteria in Crible 2017b).

Table 8.1 Revised typology of repair from Levelt (1983)

Category	Definition	Criteria
Delay	arrangement of messages (D)	insertions; initial fresh starts
Error	lexical error (EL)	EL-bias when hesitation with AL
	syntactic error (ES)	intra-sentential; incl. function-words
	phonetic error (EF)	cf. misarticulation
Appropr.	generic appropriateness (A)	incl. mitigation
	ambiguity of referents (AA)	usually pronouns
	level of precision (AL)	incl. terminology
Resonance	resonance (R)	"list" effect, repetition of form, "fluent"

In addition to these main types, some repair categories can be divided into subtypes which are category-specific. Subtypes only concern AL-repairs, D-repairs and R-repairs. The former can either be related to terminology (i.e. the repair defines a specialized term or specifies a generic statement with a specialized term) or not, when the degree of precision is at stake but involves terms of an equal level of specialization, as in (10) where the speaker defines more precisely what he means by "correctes".

(10) avoir des constructions grammaticales correctes (0.190) c'est-à-dire des constructions grammaticales qui répondent à l'ensemble des règles (0.510) qui sont généralement admises pour la langue
to use correct grammatical constructions (0.190) that is grammatical constructions that meet all the rules (0.510) which are generally followed for the language
(FR-intf-02)

D-repairs can be of three types, which correspond to three fluencemes: false-starts (i.e. interruption of a structure and utter replacement by fresh material with little or nothing in common), local linearity issue (i.e. insertion of one or two words, related to the ordering of words in an utterance) and global linearity issue (i.e. insertion of longer stretches of words for background information or coherence, related to the ordering of information). Each type is respectively illustrated in Examples (11)–(13) below.

(11) it's more of the Liverpool *acc-* but I can certainly tell the difference
(EN-intf-03)

(12) donc euh *on ne peut pas dire qu'i- (0.440) maintenant malheureusement on peut pas dire qu'il* y ait un français
so uh we cannot say th- (0.440) now unfortunately we cannot say that *there is one French language* (FR-intf-01)

(13) but we would always do our utmost (0.690) *to particularly for parents who've travelled from a long distance (0.290) to* find them accommodation (EN-intf-03)

Finally, R-repairs can either be used to create lists or parallelisms. Lists are the unmarked form and simply consist of additions or enumerations of material with a common structure. Parallel R-repairs express a stronger sense of contrast or mirroring between two or several elements, as in exclusive alternatives (Example (14)).

(14) they either go home a (0.130) *a week or two before or a week or two after* the (0.330) due date (EN-intf-03)

While this revised analytical grid for repairs has a broader coverage than the original proposal by Levelt (1983), most modifications correspond to additional types or subtypes (for instance the "local linearity" subtype of D-repairs, or the whole

R-repair category) and can therefore be retrieved and isolated for better comparability with Levelt's results. However, I believe that these revisions help in providing a more comprehensive yet more fine-grained overview of the ambivalence and functional flexibility of repairs.

8.2.3 Relation to annotated fluencemes

In line with the general approach of this study, the present analysis pays particular attention to DMs, their presence, position and function, focusing on those occurring in the editing phase. Each DM is coded for its position with respect to the structure of the repair:

- editing phase, when the DM is located between the original and the new utterance, including when the latter starts with a DM, as in "it's more of the Liverpool ac- *but* I can certainly tell the difference" (EN-intf-03);
- part of the repair, when the original and/or new utterances contain a DM, as in "the monitors go off *wh- even when* we put our hands in" (EN-intf-03);
- periphery, when at least one DM is included at any other place in the sequence containing the repair, as in "a lot of them *actually* head down there head down to the Barbican" (EN-intf-02);
- N/A if no DM is present in the sequence.

Focusing on DMs in the editing phase (including when they are the first word of the new segment), I retrieve from the original annotations the function(s) of the DM(s) in their order of appearance. No other information about DMs is either added or retrieved for this analysis.

Apart from DMs, the presence of some fluencemes (lexical and parenthetical insertions, truncations, false-starts) in the repair is made explicit, strictly following the existing annotations. Modified repetitions (RMs) are also identified when they are central to the internal structure of the repair. I took the liberty of noting the presence of coordinating conjunctions (CC) which are recurrently present in repairs (especially R-repairs), even though they do not qualify as fluencemes (except when they are inter-sentential, in which case they are considered to be DMs and annotated as such).

8.2.4 Intra-annotator agreement

Coding consistency was checked to make sure that no repair occurrence had been overlooked in the transcripts. A second round of blind coding was carried out in order to provide a measure of intra-annotator reliability for repair types. Intra-annotator agreement appears to be quite high with a kappa-score of $\kappa = 0.867$ and 89.37% of agreement across all repair types. This score is considerably higher than for the domains and functions of DMs carried out on a sample of the whole *DisFrEn* corpus. Overall, R-repairs are the most replicable category, especially considering their high frequency, while the most striking source of disagreement is the hesitation between D- and ES-repairs, which together account for half of all disagreements. For each case of disagreement, a final gold-standard value was established and then implemented in the dataset. All in all, the coding of repair is quite robust and replicable.

8.3 Repair categories across languages

I will start with general considerations on the distribution of the different (sub)categories of repair, in order to identify possible crosslinguistic differences. Table 8.2 shows that disfluent repairs (E, D) are the most frequent in the data, followed by fluent R-repairs, then A-repairs (intermediary on the fluency-disfluency scale), in both English and French. This first result contradicts the findings from Chapter 6 where the paradigmatic annotation of fluencemes revealed a higher frequency of the most ambivalent members (pauses, DMs), while typically disfluent fluencemes such as false-starts or explicit editing terms were much less frequent. This difference can be explained by the fact that the analysis in the present chapter only targets overt repairs, as opposed to the wider scope of the annotation which includes fluencemes related to both overt and covert repair. In other words, when considering overt and covert repairs simultaneously, potentially fluent uses are more frequent, whereas within overt repairs, the reverse situation is observed.

Almost half of all repairs (49%) belong to either D-repairs or ES-repairs, which could be merged into a coarse-grained category of "structural" repairs: it would seem that issues of linearization and linearity represent a very important proportion of all overt repairs in the data, as opposed to repairs related to finding "the right word" (lexical error "EL" + appropriateness "*A*" = 32%) and those related to fluent strategies (resonance "*R*" = 17%). This focus of monitoring on form rather than content is, in my view, evidence of the distinctive nature of unplanned speech, where speakers have to order complex information into the linear phonological

Table 8.2 Proportions of repair categories and subtypes by language

Repair category	EN %	FR %	Total %
Error (E)	**36.97%**	**42.08%**	**39.78%**
Lexical (EL)	20.00%	19.31%	19.62%
Syntactic (ES)	14.55%	21.78%	18.53%
Phonetic (EF)	2.42%	0.99%	1.63%
Delay (D)	**30.91%**	**30.69%**	**30.79%**
false-start	19.39%	24.26%	22.07%
global linearity	6.67%	2.97%	4.63%
local linearity	4.85%	3.47%	4.09%
Resonance (R)	**16.36%**	**17.82%**	**17.17%**
list	12.73%	12.87%	12.81%
parallel	3.64%	4.95%	4.36%
Appropriateness (A)	**15.76%**	**9.41%**	**12.26%**
Level of precision (AL)	7.88%	3.96%	5.72%
terminology	4.24%	0.50%	2.18%
N/A	3.64%	3.47%	3.54%
Ambiguity (AA)	5.45%	2.48%	3.81%
Generic (A)	2.42%	2.97%	2.72%
Total	**100%**	**100%**	**100%**
Total occ.	**165**	**202**	**367**

channel as it unfolds, while writers can spend more efforts on other, more subtle aspects of language such as lexical choice. Following this line of reasoning, monitoring for linearity and structure seems to be the priority in speech, which can be explained by the high temporal constraints on spoken production as opposed to the a-temporal nature of writing. In this view, Levelt's (1981) argument that these issues are equally present in the two modalities might be overlooking the time-bound character of speech.

The results for R-repairs show that the unmarked form of fluent resonances ("list") is more frequent than the more elaborate cases of parallels which have an added value of contrast or mirroring. This fluent device therefore seems to be a major resource for simple enumerations or additions, by recycling parts of an utterance to move forward on the syntagmatic axis, rather than for more discourse-functional strategies such as contrastive relations. It seems that the lower frequency of parallels compared to lists can be explained by their more specific meaning which involves some level of planning, all the more surprising in the register of interviews characterized by long speech turns and an intermediary degree of preparation.

From a contrastive perspective, Table 8.2 shows that repairs are slightly (not significantly) more frequent overall in the French subcorpus (202 occurrences vs.

165 in English; $LL = 1.94$, $p > 0.05$), although proportions of repair types are very similar apart from the small differences mentioned at the beginning of this section. In addition, we can see that the largest gap in raw frequencies between the two languages concerns ES-repairs and false-starts, which are both part of the "structural" repairs (either utterance-internal or utterance-initial, respectively): in this data, the French speakers thus seem to show more trouble at this planning level of speech production than the English speakers. Apart from these slight preferences, the two languages seem to behave in a strikingly similar way, which is consistent with the findings of Chapters 5 to 7 regarding the distribution and clustering of DMs and fluencemes.

8.4 DMs in repairs

In the interviews subcorpus, 134 DMs are involved in a repair sequence, 85 of which occur in the editing phase, i.e. between the *reparandum* and the *reparans*. All in all, few DMs are involved in repairs: only 7% of the 1,917 fluenceme sequences which contain a DM are cases of repairs. A similarly low frequency was already observed in Pallaud et al. (2013a), who found that only 10% of "disfluent interruptions" include DMs. This first result indicates that DMs mostly occur independently or clustered with other, mostly simple fluencemes (cf. the "conceptual frequency" of DMs and other fluencemes in Chapter 6).

In number of sequences, 130 repairs contain one or several DM(s) in various positions, 83 of which are located in the editing phase, including 32 occurrences of *reformulation*, *specification* and *enumeration*. It thus appears that 35% of overt repairs include DMs. In other words, the occurrence of DMs does not imply the occurrence of a repair (only 7% of all DMs), but repairs, when they occur, do seem to contain DMs, although only in 1/3 of the time. The tendency of DMs to cluster with pauses and, to a lesser extent, identical repetitions (as shown in the previous chapters) rather shows that DMs are more related to covert than overt repair. Since both overt and covert can be shown to perform fluent and disfluent roles, no further conclusion can be drawn from this first observation of frequency alone.

8.4.1 Position of the DMs

In order to associate DMs with a particular degree of fluency, we can look at their distribution in different repair types. Table 8.3 shows the proportions and frequency of DMs across various positions in the repair, if any, cross-tabulated by repair type. As a reminder, three positions are possible within a repair: in the editing phase

(EP), part of the repair itself (repair), or anywhere else in the sequence (periphery). In case of a sequence containing several DMs, the table was simplified according to the following bias, ranked by degree of centrality in the repair: editing phase > repair > periphery. For instance, if a sequence contains a DM in the editing phase and another in the periphery, only the editing phase is counted.

Table 8.3 Distribution of DMs across repair types and positions in the repair (if any)

	Presence of a DM	Position of the DM		
		EP	Periphery	Repair
Delay	**43% (49)**	**31**	**15**	**3**
Error	**30% (44)**	**25**	**15**	**4**
phonetic (EF)	17% (1)	0	1	0
lexical (EL)	31% (22)	17	4	1
syntactic (ES)	31% (21)	8	10	3
Appropriateness	**40% (18)**	**11**	**6**	**1**
generic (A)	60% (6)	4	1	1
ambiguity (AA)	43% (6)	2	4	0
precision (AL)	29% (6)	5	1	0
Resonance	**30% (19)**	**16**	**2**	**1**
Total	**35% (130)**	**83**	**38**	**9**

We can see that, when a DM is present in the repair, it is mostly located in the editing phase "EP" position, less so in the periphery, and very rarely in the repair itself, which means that DMs are more often part of the solution (signalling the interruption or beginning of the new utterance) than of the problem (being repaired themselves). A substantial proportion of DMs in the periphery (2/3) concern typically disfluent and structural repairs (D + ES). We can wonder whether it is precisely the DM that triggers or causes the repair, that is, whether the presence of a DM in the local context is a symptom of poor planning. Qualitative analysis of these cases reveals that many of the examples are utterances which begin with a DM, often *and*, *so* or *well* in English, as in Examples (15) and (16).

(15) anything that'll (0.200) could possibly go wrong we are tested on and we have to cover (0.840) *so the uh* it's been it's been fun (EN-intf-02)

(16) <BB_1> are you responsible for (0.230) organising that or somebody in your department (EN-intf-03)
<BB_4> *well it* as individual nurses we are allocated to care for babies

These two examples of D-repairs show initial DMs (in the utterance and in the turn, respectively) leading to an interruption after the next word, a function-word in both cases ("the", "it"). This result on interruptions corroborates Clark & Wasow's

(1998: 208) findings on repetitions and their model of speech production in four stages: (1) initial or preliminary commitment to abide to the "temporal imperative" of speech even though the utterance is not entirely planned; (2) suspension of speech, usually after the first (function) word; (3) hiatus (such as the filled pause "uh" in Example (15), absent in Example (16)) and (4) restart. The frequency of this pattern in my data and its compatibility with Clark & Wasow's (1998) model suggest that DMs, similarly to identical repetitions, might be used by speakers as an automatic strategy to hold the floor and maintain the flow of speech active under time pressure, even though the full plan of the utterance is not ready yet and might be modified.

Although we can see in Table 8.3 that each repair type occurs more frequently without a DM, DMs still appear to be particularly frequent in the EP of D-repairs (27% of the sequences), closely followed by R-repairs (25%) and EL-repairs (24%), as in the following examples, respectively.

(17) find somebody in the hospital first (0.370) to see *if you know because* it's easier (EN-intf-03)

(18) they all want to come and have a go *and* they all want to (0.247) chat and talk (EN-intf-02)

(19) tous Liégeois (0.640) dont il y a plus qu'un qui vit (0.320) *enfin* deux
all from Liège (0.640) of whom only one still lives (0.320) enfin '*well' two* (FR-intf-03)

It appears that, for the editing phase position, DMs occur in similar proportions across very different types of repairs, respectively at the disfluent, fluent and intermediary ends of the (dis)fluency scale, and can therefore not be associated to a particular degree of fluency. Proportions of DMs in the different types of A-repairs are not relevant to analyze given the small number of occurrences. Overall, no major pattern of association between DMs and repair types emerge from the sole observation of their presence or absence in the editing phase, which suggests taking into account more information, such as the particular lexemes and their functions.

8.4.2 DM lexemes

A first observation of the DM lexemes in the corpus confirms the contrastive hypothesis inspired by previous studies comparing English and Romance languages: Romance languages are more verbose and make use of more complex and more ambiguous markers than English. The list of DMs located in the editing phase with their raw frequency is the following for the two languages:

- in English: *and* (6); *you know* (5); *because* (4); *or* (4); *well* (3); *actually* (2); *but* (2); *I mean* (1); *so* (1); *then* (1); *when* (1);
- in French: *enfin* ('I mean', 13); *et* ('and', 7); *hein* ('you know', 6); *ou* ('or', 6); *bon* ('well', 4); *c'est-à-dire* ('that is to say', 3); *mais* ('but', 3); *quand* ('when', 3); *alors* ('then', 2); *etcetera* ('etcetera', 2); *puis* ('then', 2); *bon ben* ('well', 1); *donc* ('so', 1); *du moins* ('at least', 1); *en fait* ('actually', 1); *en tout cas* ('anyway', 1); *et puis* ('and then', 1); *je dirais* ('I would say', 1); *voilà* ('there', 1); *vous savez* ('you know', 1).

We can see that the French list is twice as long as the English one, which mostly contains conjunctions, adverbs and a few expressions more specific to spoken conversation (*well, I mean*). This observation supports the hypothesis of the verbosity of Romance languages (here illustrated by the heterogeneity in the list of DM types) as opposed to the tendency of English to use specialized forms. Many lexemes are *hapax legomena*, and the most frequent do not all have a core reformulative meaning. In English, *well* and *or* could be expected, but *and*, *you know* and *because* are not, while *I mean* has only one occurrence. In French, *enfin* ('I mean') and *ou* ('or') are typical markers of reformulation, unlike *et* ('and'), *hein* ('you know') or *bon* ('well') which are all more frequent than the typical *c'est-à-dire* ('that is to say'). These "unexpected" DMs are actually motivated by a number of reasons related to their function and the subtype of repair they occur in.

8.4.3 Potentially Disfluent Functions in repairs

Table 8.4 reports on the functions of DMs in the editing phase, counting each function label as an individual occurrence in case of sequences containing several DMs or DMs expressing two functions (hence a total of 95 instead of 83). It should also be reminded that some occurrences of these DMs are not "editing terms" *per se* but are located in the editing phase, sometimes as the first word of the new utterance.

DMs expressing *reformulation* (mostly *or* in English, *enfin* and *ou* in French) are the most frequent, which is a natural result given the nature of the repair phenomenon. They mostly occur in EL-repairs (13 occurrences), then A-repairs (8) and a few cases of structural repairs (6 in ES + D combined). The association between the *reformulation* function and lexical repairs is in part due to the very definition of the label, and suggests that this function should be situated at an intermediary degree on the (dis)fluency scale, which tends to confirm its categorization as a "Potentially Disfluent Function" (cf. Section 7.3). A typical example of *reformulation* in EL-repair can be found in Example (20).

(20) and they in fact were responsible (0.570) *or* added to contributed to (0.220) to the abdication the uh abolition (0.400) of slavery (EN-intf-05)

Table 8.4 Functions and frequent lexemes of DMs in the editing phase

Function	Nb of occ.	Lexemes FR	Lexemes EN
Reformulation	27	*enfin; ou*	*or; well*
Addition	14	*et*	*and*
Monitoring	12	*hein; vous savez*	*you know*
Specification	8	*enfin; c'est-à-dire*	*actually*
Topic-resuming	5	*donc*	*but; so*
Punctuation	4	*bon*	–
Temporal	4	*alors; quand*	–
Cause	3	*comme*	*because*
Opposition	3	*mais; bon*	*but*
Alternative	3	*ou*	*or*
Emphasis	2	*du moins*	–
Motivation	2	–	*because*
Condition	2	*quand*	–
Ellipsis	2	*etcetera*	–
Hedging	1	*je diras*	–
Closing	1	*voilà*	–
Concession	1	*mais*	–
Contrast	1	–	*and*
Total	**95**	–	–

It might be surprising to see that, after *reformulation*, *addition* is the second most frequent function in the editing phase of repairs: closer investigation of these cases reveals that 10 out of the 14 occurrences are R-repairs of the subtype "list", where the DM (usually *and* or *et*) enumerates by basic addition the different members in the list, as in (21).

(21) les Français maîtrisent bien leur langue euh les Belges maîtrisent mon avis bien leur langue *et* les les Canadiens maîtrisent bien leur langue
the French know their language well uh the Belgians in my opinion know their language well et *'and' the the Canadians know their language well* (FR-intf-01)

The third most frequent function, *monitoring*, is also particularly interesting: 11 out of the 12 occurrences occur in rather disfluent repair types (D, ES and EL), with only one exception in an R-repair. This association between the *monitoring* function and disfluent repairs seems to confirm (1) its classification as a "Potentially Disfluent Function" and (2) the corpus results in Section 7.2.1 which showed the association of interpersonal DMs to F-sequences, as in Example (22) where "you know" co-occurs with two truncations and a filled pause to signal trouble in lexical access (EL).

(22) and we have (1.080) been sort of starting (0.300) having p- *you know* mu- uh information leaflets in their (0.350) languages (EN-intf-03)

The pattern illustrated in this example points to the speakers' strategy to call for attention or help when they are in trouble (cf. Beeching (2016: 99) on the function of *you know* to "invite the collaboration of th[e] interlocutor to find the right words"). *Reformulation* and *monitoring*, which are two of the "Potentially Disfluent Functions" (PDFs), have now been situated on the intermediary and disfluent ends of the scale, respectively. The third of these functions, namely *punctuation*, can in turn be connected with disfluency as well, with one case of EL and three of D-repairs, as in (23).

(23) il y a beaucoup de *bon* il y a d'abord des fautes d'orthographe
there are a lot of bon*'well' first there are spelling errors* (FR-intf-01)

Here, the speaker interrupts the original utterance with a false-start and restarts after the DM "bon" with the same presentational structure ("il y a") but the presence of the structuring DM "d'abord" ('first') indicates a change of plan, probably in the linear order of ideas he wants to develop. The relation between *punctuation* and disfluent repairs (structural or lexical) is similar to that of the *monitoring* function. In this perspective, *monitoring* and *punctuation* are similar, which supports the proposal in Crible & Degand (in press) to categorize them as two variants of the same function, one interpersonal and the other sequential.

8.4.4 Specification and enumeration

The *specification* function (fourth most frequent) can be interpreted in relation to the expansive nature of some reformulations: apart from cases of D-repairs, *specification* DMs tend to occur in AL-repairs, as in the next example.

(24) <VAL_2> pensez-vous que l'accent peut (0.327) influencer la façon dont on est perçu (FR-intf-01)
do you think that accent can (0.327) influence the way we are perceived
<VAL_3> perçu de quelle manière *enfin* dans dans quel dans quel
perceived in what way enfin *'I mean' in in what in what*
<VAL_2> premier contact
first contact

In (24), "enfin" ('I mean') introduces a reformulation of the original question "de quelle manière" ('in what way') by more appropriate and more specific terms: we can suppose that the speaker was going to say "dans quel sens" ('in what sense')

before being interrupted by the interviewer. Other examples of *specification* are more related to definitions of concepts (cf. Example (10)). On the other hand, this association between DMs and precision repairs (AL) does not apply to the subtype of "terminology" repairs, where the semantic equivalence between *reparandum* and *reparans* might be strong enough without needing to be marked by an additional signal (here, a DM). Overall, the fact that not many DMs of *specification* are involved in repairs means that the specification applies at a higher level of discourse which escapes the present definition of repair, as in the following example which was not selected as an occurrence of repair:

(25) we're about uh (0.620) twelve miles South of Bristol (0.190) which is a large city in England (0.490) *in fact* as I think it's the sixth largest city in England (EN-intf-06)

While the DM "in fact" signals a specification here, it does not replace one utterance or term by the other but adds new information, from "a large city" to the exact ranking "the sixth largest city". Examples like these might be considered borderline cases of repair, especially if one adopts the approach to reformulation in Cuenca & Bach (2007) or Ciabarri (2013), where "expansion" is one type of reformulation.

A similar observation can be made for the *enumeration* function, which is completely absent from the dataset in spite of the hypothesis regarding its relation to R-repairs and lists in particular. List members appear to be connected by other DMs such as *and* in their basic additive function, which can be explained by the tendency of spoken language to be underspecified and to rely on context to disambiguate polyfunctional forms. On the other hand, enumerating DMs (typically *first of all*, French *d'abord*) are used to connect either longer stretches of discourse such as descriptions or members of a list which are not necessarily built on the criterial "anchor" structure of R-repairs, as in Example (26) which shows no formal resonance.

(26) oh God you just don't *first of all* you don't score so much *and secondly* you only get rid of two letters (EN-conv-08)

I suggest interpreting this absence of *enumeration* in R-repairs in a similar line of reasoning as for the "terminology" repairs: formal resonances between list members in R-repairs are sufficient to signal their connection and do not require additional marking by specific DMs.

In sum, the mapping of DM functions and repair types reveals very interesting and meaningful associations, of which I repeat the most frequent here: DMs with a reformulative function mostly occur in EL-repairs; additive DMs in "lists"

R-repairs; monitoring DMs in disfluent repairs. Occurrences of the other functions are too rare to identify similar meaningful patterns.

8.5 DMs and modified repetitions

Mapping the occurrence of DMs and modified repetitions (RMs) in repairs, it appears that 70% of the DMs in the editing phase do not co-occur with an RM, which would confirm my hypothesis on the redundancy of these two fluencemes. The cases where they do co-occur correspond to the most frequent categories overall, namely the *reformulation* and *addition* function in EL- and R-repairs, respectively. Table 8.5 shows the cross-tabulation of RMs and DMs in repairs, counting each value individually in case of multiple DMs occurring in different slots of the same repair.

Table 8.5 Presence and position of DMs and RMs

	No RM	**RM**	**Total**
No DM	**53% (113)**	**47% (100)**	**213**
DM	**62% (80)**	**48% (49)**	**129**
Editing phase	67% (51)	33% (25)	76
Repair	67% (8)	33% (4)	12
Periphery	51% (21)	49% (20)	41
Total	**56% (193)**	**44% (149)**	**100% (342)**

We can start by noting that DMs are equally absent whether there is an RM in the repair or not. When DMs co-occur with RMs, they are mostly located either in the editing phase or in the periphery (45 out of 49 co-occurring DMs). We can further note that their co-occurrence takes up half (49%) of all peripheral DMs (against only one third in the EP and repair positions), which could possibly indicate that the repulsive effect between DMs and RMs requires the former to perform a central role in the repair, as opposed to peripheral DMs which are more "coincidental" as in Example (27).

(27) ça va (0.560) euh lasser les gens *parce que* on voit une fois on rigole on voit deux fois on rigole encore on voit trois fois on dit pff
it will (0.560) uh bore people parce que *'because' you see it once you laugh you see it twice you laugh again you see it three times you say pff* (FR-intf-03)

By contrast, the proportion of repairs containing a DM is slightly and relatively higher when no RM is involved: repairs with DMs and no RMs take up 62%, which is more than both the proportion of co-occurrence (48%) and that of joint absence (no DM, no RM, 53%). This result suggests that, without the formal cues of a RM to relate *reparandum* and *reparans*, speakers tend to use DMs as if to compensate

the absence of formal repetition, especially to mark the interruption point (51 out of 76 DMs in the editing phase) as in Example (28).

(28) they constr- constructed a huge amount of them here (0.300) *actually* at Qu- Queen Anne's battery (EN-intf-02)

To sum up so far, the absence of DMs does not seem to have any effect on the absence or presence of an RM, but the presence of a DM tends to trigger the absence of an RM (or vice versa), especially when the DM occurs in the editing phase.

Lastly, this "repulsive" effect between DMs and RMs is no longer observed when we focus on the combination of modified repetitions with propositional substitutions (RM + SP): RM + SP patterns are equally frequent with or without a DM. Both cases are illustrated in Examples (29) and (30).

(29) the mums remember you *and* the dads remember you (EN-intf-03)

(30) is it going to look like dad is it going to look like mum (EN-intf-03)

In (29), the two constructions of "the … remember you" are labelled as RM and the SP applies to "mums" and "dads" while the DM "and" connects the two list members. In (30), the structure is quite similar with only one word in each segment being affected by the SP (here "dad" and "mum") and no DM occurs in the editing phase or elsewhere in the repair. The similarity of these two examples, which are both R-repairs, tends to suggest that the presence or absence of the DM could be a structural possibility for each of them, as in the reconstructed version "the mums remember you the dads remember you" which does not lead to major changes in interpretation effects.

Compared to RMs alone, the resonances between the original and the new utterances of a repair are stronger in RM + SPs since they combine partial repetition with semantic substitution. Yet, paradoxically, these stronger resonances do not exclude the extra marking of a DM, while RMs alone tend to be negatively affected by the presence of a DM, especially in the editing phase. It might be that the other patterns including RMs, which are mostly cases of truncations (RM + TR) or insertions (RM + IL), are particularly incompatible with the presence of DMs, possibly because they are more intra-sentential, as opposed to the inter-sentential nature of DMs.

8.6 Summary

A first conclusion of this chapter is, once more, the similarity of English and French texts in terms of the distribution of repairs, which echoes the relative absence of major crosslinguistic differences observed in the previous chapters of this book

(with a few notable exceptions). Only the heterogeneity and number of DMs in the editing phase of repairs were found to be much larger in French than in English, a result which confirms the expectation based on former contrastive studies.

The second general conclusion is that repairs linked to issues of structure (either micro-planning, i.e. local ordering of elements within utterances, or macro-planning, i.e. higher-order arrangement of messages and ideas) are the most frequent and appear to be the priority speakers attend to. In order to fully answer whether this finding is an indication of the higher pressure of temporality in speech than in writing, one would have to monitor the editing process of writers (e.g. Flower & Hayes 1981; Leijten & van Waes 2013). What we can say at this stage is that, according to research on reformulation in written texts, these operations do not target issues of structure but rather of lexical precision or inference management, since structuring and organizational issues are elements of the editing process which are not apparent in the final written product.

The hypothesis on the redundancy and repulsive effect between DMs and RMs was confirmed with a very small number of co-occurrences, although this finding was refined when taking into account the particular type of fluenceme sequence: DMs were indeed absent from modified repetitions containing a truncation (RM + TR) and a lexical insertion (RM + IL), which often correspond to intra-sentential repairs, yet no such repulsive effect could be inferred from the occurrences of propositional substitutions (RM + SP) which are, in turn, more linked to initial re-starts and therefore compatible with the inter-sentential nature of DMs.

The relative absence of DMs in overt repairs indicates their strong link to covert repair, a distinction which I have proposed to map with Ginzburg et al.'s (2014) "backward-" vs. "forward-looking" disfluencies: DMs announce some "work in progress" and upcoming material (cf. also their use to re-start after a syntactic interruption), and are therefore part of the solution, or at least a sign of the search for the solution, instead of being part of the problem, that is in need of repairing. Moreover, the analysis also shows that DMs are often involved in the periphery of disfluent structural repairs, usually as the first word of the interrupted utterance: initiating an utterance with a DM without a full plan in mind allows speakers to hold the floor under time pressure and create an impression of connectivity with previous discourse, even though it often leads to re-starts.

All in all, DMs appear to be used strategically to maintain the illusion of fluency. This general statement can be refined by taking into account the functions that DMs express in meaningful (yet rare) patterns emerging from the data: reformulative DMs occur in intermediate lexical repairs, while monitoring and punctuating DMs are closer to disfluent structural repairs, thus confirming their categorization as Potentially Disfluent Functions. The more frequent the DM function in the corpus,

the more it is involved in fluent repairs (cf. the high frequency of additive DMs in *DisFrEn* and their presence in R-repairs, as opposed to the lower frequency of *monitoring* DMs and their occurrence in D-repairs), which corroborates the fluency-as-frequency hypothesis of this research and the usage-based assumption of the central role of frequency in language.

Taking a step back from the scale of fluency, the results of this chapter also provide some empirical validation of theoretical groupings and categories, following the usage-based principle that structures or expressions that behave in a similar way should be grouped in the same category. Two patterns were confirmed:

– "Potentially Disfluent Functions" grouping *reformulation*, *monitoring* and *punctuation*, which appear to be among the most frequent functions in the editing phase of disfluent repairs;
– the *monitoring-punctuation* pair which is re-coded as one [punctuating] function in two variants of different domains in the revised functional taxonomy by Crible & Degand (in press).

What these results also confirm is the intuition that the notion of reformulation, as defined in formal or contrastive linguistics, is narrower than the notion of repair in the present approach, which also includes issues of structure or linearization, as well as list constructions or other fluent effects. I would also like to suggest that, in a way, repair is narrower than reformulation, in the sense that repair targets "local" discourse moves, not only within the same speaker turn but also in a coherent span of text (resonances are not identified as repairs if many unrelated utterances were produced between the different segments). While more long-distance repairs have been taken into account in other works (e.g. the notion of "diagraph" in Du Bois 2014), the present annotation of DMs also shows that certain functions of DMs (namely *specification* and *enumeration*), although conceptually related to repair, do not always occur in sequences formally marked as repair.

8.7 Interim discussion: Low quantity, high quality?

Two elements of methodological discussion should be addressed before turning to the general conclusion of the book (Chapter 9), namely the qualitative nature of the coding procedure, and the absence of statistical validation. The coding scheme used throughout this chapter is more qualitative than the corpus-based approach adopted so far: even though the functional annotation of DMs is already challenging and arguably subjective, the identification of repair types relies heavily on

a much deeper interpretation of the speaker's motives and intentions, as well as some normative evaluation of the degree of "error" involved in a repaired utterance.

I would like to suggest that the main difference between the method described in Chapter 4 and the procedure detailed in Section 8.2 of this chapter corresponds to the difference between corpus annotation and discourse analysis: while both involve some coding of linguistic phenomena, the relation to the text and to the speaker's intentions is stronger in the latter approach. When annotating the functions of DMs, the researcher does not reconstruct the original message but analyzes the output, in this case the relation between the DM and its context: offline annotation does not equate to online interpretation, and at no point during the analysis is it assumed that function labels are identified and used by the participants of an interaction with the same level of precision. With discourse analysis, on the other hand, the analyst aims at making sense of the observed output with respect to the participants' own reactions and interpretations of the on-going interaction, grounding the analysis in ethnomethodology and sociology.

In the context of coding repair types, the analyst heavily relies upon their world knowledge and experience as a member of a linguistic community. As a social science, linguistics should not shy away from such methods where the analyst is more subjectively involved, provided necessary precautions are taken during the interpretation of the results. Furthermore, the combination of "objective" and systematic corpus methods with more "subjective" approaches to the same data overcomes the limitations of each individual method and provides a richer background for the investigation of the shared object of study, in line with the goal of triangulation and converging evidence promoted by Marchi & Taylor (2009) or researchers in cognitive semantics (Glynn 2010).

A corollary to the qualitative nature of the present analysis is the lack of statistical validation of the results. In corpus linguistics terms and against the current "big data" trend, a sample of 367 occurrences of repair sequences is particularly small, especially in the perspective of finding recurrent patterns of association between variables. With so few data, powerful statistical models become irrelevant since the sample fails to meet the requirements of observed and expected frequencies, running the risk of over-generalizing or, at the other extreme, overlooking potentially interesting – albeit rare – observations. Frequency information remains the basis of my results since I attempt to quantify observed patterns. The analysis, therefore, while not a statistical one, still qualifies as quantitative-qualitative.

Although the scientific value of a study should not be entirely measured by the statistical significance of the results and qualitative studies do present their indubitable advantages, it has become standard practice in the field to evaluate the strength of the observed associations between variables, in any attempt to model (and even predict) specific linguistic behaviors. Therefore, the conclusions

presented in this chapter should be considered tentative and in want of further (statistical, experimental) validation.

I would like to conclude on the richness and flexibility of corpora, which offer complementary methods ranging from purely corpus-driven automatic extraction of statistical patterns to more and more qualitative corpus-based analysis either through (manual) annotation of relatively large amounts of data or as sampled material for more discourse-analytic approaches. I hope that this chapter has illustrated the merits of smaller-scale studies combining quantitatively low samples with qualitatively high interpretations, especially since it provided converging yet independent evidence for some major results from the previous chapters.

CHAPTER 9

Conclusion

9.1 Summary of the main findings

The present usage-based contrastive study of discourse markers and (dis)fluency across registers pursued a three-fold objective: (1) to provide a bottom-up description of the category of DMs in English and French covering their positional, functional and co-occurring behavior (Chapter 5); (2) to situate DMs within the wider typology of fluencemes through paradigmatic analysis of distribution and clustering tendencies (Chapter 6); to uncover fluent and disfluent uses of DMs based on the converging evidence of their linguistic features, their contextual variation and the types of fluenceme sequences in which they occur (Chapters 7 and 8). This first section of the concluding chapter summarizes the main results of this book.

Starting with the contrastive and variationist description of discourse markers, the results tend to show a systematically greater impact of register (e.g. conversation vs. news) and situational features (e.g. prepared vs. non-prepared) over language (i.e. English vs. French) on the distribution and behavior of DMs. The overall frequency of 54 DMs per thousand words was found to decrease from informal registers (conversations, phone calls) to intermediary (interviews) and formal settings (political speeches, news broadcasts). Beyond mere frequency, the specific types (positions, functions) of DMs also vary according to external context. For instance, the four functional domains in the DM taxonomy each favor one type of setting, namely sequential (text-structuring) DMs in spontaneous settings, rhetorical (subjective) DMs in argumentative discourse, ideational (objective) DMs in factual discourse and interpersonal (intersubjective) DMs in interactive dialogues.

Major crosslinguistic differences include, among others, the higher frequency of French utterance-final interpersonal DMs (e.g. *quoi* 'you know', *hein* 'right', *tu vois* 'you see') and the higher frequency of left-integrated ideational DMs in English (e.g. *although*). These differences in quantity and types of DMs favored in each language are counter-balanced by a striking similarity in major form-function patterns as well as the top-five most frequent expressions, viz. *and* / *et* 'and', *but* / *mais* 'but', *so* / *donc* 'so', *well* / *alors* 'so/well', *you know* / *hein* 'right'. Some caveats in the comparability of the corpus prevent us from generalizing these findings beyond the data used in this study.

All in all, the following patterns were identified from the integration of independent variables and through various quantitative (statistical) modeling techniques:

- coordinating conjunctions in pre-field (initial, non-integrated) position marking discourse structure (e.g. *and*, *et*);
- subordinating conjunctions in both left- and right-integrated position signaling discourse relations (e.g. *because, parce que*);
- adverbs in medial position expressing speakers' meta-comments (e.g. *actually, enfin*);
- interjections as independent units serving interactional (speech-segmenting, interpersonal) purposes (e.g. *okay*).

The wide, onomasiological coverage of the DM category in *DisFrEn* allows us to identify coordinating conjunctions (e.g. *and*, *but*) as the most frequent type of DMs, while adverbs (e.g. *so*, *well*) are more representative of the polyfunctionality of the category, with a substantial frequency in all four domains of the taxonomy. In addition, the centrality of a number of formal and functional features, which are often listed as criterial in many definitions of the DM category (namely initiality, structural function and co-occurrence), was confirmed and quantified, thus drawing a corpus-based portrait of DMs while at the same time uncovering their less typical uses.

Turning to the relation between DMs and (dis)fluency, the endeavor to situate DMs within the typology of fluencemes and to uncover patterns where DMs are more or less fluent was partially met within the potential and limitations of corpus-based research to access cognitive, perceptive information. What can be asserted with high confidence from our results is the prominent place of DMs as the second most frequent fluenceme in the corpus after unfilled pauses, with which they frequently cluster. This result is particularly telling of the merits of a broad coverage of (dis)fluent devices including functionally ambivalent elements such as pauses and DMs, as opposed to the bulk of annotation models where such ambivalent elements are highly restricted, if not excluded altogether.

The formal approach to fluenceme identification revealed a number of objective cues to rather disfluent types of sequences, namely mid-size sequences mixing several types of fluencemes (i.e. simple and compound), especially when they occur in registers where they are relatively infrequent (e.g. mixed sequences of substitutions in phone calls). Some registers showed a particular attraction to one sequence type or another, such as interruptions in conversations or identical repetitions in radio interviews, for instance.

The integration of DM-level and sequence-level variables suggested a tentative scale of potentially fluent and potentially disfluent uses of DMs. In particular,

the discourse-structuring function of sequential DMs, added to their tendency to co-occur with pervasive and highly ambivalent fluencemes such as pauses and their frequent occurrence in initial position of hierarchically larger units (i.e. speech turns) all converge in ranking this domain of use as (generally and potentially) fluent. On the other hand, the attraction of interpersonal DMs to the final periphery and to more disruptive fluencemes such as false-starts and truncations suggests a rather disfluent interpretation of this domain.

Such a negative diagnosis was also confirmed for the hypothesized group of "Potentially Disfluent Functions" (viz. *monitoring*, *punctuation* and *reformulation*), which share a strong association to informal, interactive settings and to the aforementioned objective cues of disfluency, a result which was corroborated by the analysis in Chapter 8. However, these patterns were only identified at a very coarse-grained level of analysis and should be viewed as generalizations in want of further validation, especially given the high variability of some uses (e.g. DMs in the rhetorical domain) which remain challenging to situate on the targeted scale of (dis)fluency.

Lastly, the analysis of repairs based on Levelt (1983) revealed that, in the settings of face-to-face interviews, English and French speakers tend to attend primarily to issues of structure (micro- and macro-planning) rather than issues of lexical adequacy. This attention to form over content was argued to be a consequence of the time pressure in unplanned speech. As for the role of DMs in repair, the results tend to suggest a stronger association to covert than overt repair, that is, DMs seem to belong to the (search for a) solution rather than being part of the problem. In other words, DMs maintain the illusion of fluency, except in specific uses where their function stresses the type of ongoing repairing operation (e.g. *monitoring* in structural repairs, *reformulation* in lexical-search repairs, *addition* in resonance repairs).

As a final, general result synthesized from all four empirical chapters, I would like to point to the crucial role of the beginning and ending (i.e. peripheries) of utterances, which are respectively related to planning and monitoring. The initial position was identified as the most frequent slot for DMs and in particular the typical *locus* of fluent clusters of sequential DMs and pauses. Final position, on the other hand, was associated with interpersonal DMs, which are themselves connected to more disfluent contexts of use. I take the cognitive prevalence of these positions or slots in a linguistic unit as further evidence of the time-sensitive dynamics of speech. Speakers make planning decisions either before or right after the beginning of an utterance, then proceed on "auto-pilot" mode once the final plan is decided, and finally look back on the final output to check its adequacy to intentions and rules as well as its appropriate reception by the hearer. This tentative

model is very much in line with the notions of "temporal patterns" in Greene & Capella (1986), "temporal cycles" in Roberts & Kirsner (2000) and Pawley & Syder's (1975) "one-clause-at-a-time" hypothesis. The overwhelming presence of "time" in these works and in the underlying view of language recalls the introductory quote of this book by Carter & McCarthy (2006), which I repeat here for convenience: "Spoken language exists in time, not space" (2006: 193). Yet, I would like to suggest that this final result on the paramount importance of both peripheries (i.e. spatial) and rhythm (i.e. temporal) in spoken discourse and (dis)fluency in fact reconciles time with space, in accordance with the "spacetime continuum" metaphor with which this book started.

9.2 General discussion

The results summarized above raise a number of theoretical and methodological issues. The starting assumption of the present approach to (dis)fluency states that all fluencemes are ambivalent, that is, the same abstract structure (e.g. a pause or DM) can be used and perceived either fluently or disfluently depending on a wide range of linguistic and other factors. Although corpus data can never pretend to cover the full range of possible uses for a given form, the results of this study seem to suggest that some fluencemes are, in fact, less fluent than others as a general rule. In particular, false-starts and truncations were consistently associated with cues of disfluency from multiple independent sources of evidence (e.g. occurrence in mid-size mixed sequences, clustered with DMs expressing "Potentially Disfluent Functions"), as opposed to pauses or discourse markers, whose functional ambivalence was repeatedly illustrated. This does not mean that fluent uses of interruptions do not exist, nor that all cases of interruptions would be perceived as disfluent in context. Nonetheless, robust statistical tendencies clearly suggest significant associations between formal objective cues of fluency and disfluency and specific types of fluencemes.

Another related endeavor aimed at distinguishing fluent from disfluent (uses of) DMs, paying particular attention to their wide range of functions. The analyses from Chapter 7 revealed that, while it is possible to identify potentially disfluent functions of DMs based on the combination of several cues (e.g. rarity in formal registers, co-occurrence with non-ambivalent fluencemes, conceptual relation to disfluency, high frequency in mid-size mixed sequences), the reverse (i.e. identifying potentially fluent functions) is more challenging to carry out on a large scale given the great variability of DMs. This variation is indeed more problematic for fluent DMs since, according to the fluency-as-frequency hypothesis, fluent uses should be very frequent. A higher frequency usually implies a more widespread use

in many different contexts, restricting general interpretations to quite abstract patterns of use. For instance, clusters of sequential DMs and pauses in initial position were identified as a rather high-fluency pattern, yet it would be quite speculative to make such a diagnosis for all its 1,326 instantiations in *DisFrEn*.

Furthermore, high frequency does not necessarily imply widespread use or high fluency. A case in point is *quoi* 'right', which is the sixth most frequent DM in the French data but is highly restricted to conversational registers. This particular expression combines several potentially disfluent features such as its frequent interpersonal function and final position, yet a strict compliance with the fluency-as-frequency hypothesis would suggest a high degree of fluency. Similarly, some high-frequency DMs such as *and* are semantically and pragmatically underspecified, which could result in a greater interpretation cost for the hearer, who is given few cues to disambiguate the intended meaning. It is quite reasonable to imagine that the repeated, pervasive use of *quoi* 'right' or *and* would hinder communicative success and generate negative impressions of disfluency in the hearer's ears. In sum, the high variability, underspecification and resulting lack of recipient design (Mustajoki 2012) of very frequent DMs and schemas constitute limitations to the fluency-as-frequency hypothesis proposed in this study.

More generally, the tools and methods at the corpus linguist's disposal remain limited in their potential to access cognitive or perceptive aspects of language. Beside the shortcomings of a frequentist approach discussed above, the observed patterns remain speaker-based, that is, they only strive to reproduce production mechanisms from the speaker's viewpoint and are utterly silent with respect to the reception of these patterns by hearers. This dependency on observable linguistic features is, therefore, limited to a partial picture of (dis)fluency, which has been amply described as a multi-faceted phenomenon mixing surface features ("productive fluency" in Götz 2013, "utterance fluency" in Segalowitz 2010) with other more holistic measures, as well as individual, even physical and affective factors which remain outside the analyst's control. As Freed (2000: 262) puts it, "the popular notion of fluency includes but is surely far broader than the narrow construct associated with a small cluster of hesitation and repair phenomena". Only a deeply multidisciplinary, multi-method approach to fluency combining corpus data, experimental paradigms, sociolinguistic questionnaires and possibly other tools could substantially broaden our understanding of what makes speech fluent or disfluent – and maybe not even then, especially considering the challenge of inter-operability in making these different approaches communicate.

While the present corpus-based study can only provide a partial picture of (dis)fluency in general and of the (dis)fluency of DMs in particular, it is, however, far-reaching in terms of the description of the DM category. The present endeavor to aim at an exhaustive portrait of DMs, as opposed to the majority of case studies

in the field, motivates the resort to corpus-based analysis, since only corpora can provide such a broad coverage of complex linguistic categories, provided they are thoroughly explored through bottom-up and informative annotation procedures. What this extensive-intensive approach to DMs further reveals is that DMs fulfil many different functions, only a handful of which bear a direct connection to fluency.

9.3 Implications and research avenues

Although this study of native (dis)fluency is more descriptive than applied, its methodological and empirical contributions have a number of implications for the fields of discourse markers and fluency research as well as for more concrete applications beyond academia. First, *DisFrEn* is, to my knowledge and to date, the only dataset of any spoken language to be fully annotated for DMs, their position and function beyond the restrictions discussed in the literature review, thus adding to "the small class of corpora featuring discourse and pragmatic annotation" (Rühlemann & O'Donnell 2012: 315). As such, the annotations can be queried for any type of research question involving the linguistic variables covered by the coding scheme beyond the questions already investigated in the present work.

The functional taxonomy specifically designed for *DisFrEn* has already been applied to other spoken languages (Kinshasa Lingalá by Nzoimbengene 2016; Slovene by Dobrovoljc 2016) as well as writing (Crible & Zufferey 2015), gestures (Bolly 2015) and Belgian French Sign Language (Gabarró-López forthc.), by both expert and naïve coders (Crible & Degand in press). Were the annotations in these different corpora sufficiently reliable and comparable, they would constitute a very rich resource for crosslinguistic discourse analysis. Future contrastive research might make use of comparable annotations and uncover language-specific vs. universal types and uses of discourse markers and their clustering with (dis)fluent devices (see Pascual & Crible 2017 for a comparison with Spanish).

In addition, it would be highly relevant to extend the present method and analysis to multimodal data, either in the form of gesture analysis (cf. the work by Bolly and colleagues, e.g. Gerstenberg & Bolly 2015) or in computer-mediated interfaces involving both speech and writing at the same time, as in videogame communication (Collister 2013). Comparison with written data alone, although restricted to a common core of relational discourse markers, also constitutes a fruitful avenue (e.g. Ciabarri 2013; Fox Tree 2014; Lapshinova-Koltunski et al. 2015) which could benefit from the large-scale and bottom-up coverage of the DM category and their functions as proposed here. *DisFrEn* could also be used as a reference corpus or

basis for comparison with more specific data types such as business English or French, pathological language or human-machine communication.

Enhancing the amount of annotated data could be particularly useful to computational applications making use of the observed patterns of DMs (e.g. part-of-speech tag, syntactic position) as reliable cues in the perspective of automatic sense disambiguation or machine translation (cf. the works of Popescu-Belis and colleagues, e.g. Meyer et al. 2012, Popescu-Belis et al. 2012), an endeavor which is still in its infancy in written data, let alone in speech. The wide coverage of fluencemes in *DisFrEn* also provides natural language processing approaches with training data for automatic disfluency detection, including ambivalent structures such as modified repetitions.

Another obvious area which could benefit from the contributions of this work is second-language studies and learner corpus research, where the study of discourse markers or connectives is already a strong area of interest. This trend of investigation is represented by, e.g., Granger & Petch-Tyson (1996), Müller (2005), Denke (2009) or Gilquin (2016). Like most DM research in native language, these L2 studies either focus on connectives (or subtypes thereof), especially in written data, or on a selection of spoken DMs, usually without deeper levels of analysis (such as information on position or meaning-in-context), which can be explained by the already complex task of working with non-native data. The unique features of learner language probably exclude any direct application of the coding scheme used in the present corpus of native speech, yet the functional categories should, in principle, exist in English or French as a foreign language as well and could definitely serve as a basis for a revised model to be used in future research. In any case, the crosslinguistic portrait of the variation and combination of DMs with (dis)fluency devices provides a basis for quantitative and qualitative comparison with any L2 and other corpus looking into the complex mechanisms of spoken interaction.

Other promising research avenues can address the limitations of this research to further the validity and theoretical reach of the results, such as the need to include sociolinguistic metadata to check for any effect of age, gender or socio-economic background on the distribution of DMs and fluencemes, the addition of prosodic analysis beyond the mere identification of filled and unfilled pauses to refine the patterns and local contexts of DM use, or the combination with other methods, for instance experimental paradigms, to shed complementary light on the corpus-based patterns presently identified.

Overall, I hope that this research has somehow enhanced our understanding of discourse markers and fluencemes, these complex categories which are so frequent and necessary to any type of formal and casual language and yet still escape comprehensive modeling.

Bibliography

Abeillé, Anne, Lionel Clément, and François Toussenel. 2003. "Building a Treebank for French." In *Treebanks: Building and Using Parsed Copora*, ed. by Anne Abeillé, 165–188. Dordrecht: Kluwer Academic Publishers. doi:10.1007/978-94-010-0201-1_10

Aijmer, Karin, and Anne-Marie Simon-Vandenbergen (eds). 2006. *Pragmatic Markers in Contrast*. Amsterdam: Elsevier.

Aijmer, Karin, and Anne-Marie Simon-Vandenbergen. 2011. "Pragmatic Markers." In *Discursive Pragmatics*, ed. by Jan Zienkowski, Jan-Ola Östmann and Jef Verschueren, 223–247. Amsterdam: John Benjamins. doi:10.1075/hoph.8.13aij

Altenberg, Bengt. 2006. "The Function of Adverbial Connectors in Second Initial Position in English and Swedish." In *Pragmatic Markers in Contrast*, ed. by Karin Aijmer, and Anne-Marie Simon-Vandenbergen, 11–37. Oxford: Elsevier.

Andersen, Gisle. 1997. "They Like Wanna See Like How We Talk and All That. The Use of Like as a Discourse Marker in London Teenage Speech." *Corpus-Based Studies in English*, ed. by Magnus Ljung, 37–48. Amsterdam: Rodopi.

Anscombre, Jean-Claude, and Oswald Ducrot. 1983. *L'Argumentation dans la Langue*. Liège-Bruxelles: Mardaga.

Arnold, Jennifer E., Maria Fagnano, and Michael K. Tanenhaus. 2003. "Disfluencies Signal theee, um, New Information." *Journal of Psycholinguistic Research* 32 (1): 25–36. doi:10.1023/A:1021980931292

Arnold, Jennifer E., and Michael K. Tanenhaus. 2011. "Disfluency Effects in Comprehension: How New Information Can Become Accessible." In *The Processing and Acquisition of Reference*, ed. by Edward A. Gibson, and Neal J. Perlmutter, 197–217. Cambridge, MA: MIT Press. doi:10.7551/mitpress/9780262015127.003.0008

Asher, Nicholas, and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge: Cambridge University Press.

Auer, Peter. 1996. "The Pre-Front Field in Spoken German and its Relevance as a Grammatical Position." In *Pragmatics* 6 (3): 223–259. doi:10.1075/prag.6.3.03aue

Auer, Peter. 2005. "Delayed Self-Repairs as a Structuring Device for Complex Turns in Conversation." In *Syntax and Lexis in Conversation: Studies on the Use of Linguistic Resources in Talk-in-interaction*, ed. by Auli Hakulinen, and Margret Selting, 75–102. Amsterdam: John Benjamins. doi:10.1075/sidag.17.06aue

Auer, Peter. 2009. "On-Line Syntax: Thoughts on the Temporality of Spoken Language." *Language Sciences* 31 (1): 1–13. doi:10.1016/j.langsci.2007.10.004

Auer, Peter, and Stefan Pfänder. 2007. "Multiple Retractions in Spoken French and Spoken German. A Contrastive Study in Oral Performance Styles." *Cahiers de Praxématique* 48: 57–84.

Balthasar, Lukas, and Michel Bert. 2005. "La base de données 'Corpus de langues parlées en interaction' (CLAPI): Genèse, état des lieux et perspectives [The database 'Corpus of spoken languages in interaction': Genesis, current state and perspectives]." *Lidil* 31.

Barr, Dale J., and Mandana Seyfeddinipur. 2010. "The Role of Fillers in Listener Attributions for Speaker Disfluency." *Language and Cognitive Processes* 25 (4): 441–455. doi:10.1080/01690960903047122

Barth, Danielle, and Vsevolod Kapatsinski. 2018. "Evaluating Logistic Mixed-Effects Models of Corpus Data." In *Mixed-Effects Regression Models in Linguistics*, ed. by Dirk Speelman, Kris Heylen, and Dirk Geeraerts. Berlin: Springer.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2014. "lme4: Linear Mixed-Effects Models Using Eigen and S4." R package version 1.0–6. http://CRAN.R-project.org/package=lme4.

Beeching, Kate. 2007. "La co-variation des marqueurs discursifs *bon*, *c'est-à-dire*, *enfin*, *hein*, *quand même*, *quoi* et *si vous voulez*: Une question d'identité? [The co-variation of discourse markers *bon*, *c'est-à-dire*, *enfin*, *hein*, *quand même*, *quoi* and *si vous voulez*: A question of identity?]." *Langue Française* 154 (2): 78–93.

Beeching, Kate. 2016. *Pragmatic Markers in British English. Meaning in Social Interaction*. Cambridge: Cambridge University Press.

Beeching, Kate. 2017. "Just a Suggestion: *just/e* in French and English." In *Discourse Markers, Pragmatics Markers and Modal Particles: New Perspectives*, ed. by Chiara Fedriani, and Andrea Sanso, 465–487. Amsterdam: John Benjamins. doi:10.1075/slcs.186.18bee

Beeching, Kate, and Ulrich Detges (eds). 2014. *The Role of the Left and Right Periphery in Semantic Change: Crosslinguistic Investigations of Language and Language Change*. Leiden: Brill.

Beliao, Julie, and Anne Lacheret. 2013. "Disfluency and Discursive Markers: When Prosody and Syntax Plan Discourse." In *Proceedings of Disfluency in Spontaneous Speech (DiSS)*, ed. by Robert Eklund: 5–8.

Bertrand, Roxanne, Philippe Blâche, Robert Espesser, Gaëlle Ferré, Christine Meunier, Béatrice Priego-Valverde, and Stéphane Rauzy. 2008. "Le CID – Corpus of Interactional Data – Annotation et exploitation multimodale de parole conversationnelle [The CID – Corpus of Interactional Data – Annotation and multimodal exploitation of conversation speech]." *Traitement Automatique des Langues* 49 (3).

Besser, Jana, and Jan Alexandersson. 2007. "A Comprehensive Disfluency Model for Multi-Party Interaction." In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, ed. by Simon Keizer, Harry Bunt, and Tim Paek: 182–189.

Biber, Douglas, Susan Conrad, and Randi Reppen. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Philadelphia: John Benjamins.

Blakemore, Diane. 1993. "The Relevance of Reformulations." *Language and Literature* 2 (2): 101–120.

Blakemore, Diane. 2002. *Relevance and Linguistic Meaning. The Semantics and Pragmatics of Discourse Markers*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511486456

Blanche-Benveniste, Claire. 2003. "Le recouvrement de la syntaxe et de la macro-syntaxe [The mapping of syntax and macro-syntax]." In *Macro-syntaxe et Pragmatique. L'Analyse Linguistique de l'Oral*, ed. by Antonietta Scarano, 53–75. Rome: Bulzoni.

Blanche-Benveniste, Claire, Mireille Bilger, Christine Rouget, Karel Van Den Eynde, and Piet Mertens. 1990. *Le Français Parlé. Etudes Grammaticales*. Paris: CNRS.

BNC Consortium. 2007. "The British National Corpus, Version 3 (BNC XML Edition)." Distributed by Oxford University Computing Services. http://www.natcorp.ox.ac.uk/.

Bolly, Catherine. 2015. "Towards Pragmatic Gestures: From Repetition to Construction in Multimodal Pragmatics." Paper presented at the 13th International Cognitive Linguistics Conference (ICLC-13), July 20–25, Newcastle, UK.

Bolly, Catherine, and Ludivine Crible. 2015. "From Context to Functions and Back Again: Disambiguating Pragmatic Uses of Discourse Markers." Paper presented at the International Pragmatics Association (IPrA) Conference, July 26–31, Antwerp, Belgium.

Bolly, Catherine, Ludivine Crible, Liesbeth Degand, and Deniz Uygur-Distexhe. 2015. "MDMA. Identification et annotation des marqueurs discursifs 'potentiels' en contexte. [MDMA. Identification and annotation of 'potential' discourse markers in context]." *Discours* 15.

Bolly, Catherine, Ludivine Crible, Liesbeth Degand, and Deniz Uygur-Distexhe. 2017. "Towards a Model for Discourse Marker Annotation in Spoken French: From Potential to Feature-Based Discourse Markers." In *Discourse Markers, Pragmatics Markers and Modal Particles: New Perspectives*, ed. by Chiara Fedriani, and Andrea Sanso, 73–99. Amsterdam: John Benjamins. doi:10.1075/slcs.186.03bol

Bolly, Catherine, and Anaïs Thomas. 2015. "Facing Nadine's Speech. Multimodal Annotation of Emotion in Later Life." *Linköping Electronic Conference Proceedings* 110: 23–32.

Bortfeld, Heather, Silvia D. Leon, Jonathan E. Bloom, Michael F. Schober, and Susan E. Brennan. 2001. "Disfluency Rates in Conversation: Effects of Age, Relationship, Topic, Role and Gender." *Language and Speech* 44: 123–147. doi:10.1177/00238309010440020101

Bosker, Hans Rutger, Hugo Quené, Ted J. M. Sanders, and Nivja H. 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. doi:10.1016/j.jml.2014.05.004

Boula de Mareüil, Philippe, Gilles Adda, Martine Adda-Decker, Claude Barras, Benoît Habert, and Patrick Paroubek. 2013. "Une étude quantitative des marqueurs discursifs, disfluences et chevauchements de parole dans des interviews politiques [A quantitative study of discourse markers, disfluencies and overlaps in political interviews]." *TIPA Travaux Interdisciplinaires sur la Parole et le Langage* 29.

Boula De Mareüil, Philippe, Benoît Habert, Frédérique Bénard, Martine Adda-Decker, Claude Barras, Gilles Adda, and Patrick Paroubek. 2005. "A Quantitative Study of Disfluencies in French Broadcast Interviews." In *Proceedings of Disfluency In Spontaneous Speech (DISS) Workshop*, 10–12 September 2005, Aix-en-Provence, France: 27–32.

Brédart, Serge. 1991. "Word Interruption in Self-Repairing." *Journal of Psycholinguistic Research* 20 (2): 123–138.

Brennan, Susan E., and Michael F. Schober. 2001. "How Listeners Compensate for Disfluencies in Spontaneous Speech." *Journal of Memory and Language* 44: 274–296. doi:10.1006/jmla.2000.2753

Brinton, Laurel. 1996. *Pragmatic Markers in English. Grammaticalization and Discourse Functions*. New York: Mouton de Gruyter. doi:10.1515/9783110907582

Briz, Antonio, and Salvador Pons Bordería. 2010. "Unidades, marcadores discursivos y posición [Units, discourse markers and position]." In *Los Estudios sobre Marcadores del Discurso*, ed. by Oscar Loureda, and Esparanza Acin, 523–557. Madrid: Acro/Libros.

Briz, Antonio, and Val.Es.Co Group. 2003. "Un sistema de unidades para el estudio del lenguaje coloquial [A system of units for the study of colloquial language]." *Oralia* 6: 7–61.

Broen, Patricia A., and Gerald M. Siegel. 1972. "Variations in Normal Speech Disfluencies." *Language and Speech* 15: 219–231.

Brown, Penelope, and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage*. Cambridge: Cambridge University Press.

Brumfit, Christopher J. 1984. *Communicative Methodology in Language Teaching*. Cambridge: Cambridge University Press.

Bybee, Joan. 1985. *Morphology: A Study on the Relation Between Meaning and Form*. Amsterdam: John Benjamins. doi:10.1075/tsl.9

Bybee, Joan. 2006. "From Usage to Grammar: The Mind's Response to Repetition." *Language* 82 (4): 711–733. doi:10.1353/lan.2006.0186

Candéa, Maria. 2000. *Contribution à l'étude des pauses silencieuses et des phénomènes dits "d'hésitation" en français oral spontané* [*Contribution to the study of silent pauses and so-called "hesitation" phenomena in spontaneous spoken French*]. PhD thesis, Université Paris III.

Carter, Ronald, and Michael McCarthy. 2006. *Cambridge Grammar of English*. Cambridge: Cambridge University Press.

Carlson, Lynn, Daniel Marcu, and Mary Ellen Okurowski. 2002. "RST Discourse Treebank [Corpus]." *Linguistic Data Consortium*, Phildelphia, PA.

Castellà, Josep Maria. 2004. *Oralitat i escriptura. Dues cares de la complexitat del llenguatge* [*Speech and writing. Two faces of the complexity of language*]. Barcelona: Publicacions de l'Abadia de Montserrat.

Chafe, Wallace. 1992. "The Importance of Corpus Linguistics to Understanding the Nature of Language." In *Directions in Corpus Linguistics*, ed. by Jan Svartvik, 79–97. Berlin: Mouton de Gruyter. doi:10.1515/9783110867275.79

Chambers, Francine. 1997. "What Do We Mean by Fluency?" *System* 25 (4): 535–544. doi:10.1016/S0346-251X(97)00046-8

Chanet, Catherine. 2001. "1700 occurrences de la particule *quoi* en français parlé contemporain: Approche de la 'distribution' et des fonctions en discours [1700 occurrences of the particle *quoi* in spoken contemporary French: Study of the 'distribution' and functions in discourse]." *Marges Linguistiques* 2: 56–80.

Charolles, Michel, and Danièle Coltier. 1986. "Le contrôle de la compréhension dans une activité rédactionnelle: Eléments pour l'analyse des reformulations paraphrastiques. [Comprehension control in a writing activity: Elements for the analysis of paraphrastic reformulations]." *Pratiques* 49: 51–66.

Ciabarri, Federica. 2013. "Italian Reformulation Markers: A Study on Spoken and Written Language." In *Across the Line of Speech and Writing Variation*, ed. by Catherine Bolly, and Liesbeth Degand, 113–128. Louvain-la-Neuve, Presses universitaires de Louvain.

Clark, Herbert H., and Jean E. Fox Tree. 2002. "Using *Uh* and *Um* in Spontaneous Speaking." *Cognition* 84: 73–111. doi:10.1016/S0010-0277(02)00017-3

Clark, Herbert H., and Thomas Wasow. 1998. "Repeating Words in Spontaneous Speech." *Cognitive Psychology* 37: 201–242. doi:10.1006/cogp.1998.0693

Colletta, Jean-Marc, Ramona N. Kunene, Aurélie Venouil, Virginie Kaufmann, and Jean-Pascal Simon. 2009. "Multi-Track Annotation of Child Language and Gestures." In *Multimodal corpora: From Models of Natural Interaction to Systems and Applications*, ed. by Michael Kipp, Jean-Claude Martin, Patrizia Paggio, and Dirk Heylen, 54–72. Berlin: Springer. doi:10.1007/978-3-642-04793-0_4

Collister, Lauren. 2013. *Multimodality as a Sociolinguistic Resource*. PhD thesis, University of Pittsburgh.

Connor, Ulla M., and Ana I. Moreno. 2005. "Tertium Comparationis: A Vital Component in Contrastive Research Methodology." In *Directions in Applied Linguistics: Essays in Honor of Robert B. Kaplan*, ed. by Paul Bruthiaux, Dwight Atkinson, William Eggington, William Grabe, and Vaidehi Ramanathan, 153–164. England: Multilingual Matters.

Corley, Martin. 2010. "Making Predictions from Speech with Repairs: Evidence from Eye Movements." *Language and Cognitive Processes* 25 (5): 706–727. doi:10.1080/01690960903512489

Corley, Martin, Lucy MacGregor, and David Donaldson. 2007. "It's the Way That You, er, Say It: Hesitations in Speech Affect Language Comprehension." *Cognition* 105: 658–668. doi: 10.1016/j.cognition.2006.10.010

Crible, Ludivine. 2014. "Identifying and Describing Discourse Markers in Spoken Corpora. Annotation Protocol v.8." *Technical report*, Université catholique de Louvain.

Crible, Ludivine. 2015. "Grammaticalisation du marqueur discursif complexe *ou sinon* dans le corpus de SMS belge: Spécificités sémantiques, graphiques et diatopiques. [Grammaticalization of the complex discourse marker *ou sinon* in the Belgian text-message corpus: Semantic, graphic and diatopic features]." *Le Discours et la Langue* 7 (1): 181–200.

Crible, Ludivine. 2017a. "Towards an operational category of discourse markers: A definition and its model." In *Discourse Markers, Pragmatics Markers and Modal Particles: New Perspectives*, ed. by Chiara Fedriani, and Andrea Sanso, 101–126. Amsterdam: John Benjamins. doi: 10.1075/slcs.186.04cri

Crible, Ludivine. 2017b. *Discourse Markers and (Dis)fluency across Registers: A Contrastive Usage-Based Study in English and French*. PhD thesis, Université catholique de Louvain.

Crible, Ludivine. 2017c. "Discourse markers and (dis)fluencies in English and French: Variation and combination in the DisFrEn corpus." *International Journal of Corpus Linguistics* 22 (2): 242–269.

Crible, Ludivine, and Maria Josep Cuenca. 2017. "Discourse Markers in Speech: Characteristics and Challenges for Corpus Annotation". *Dialogue and Discourse* 8 (2): 149–166.

Crible, Ludivine, and Liesbeth Degand. In press. "Reliability vs. Granularity in Discourse Annotation: What is the Trade-Off?" *Corpus Linguistics and Linguistic Theory*.

Crible, Ludivine, Liesbeth Degand, and Gaëtanelle Gilquin. 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. doi: 10.1075/lic.17.1.04cri

Crible, Ludivine, Amandine Dumont, Iulia Grosman, and Ingrid Notarrigo. 2016. "Annotation Manual of Fluency and Disfluency Markers in Multilingual, Multimodal, Native and Learner Corpora. Version 2.0." *Technical report*, Université catholique de Louvain and Université de Namur.

Crible, Ludivine, Amandine Dumont, Iulia Grosman, and Ingrid Notarrigo. Forthcoming. "(Dis)fluency across Spoken and Signed Languages: Applications of an Interoperable Annotation Scheme."

Crible, Ludivine, and Sandrine Zufferey. 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), IWCS 2015 Workshop*, ed. by Harry Bunt, 14–22.

Crystal, David. 1987. *The Cambridge Encyclopedia of Language*. Cambridge: Cambridge University Press.

Crystal, David. 1988. "Another Look at, *Well*, *You Know*…" *English Today* 4 (1): 47–49. doi: 10.1017/S0266078400003321

Cuenca, Maria Josep. 2003. "Two Ways to Reformulate: A Contrastive Analysis of Reformulation Markers." *Journal of Pragmatics* 35: 1069–1093. doi: 10.1016/S0378-2166(03)00004-3

Cuenca, Maria Josep. 2013. "The Fuzzy Boundaries Between Discourse Marking and Modal Marking." In *Discourse Markers and Modal Particles. Categorization and Description*, ed. by Liesbeth Degand, Bert Cornillie, and Paola Pietrandrea, 191–216. Amsterdam: John Benjamins. doi: 10.1075/pbns.234.08cue

Cuenca, Maria Josep, and Carme Bach. 2007. "Contrasting the Form and Use of Reformulation Markers." *Discourse Studies* 9 (2): 149–175. doi: 10.1177/1461445607075347

Cuenca, Maria Josep, and Maria Josep Marín. 2009. "Co-Occurrence of Discourse Markers in Catalan and Spanish Oral Narrative." *Journal of Pragmatics* 41: 899–914. doi: 10.1016/j.pragma.2008.08.010

Cutting, Joan. 2008. *Pragmatics and Discourse, 2nd edition*. New York: Routledge.

Danks, Joseph, and Laurel J. End. 1987. "Processing Strategies for Reading and Listening." In *Comprehending Oral and Written Language*, ed. by Rosalind Horowitz, and S. Jay Samuels, 271–294. San Diego: Academic Press.

Danlos, Laurence, Margot Colinet, and Jacques Steinlin. 2015. "FDTB1, première étape du projet 'French Discourse Treebank': repérage des connecteurs de discours en corpus [FDTB1, first step of the project 'French Discourse Treebank': identification of discourse connectives in corpus]." *Discours* 17 [online].

De Cock, Sylvie. 2000. "Repetitive Phrasal Chunkiness and Advanced EFL Speech and Writing." In *Corpus Linguistics and Linguistic Theory. Papers from the Twentieth International Conference on English Language Research on Computerized Corpora (ICAME 20)*, ed. by Christian Mair, and Marianne Hundt, 51–68. Amsterdam: Rodopi.

De Gaulmyn, Marie-Madeline. 1987. "Actes de reformulation et processus de reformulation [Reformulative acts and reformulation process]." In *L'analyse des interactions verbales. La Dame de Caluire: Une consultation*, ed. by Pierre Bange, 83–98. Bern: Peter Lang.

Defour, Tine, Ulrique D'Hondt, Anne-Marie Vandenbergen, and Dominique Willems. 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, Liesbeth. 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*, ed. by Kate Beeching, and Ulrich Detges, 151–178. Leiden: Brill.

Degand, Liesbeth, Laurence J. Martin, and Anne-Catherine Simon. 2014. "Unités discursives de base et leur périphérie gauche dans LOCAS-F, un corpus oral multigenres annoté [Basic discourse units and their left periphery in LOCAS-F, a spoken multigenre annotated corpus]." In *Proceedings of CMLF 2014–4ème Congrès Mondial de Linguistique Française 2014*, Berlin, Germany: EDP Sciences.

Degand, Liesbeth, and Anne-Catherine Simon. 2009. "On Identifying Basic Discourse Units in Speech: Theoretical and Empirical Issues." *Discours* 4.

Degand, Liesbeth, and Anne-Marie Simon-Vandenbergen. 2011. "Grammaticalization and (Inter) subjectification of Discourse Markers." *Linguistics* 49: 287–294. doi: 10.1515/ling.2011.008

Demirşahin, Işın, and Deniz Zeyrek. 2014. "Annotating Discourse Connectives in Spoken Turkish." In *Proceedings of LAW VIII – The 8th Linguistic Annotation Workshop*: 105–109. doi: 10.3115/v1/W14-4916

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

Deppermann, Arnulf, and Susanne Günthner. 2015. *Temporality in Interaction*. Amsterdam: John Benjamins.

Diessel, Holger, and Martin Hilpert. 2016. "Frequency Effects in Grammar." In *Oxford Research Encyclopedia of Linguistics*, ed. by Mark Aronoff. New York: Oxford University Press.

Dister, Anne. 2007. *De la transcription à l'étiquetage morphosyntaxique – Le cas de la banque de données textuelles orales VALIBEL* [*From transcription to morphosyntactic tagging – The case of the spoken text databank VALIBEL*]. PhD thesis, Université catholique de Louvain.

Dister, Anne, Michel Francard, Philippe Hambye, and Anne-Catherine Simon. 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) [From corpus to databank. Sound, texts and metadata. The evolution of the spoken text databank VALIBEL (1989–2009)." *Cahiers de Linguistique* 33 (2), 113–129.

Dobrovoljc, Kaja. 2016. "Annotation of Multi-Word Discourse Markers in Spoken Slovene." Poster presented at Discourse Relational Devices Conference (LPTS2016), Linguistic & Psycholinguistic Approaches to Text Structuring, January 24–26, Valencia, Spain.

Dostie, Gaëtane. 2004. *Pragmaticalisation et marqueurs discursifs. Analyse sémantique et traitement lexicographique* [*Pragmaticalization and discourse markers. Semantic analysis and lexicographic treatment*]. Bruxelles: De Boeck.

Dostie, Gaëtane. 2013. "Les associations de marqueurs discursifs – De la cooccurrence libre à la collocation [Associations of discourse markers – From free co-occurrence to collocation]." *Linguistik Online* 62 (5).

Du Bois, John. 2014. "Towards a Dialogic Syntax." *Cognitive Linguistics* 25 (3): 359–410. doi: 10.1515/cog-2014-0024

Du Bois, John, Wallace Chafe, Charles Meyer, Sandra A. Thompson, Robert Englebretson, and Nii Martey. 2000–2005. *Santa Barbara Corpus of Spoken American English, Parts 1–4*. Philadelphia: Linguistic Data Consortium.

Duez, Danielle. 1991. *La pause dans la parole de l'homme politique* [*Pauses in the speech of politicians*]. Paris: Editions du CNRS.

Dupont, Maïté. 2015. "Word Order in English and French. The Position of English and French Adverbial Connectors of Contrast." *English Text Construction* 8 (1): 88–124. doi: 10.1075/etc.8.1.04dup

Durand, Jacques, Bernard Laks, and Chantal Lyche. 2002. "La phonologie du français contemporain: Usages, variétés et structure [Phonology of contemporary French: Uses, varieties and structure]." In *Romanistische Korpuslinguistik – Korpora und gesprochene Sprache / Romance Corpus Linguistics – Corpora and Spoken Language*, ed. by Claus D. Pusch, and Wolfgand Raible, 93–106. Tübingen: Gunter Narr Verlag.

Durand, Jacques, Bernard Laks, and Chantal Lyche. 2009. "Le projet PFC: Une source de données primaires structurées [The PFC project: A resource for structured primary data]." In *Phonologie, variation et accents du français*, ed. by Jacques Durand, Bernard Laks, and Chantal Lyche, 19–61. Paris: Hermès.

Dutrey, Camille, Sophie Rosset, Martine Adda-Decker, Chloé Clavel, and Ioana Vasilescu. 2014. "Disfluences dans la parole spontanée conversationnelle: Détection automatique utilisant des indices lexicaux et acoustiques [Disfluencies in spontaneous conversational speech: Automatic detection using lexical and acoustic cues]." In *Proceedings of the XXXe Journées d'Etude sur la Parole (JEP'14)*: 366–373.

Ejzenberg, Roseli. 2000. "The Juggling Act of Oral Fluency: A Psycho-Sociolinguistic Metaphor." In *Perspectives on Fluency*, ed. by Heidi Riggenbach, 288–313. Ann Arbor: The University of Michigan Press.

Eklund, Robert. 2004. *Disfluency in Swedish Human-human and Human-machine Travel Booking Dialogues*. PhD thesis, Linköping Studies in Science and Technology.

Eklund, Robert, and Elizabeth Shriberg. 1998. "Crosslinguistic Disfluency Modeling: A Comparative Analysis of Swedish and American English Human-Human and Human-Machine Dialogs." In *Proceedings of the 5th International Conference on Spoken Language Processing*.

Erard, Michael. 2004. "Just Like, Er, Words, Not, Um, Throwaways." *The New York Times*, 2 January 2004: A 13 & A 15.

Erman, Britt. 2001. "Pragmatic Markers Revisited with a Focus on You Know in Adult and Adolescent Talk." *Journal of Pragmatics* 33: 1337–1359. doi:10.1016/S0378-2166(00)00066-7

Esser, Jürgen. 1993. *English Linguistic Stylistics*. Tübingen: Niemeyer.

Estellés Arguedas, María, and Salvador Pons Bordería. 2014. "Absolute Initial Position." In *Discourse Segmentation in Romance Languages*, ed. by Salvador Pons Bordería, 121–155. Amsterdam: John Benjamins.

Fernández Polo, Fransisco Javier. 1999. *Traducción y retórica contrastiva. A propósito de la traducción de textos de divulgación científica del inglés al español* [*Translation and contrastive rhetoric. On the translation of texts for scientific dissemination from English to Spanish*]. Santiago de Compostela: Universidade de Santiago de Compostela. Anexo de Moenia 6.

Fiksdal, Susan. 2000. "Fluency as a Function of Time and Rapport." In *Perspectives on Fluency*, ed. by Heidi Riggenbach, 128–140. Ann Arbor: The University of Michigan Press.

Fillmore, Charles. 1979. "On Fluency." In *Individual Differences in Language Ability and Language Behavior*, ed. by Charles Fillmore, Daniel Kempler, and William S-Y. Wang, 85–102. New York: Academic Press.

Fillmore, Charles. 2000. "On Fluency." In *Perspectives on Fluency*, ed. by Heidi Riggenbach, 43–60. Ann Arbor: The University of Michigan Press.

Fischer, Kerstin. 2006. "Towards an Understanding of the Spectrum of Approaches to Discourse Particles: Introduction to the Volume." In *Approaches to discourse particles*, ed. by Kerstin Fischer, 1–20. Amsterdam: Elsevier.

Fischer, Kerstin. 2014. "Discourse Markers." In *Pragmatics of Discourse*, ed. by Klaus Schneider, and Anne Barron, 271–294. Berlin: De Gruyter. doi:10.1515/9783110214406-011

Fleischman, Suzanne, and Marina Yaguello. 2004. "Discourse Markers across Languages. Evidence from English and French." In *Discourse across Languages and Cultures*, ed. by Carol Lynn Moder, and Aida Martinovic-Zic, 129–148. Philadelphia: John Benjamins. doi:10.1075/slcs.68.08fle

Flower, Linda, and John R. Hayes. 1981. "A Cognitive Process Theory of Writing." *College Composition and Communication* 32 (4): 365–87. doi:10.2307/356600

Foster, Pauline, Alan Tonkyn, and Gillian Wigglesworth. 2000. "Measuring Spoken Language: A Unit for all Reasons." *Applied Linguistics* 21 (3): 354–375. doi:10.1093/applin/21.3.354

Fox, Barbara A., Makoto Hayashi, and Robert Jasperson. 1996. "Resources and Repair: A Cross-Linguistic Study of Syntax and Repair." In *Interaction and Grammar*, ed. by Elinor Ochs, Emanuel A. Schegloff, and Sandra A. Thompson, 185–237. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511620874.004

Fox Tree, Jean E. 2001. "Listeners' Uses of *Um* and *Uh* in Speech Comprehension." *Memory and Cognition* 29 (2): 320–326. doi:10.3758/BF03194926

Fox Tree, Jean E. 2014. "Discourse Markers in Writing." *Discourse Studies* 17 (1): 64–82. doi:10.1177/1461445614557758

Fraser, Bruce. 1996. "Pragmatic markers." *Pragmatics* 6 (2): 167–190. doi:10.1075/prag.6.2.03fra

Fraser, Bruce. 1999. "What Are Discourse Markers?" *Journal of Pragmatics* 31: 931–952. doi:10.1016/S0378-2166(98)00101-5

Freed, Barbara F. 2000. "Is Fluency, like Beauty, in the Eyes (and Ears) of the Beholder?" In *Perspectives on Fluency*, ed. by Heidi Riggenbach, 243–265. Ann Arbor: The University of Michigan Press.

Gabarró-López, Sílvia. Forthcoming. "Marqueurs du discours en langue des signes de Belgique francophone (LSFB) et langue des signes catalane (LSC): Les 'balise-listes' et les 'palm-ups' [Discourse markers in Belgian French sign language (LSFB) and Catalan sign language (LSC): 'buoys' and 'palm-ups']." In *Marcadores del discurso y lingüística contrastiva en las lenguas románicas*, ed. by Oscar Loureda, Guillermo Álvarez Sellán, and Martha Rudka. Madrid: Iberoamericana Vervuert.

Geluykens, Ronald. 1994. *The Pragmatics of Discourse Anaphora in English. Evidence from Conversational Repair*. Berlin: Mouton de Gruyter. doi:10.1515/9783110846171

Gerstenberg, Annette, and Catherine Bolly. 2015. "Functions of Repetition in the Discourse of Elderly Speakers: The Role of Prosody and Gesture." Paper presented at the 14th International Pragmatics Conference (IPrA), July 26–31, Antwerp, Belgium.

Gilquin, Gaëtanelle. 2016. "Discourse Markers in L2 English. From Classroom to Naturalistic Input." In *New Approaches to English Linguistics: Building Bridges*, ed. by Olga Timofeeva, Anne-Christine Gardner, Alpo Honkapohja, and Sarah Chevalier, 213–249. Amsterdam: John Benjamins. doi:10.1075/slcs.177.09gil

Gilquin, Gaëtanelle, and Sylvie De Cock. 2011. "Errors and Disfluencies in Spoken Corpora. Setting the Scene." *International Journal of Corpus Linguistics* 16 (2): 141–172. doi:10.1075/ijcl.16.2.01gil

Ginzburg, Jonathan, Raquel Fernandez, and David Schlangen. 2014. "Disfluencies as Intra-Utterance Dialogue Moves." *Semantics & Pragmatics* 7: 1–64. doi:10.3765/sp.7.9

Glynn, Dylan. 2010. "Corpus-Driven Cognitive Semantics. Introduction to the Field." In *Quantitative Methods in Cognitive Semantics: Corpus-Driven Approaches*, ed. by Dylan Glynn, and Kerstin Fischer, 1–41. Berlin: De Gruyter Mouton. doi:10.1515/9783110226423.1

Goldman, Jean-Philippe, Tea Prsir, and Antoine Auchlin. 2014. "C-PhonoGenre: A 7-Hour Corpus of 7 Speaking Styles in French: Relations between Situational Features and Prosodic Properties." In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC'14)*: 302–305.

Goldman-Eisler, Frieda. 1968. *Psycholinguistics: Experiments in Spontaneous Speech*. London: Academic Press.

Gómez González, María de los Ángeles. 2014. "Canonical Tag Questions in English, Spanish and Portuguese. A Discourse-Functional Study." *Languages in Contrast* 14 (1): 93–126. doi:10.1075/lic.14.1.06gom

González, Montserrat. 2005. "Pragmatic Markers and Discourse Coherence Relations in English and Catalan Oral Narrative." *Discourse Studies* 77 (1), 53–86. doi:10.1177/1461445605048767

Götz, Sandra. 2013. *Fluency in Native and Nonnative English Speech*. Amsterdam: John Benjamins. doi:10.1075/scl.53

Granger, Sylviane, and Stephanie Petch-Tyson. 1996. "Connector Usage in the English Essay Writing of Native and Non-Native EFL Speakers of English." *World Englishes: Journal of English as an International and Intranational Language* 15 (1): 17–27. doi:10.1111/j.1467-971X.1996.tb00089.x

Greene, John O., and Joseph N. Cappella. 1986. "Cognition and Talk: The Relationship of Semantic Units to Temporal Patterns of Fluency in Spontaneous Speech." *Language and Speech* 29 (2): 141–157.

Grice, Herbert P. 1957. "Meaning." *The Philosophical Review* 66: 377–388. doi:10.2307/2182440

Gries, Stefan. 2011. "Corpus Data in Usage-Based Linguistics: What's the Right Degree of Granularity for the Analysis of Argument Structure Constructions?" In *Cognitive Linguistics: Convergence and Expansion*, ed. by Mario Brdar, Stefan Gries, and Milena Žic Fuchs, 237–256. Amsterdam: John Benjamins. doi:10.1075/hcp.32.15gri

Gries, Stefan, and Anatol Stefanowitsch (eds). 2006. *Corpora in Cognitive Linguistics: Corpus-Based Approaches to Syntax and Lexis*. Berlin: Mouton de Gruyter. doi:10.1515/9783110197709

Grosjean, François, and Alain Deschamps. 1975. "Analyse contrastive des variables temporelles de l'anglais et du français: Vitesse de parole et variables composantes, phénomènes d'hésitation [Contrastive analysis of temporal variables in English and French: Speech rate and other variables, hesitation phenomena]." *Phonetica* 31: 144–184. doi:10.1159/000259667

Grosman, Iulia. 2016. "How Do French Humorists Manage their Persona across Situations? A Corpus Study on their Prosodic Variation." In *Metapragmatics of Humor: Current Research Trends*, ed. by Leonor Ruiz-Gurillo, 147–175. Amsterdam: John Benjamins. doi:10.1075/ivitra.14.08gro

Guillemin-Flescher, Jacqueline. 1981. *Syntaxe comparée du français et de l'anglais* [*Comparative syntax of French and English*]. Paris: Ophrys.

Gülich, Elisabeth, and Thomas Kotschi. 1987. "Les actes de reformulation dans la consultation La dame de Caluire [Reformulative acts in the consultation The Lady of Caluire]." In *L'Analyse des interactions verbales. La dame de Caluire: Une Consultation*, ed. by Pierre Bange, 15–81. Bern: Peter Lang.

Gülich, Elisabeth, and Thomas Kotschi. 1995. "Discourse Production in Oral Communication." In *Aspects of Oral Communication*, ed. by Uta M. Quasthoff, 30–66. Berlin: Walter de Gruyter. doi:10.1515/9783110879032.30

Halliday, Michael A. K. 1970. "Functional Diversity in Language as Seen from a Consideration of Modality and Mood in English." *Foundations of Language: International Journal of Language and Philosophy* 6: 322–361.

Halliday, Michael A. K. 1987. "Spoken and Written Modes of Meaning." In *Comprehending Oral and Written Language*, ed. by Roalind Horowitz, and S. Jay Samuels, 55–82. New York: Academic Press.

Halliday, Michael A. K., and Ruqaiya Hasan. 1976. *Cohesion in English*. London: Longman.

Halliday, Michael A. K., and Ruqaiya Hasan. 1989. *Language, Context, and Text: Aspects of Language in a Social-Semiotic Perspective*. Oxford: Oxford University Press.

Hansen, Maj-Britt M. 2006. "A Dynamic Polysemy Approach to the Lexical Semantics of Discourse Markers (with an Exemplary Analysis of French *toujours*)." In *Approaches to Discourse Particles*, ed. by Kerstin Fischer, 21–41. Amsterdam: Elsevier.

Hansen, Maj-Britt M. 2008. *Particles at the Semantics/Pragmatics Interface: Synchronic and Diachronic Issues. A Study with Special Reference to the French Phrasal Adverbs*. Elsevier: Oxford.

Haselow, Alexander. 2012. "Subjectivity, Intersubjectivity and the Negotiation of Common Ground in Spoken Discourse: Final Particles in English." *Language and Communication* 32: 182–204. doi:10.1016/j.langcom.2012.04.008

Hasselgren, Alexander. 2002. "Learner Corpora and Language Testing: Small Words as Markers of Learner Fluency." In *Computer-Learner Corpora, Second Language Acquisition, and Foreign Language Teaching*, ed. by Sylviane Granger, Joseph Hung, and Stephanie Petch-Tyson, 143–173. Philadelphia: John Benjamins. doi:10.1075/lllt.6.11has

Heeman, Peter, and James Allen. 1999. "Speech Repairs, Intonational Phrases and Discourse Markers: Modeling Speakers' Utterances in Spoken Dialog." *Computational Linguistics* 25 (4): 1–45.

Hieke, Adolf E. 1985. "A Componential Approach to Oral Fluency Evaluation." *The Modern Language Journal* 69 (2): 135–142. doi:10.1111/j.1540-4781.1985.tb01930.x

Hoffmann, Sebastian. 2004. "Are Low-Frequency Complex Prepositions Grammaticalized? On the Limits of Corpus Data – and the Importance of Intuition." In *Corpus Approaches to Grammaticalization in English*, Hans Lindquist, and Christian Mair, 171–210. Amsterdam: John Benjamins. doi: 10.1075/scl.13.09hof

Hopper, Paul. 1987. "Emergent grammar." In *Proceedings of the Thirteenth Annual Meeting of the Berkeley Linguistics Society*, ed. by Jon Aske, Natasha Beery, Laura Michaelis, and Hana Filip. Berkeley, CA: Berkeley Linguistics Society.

Hopper, Paul, and Elizabeth C. Traugott. 2003. *Grammaticalization (Second Edition)*. Cambridge: Cambridge University Press.

Hothorn, Torsten, Kurt Hornik, and Achim Zeileis. 2006. "Unbiased Recursive Partitioning: A Conditional Inference Framework." *Journal of Computational and Graphical Statistics* 15 (3): 651–674. doi: 10.1198/106186006X133933

Hunt, Kellogg W. 1965. *Grammatical Structures Written at Three Grade Levels. NCTE Research Report 3*. Champaign, Ill.: NCTE.

Husserl, Edmund. 1964. *The Phenomenology of Internal Time-Consciousness*. Bloomington, IN: Indiana University Press.

Izutsu, Mitsuko N., and Katsunobu Izutsu. 2014. "Truncation and Backshift: Two Pathways to Sentence-Final Coordinating Conjunctions." *Journal of Historical Pragmatics* 15 (1): 62–92. doi: 10.1075/jhp.15.1.04izu

James, Carl. 1980. *Contrastive Analysis*. Harlow: Longman.

Jaszczolt, Katarzyna M.> 2003. "On Translating 'What is Said': Tertium Comparationis in Contrastive Semantics and Pragmatics." In *Meaning through Language Contrast*, ed. by Katarzyna M. Jaszczolt, and Ken Turner, 441–462. Amsterdam: John Benjamins. doi: 10.1075/pbns.100.26jas

Kemmer, Suzanne, and Michael Barlow. 2000. "Introduction: A Usage-Based Conception of Language." In *Usage Based Models of Language*, ed. by Michael Barlow, and Suzanne Kemmer, vii–xxviii. Stanford: CSLI.

Koch, Peter, and Wulf Osterreicher. 2001. "Gesprochene Sprache und geschriebene Sprache / Langage parlé et langage écrit [Spoken language and written language]." In *Lexikon der Romanistischen Linguistik*, ed. by Günter Holtus, Michael Metzeltin, and Christian Schmitt, 584–627. Bd. I / 2. Tübingen: Niemeyer.

Kohn, Kurt. 2012. "Pedagogic Corpora for Content and Language Integrated Learning. Insights from the BACKBONE Project." *The Eurocall Review* 20 (2).

Koponen, Matti, and Heidi Riggenbach. 2000. "Overview: Varying Perspectives on Fluency." In *Perspectives on Fluency*, ed. by Heidi Riggenbach, 5–25. Ann Arbor, MI: The University of Michigan Press.

Kormos, Judit. 2006. *Speech Production and Second Language Acquisition*. London: Lawrence Erlbaum Associates.

Krzeszowski, Tomasz P. 1981. "Tertium Comparationis." In *Contrastive Linguistics: Prospects and Problems*, ed. by Jacek Fisiak, 301–312. Berlin: Mouton de Gruyter.

Kunz, Kerstin, and Ekaterina Laphinova-Koltunski. 2015. "Cross-Linguistic Analysis of Discourse Variation across Registers." *Nordic Journal of English Studies* 14 (1): 258–288.

Labov, William. 1972. *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

Lacheret, Anne, Sylvain Kahane, and Paola Pietrandrea (eds). 2014. *Rhapsodie: A Prosodic and Syntactic Treebank for Spoken French*. Amsterdam: John Benjamins.

Lakoff, George. 1987. *Women, Fire, and Dangerous Things. What Categories Reveal about the Mind*. Chicago: The University of Chicago Press. doi: 10.7208/chicago/9780226471013.001.0001

Lakoff, George, and Mark Johnson. 1980. *Metaphors We Live By*. Chicago: University of Chicago Press.

Langacker, Ronald. 1987. *Foundations of Cognitive Grammar, Vol.1: Theoretical Prerequisites*. Stanford: Stanford University Press.

Langacker, Ronald. 1988. "An Overview of Cognitive Grammar." In *Topics in Cognitive Linguistics*, ed. by Brygida Rudzka-Ostyn, 3–48. Amsterdam: John Benjamins. doi:10.1075/cilt.50.03lan

Langacker, Ronald. 1990. *Concept, Image, and Symbol: The Cognitive Basis of Grammar*. Berlin: Mouton de Gruyter.

Langacker, Ronald. 2000. "A Dynamic Usage-Based Model." In *Usage Based Models of Language*, ed. by Michael Barlow, and Suzanne Kemmer, 1–63. Stanford: CSLI.

Lapshinova-Koltunski, Ekaterina, Anna Nedoluzhko, and Kerstin Kunz. 2015. "Across Languages and Genres: Creating a Universal Annotation Scheme for Textual Relation." In *Proceedings of LAW IX at NAACL HLT 2015*, Denver, USA.

Lee, Hye-Kyung. 2002. "Towards a New Typology of Connectives with Special Reference to Conjunction in English and Korean." *Journal of Pragmatics* 34: 851–866. doi:10.1016/S0378-2166(01)00065-0

Leijten, Mariëlle, and Luuk Van Waes. 2013. "Keystroke Logging in Writing Research Using Inputlog to Analyze and Visualize Writing Processes." *Written Communication* 30 (3): 358–392. doi:10.1177/0741088313491692

Lenk, Uta. 1998. "Discourse Markers and Global Coherence in Conversation." *Journal of Pragmatics* 30: 245–257. doi:10.1016/S0378-2166(98)00027-7

Lennon, Paul. 2000. "The Lexical Element in Spoken Second Language Fluency." In *Perspectives on Fluency*, ed. by Heidi Riggenbach, 25–42. Ann Arbor: The University of Michigan Press.

Léon, Pierre Roger. 1993. *Précis de Phonostylistique, Parole et Expressivité*. Paris: Nathan Université.

Levelt, Willem J. M. 1981. "The Speaker's Linearization Problem." *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences* 295 (1077): 305–315. doi:10.1098/rstb.1981.0142

Levelt, Willem J. M. 1983. "Monitoring and Self-Repair in Speech." *Cognition* 14: 41–104. doi:10.1016/0010-0277(83)90026-4

Levelt, Willem J. M. 1989. *Speaking: From Intention to Articulation*. Cambridge: MIT Press.

Lewis, Diana. 2006a. "Contrastive Analysis of Adversative Relational Markers, Using Comparable Corpora." In *Pragmatic Markers in Contrast*, ed. by Karin Aijmer, and Anne-Marie Simon-Vandenbergen, 139–153. Amsterdam: Elsevier.

Lewis, Diana. 2006b. "Discourse Markers in English: A Discourse-Pragmatic View." In *Approaches to Discourse Particles*, ed. by Kerstin Fischer, 43–60. Amsterdam: Elsevier.

Lindström, Jan. 2001. "Inner and Outer Syntax of Constructions: The Case of the X Och X Construction in Swedish." Paper presented at the 7th International Pragmatics Conference, July 9–14 2000, Budapest, Hungary.

Linell, Per. 1982. *The Written Language Bias in Linguistics*. Linköping, Sweden: University of Linköping.

Little, Daniel R., Raoul Oehmen, John Dunn, Kathryn Hird, and Kim Kirsner. 2013. "Fluency Profiling System: An Automated System for Analyzing the Temporal Properties of Speech." *Behavioral Research Methods* 45 (1): 191–202. doi:10.3758/s13428-012-0222-0

Liu, Kris, and Jean E. Fox Tree. 2012. "Hedges Enhance Memory but Inhibit Retelling." *Psychon Bull Rev* 19: 892–898. doi:10.3758/s13423-012-0275-1

Lopes, António, David Martins de Matos, Vera Cabarrão, Ricardo Ribeiro, Helena Moniz, Isabel Trancoso, and Ana Isabel Mata. 2015. "Towards Using Machine Translation Techniques to Induce Multilingual Lexica of Discourse Markers." *CoRR* abs/1503.09144.

Luscher, Jean-Marc. 1993. "La marque de connexion complexe [The complex connection marker]." *Cahiers de Linguistique Française* 14: 173–188.

Luscher, Jean-Marc. 1994. "Marques de connexion: Procédures de traitement et guidage référentiel [Connection markers: Processing and referential guidance]." In *Langage et pertinence. Référence Temporelle, Anaphore, Connecteurs et Métaphore*, ed. by Jacques Moeschler, Anne Reboul, Jean-Marc Luscher, and Jacques Jayez, 175–228. Nancy: Presses universitaires de Nancy.

MacGregor, Lucy, Corley, Martin, and David Donaldson. 2009. "Not All Disfluencies Are Are Equal: The Effects of Disfluent Repetitions on Language Comprehension." *Brain & Language* 111: 36–45. doi:10.1016/j.bandl.2009.07.003

Maclay, Howard, and Charles E. Osgood. 1959. "Hesitation Phenomena in Spontaneous English Speech." *Word* 15: 19–44 doi:10.1080/00437956.1959.11659682

Mahesha, P. & Vinod, D. 2012. "An Approach for Classification of Dysfluent and Fluent Speech Using K-NN and SVM." *International Journal of Computer Science, Engineering and Applications (IJCSEA)* 2 (6): 23–31. doi:10.5121/ijcsea.2012.2603

Mann, William, and Sandra A Thompson. 1988. "Rhetorical Structure Theory: Toward a Functional Theory of Text Organization." *Text* 8 (3): 243–281. doi:10.1515/text.1.1988.8.3.243

Marchi, Anna, and Charlotte Taylor. 2009. "If on a Winter's Night Two Researchers… A Challenge to Assumptions of Soundness of Interpretation." *Critical Approaches to Discourse Analysis across Disciplines* 3 (1): 1–20.

Maschler, Yael, and Deborah Schiffrin. 2015. "Discourse Markers: Language, Meaning, and Context." In *The Handbook of Discourse Analysis*. 2nd Edition, ed. by Deborah Tannen, Heidi E. Hamilton, and Deborah Schiffrin, 189–221. Hoboken, NJ: John Wiley & Sons.

McHugh, Mary L. 2012. "Interrater Reliability: The Kappa Statistic." *Biochemia Medica* 22 (3): 276–282. doi:10.11613/BM.2012.031

Merlo, Sandra, and Leticia Mansur. 2004. "Descriptive Discourse: Topic Familiarity and Disfluencies." *Journal of Communication Disorders* 37: 489–503. doi:10.1016/j.jcomdis.2004.03.002

Meteer, Marie W., Ann A. Taylor, Robert MacIntyre, and Rukmini Iyer. 1995. *Disfluency Annotation Stylebook for the Switchboard Corpus*. Technical report, Linguistic Data Consortium.

Meyer, Thomas, Andrei Popescu-Belis, Najeh Hajlaoui, and Andrea Gesmundo. 2012. "Machine Translation of Labeled Discourse Connectives." In *Proceedings of the Tenth Biennial Conference of the Association for Machine Translation in the Americas (AMTA)*.

Meyer, David, Achim Zeileis, and Kurt Hornik. 2014. *vcd: Visualizing Categorical Data*. R package version 1.3–2.

Mieskes, Margot, and Michael Strube. 2008. "A Three-Stage Disfluency Classifier for Multi Party Dialogues." In *Proceedings of the 6th International Conference on Language Resources and Evaluation*: 2681–2686.

Miskovic-Lukovic, Mirjana. 2009. "Is There a Chance That I Might Kinda Sort of Take You out to Dinner?: The Role of the Pragmatic Particles *Kind of* and *Sort of* in Utterance Interpretation." *Journal of Pragmatics* 41: 602–625. doi:10.1016/j.pragma.2008.06.014

Moniz, Helena. 2013. *Processing Disfluencies in European Portuguese*. PhD thesis, Universidade de Lisboa.

Mulder, Jean, and Sandra A. Thompson. 2008. "The Grammaticization of *but* as a Final Particle in English Conversation." In *Crosslinguistic Studies of Clause Combining: The Multifunctionality of Conjunctions*, ed. by Ritva Laury, 179–204. Amsterdam: John Benjamins. doi:10.1075/tsl.80.09mul

Müller, Simone. 2005. *Discourse Markers in Native and Non-native English Discourse*. Amsterdam: John Benjamins. doi:10.1075/pbns.138

Murillo, Silvia. 2016. "Reformulation Markers and Polyphony. A Contrastive English-Spanish Analysis." *Languages in Contrast* 16 (1): 1–30.

Mustajoki, Arto. 2012. "A Speaker-Oriented Multidimensional Approach to Risks and Causes of Miscommunication." *Language and Dialogue* 2 (2): 216–243. doi:10.1075/ld.2.2.03mus

Nakatani, Christine H., and Julia Hirschberg. 1994. "A Corpus-Based Study of Repair Cues in Spontaneous Speech." *Journal of the Acoustical Society of America* 95 (3): 1603–1616. doi:10.1121/1.408547

Nelson, Gerard, Sean Wallis, and Bas Aarts. 2002. *Exploring Natural Language: Working with the British Component of the International Corpus of English*. Amsterdam: John Benjamins. doi:10.1075/veaw.g29

Neumann, Stella. 2014. "Cross-Linguistic Register Studies. Theoretical and Methodological Considerations." *Languages in Contrast* 14 (1): 35–57. doi:10.1075/lic.14.1.03neu

Nølke, Henning. 2006. "Pour une théorie linguistique de la polyphonie: Problèmes, avantages, perspectives [For a linguistic theory of polyphony: Problems, advantages, perspectives]." In *Le Sens et ses Voix. Dialogisme et Polyphonie en Langue et en Discours*, ed. by Laurent Perrin, 243–269. Metz: Université Paul Verlaine.

Nzoimbengene, Philippe. 2016. *Les 'discourse markers' en lingála. Étude sémantique et pragmatique sur base d'un corpus de lingála de Kinshasa oral* [*Discourse markers in Lingála. Semantic and pragmatic study on a corpus of spoken Língala from Kinshasa*]. PhD thesis, Université catholique de Louvain.

O'Connell, Daniel C., and Sabine H. Kowal. 1972. "Cross-Linguistic Pause and Rate Phenomena in Adults and Adolescents." *Journal of Psycholinguistic Research* 1: 155–164. doi:10.1007/BF01068105

O'Donnell, W. R., and Loreto Todd. 1980. *Variety in Contemporary English*. London: Allen and Unwin. doi:10.4324/9780203308288

Oza, Umangi, Rashmi Prasad, Sudheer Kolachina, Suman Meena, Dipti Misra Sharma, and Aravind Joshi. 2009. "Experiments with Annotating Discourse Relations in the Hindi Discourse Relation Bank." In *Proceedings of the 7th International Conference on Natural Language Processing (ICON)*: 1–10.

Palisse, Stéphanie. 1997. *"Artisans"*, "*Assureurs*", *Conversations téléphoniques en entreprise* [*"Craftsmen", "Insurance workers", Phone business conversations*]. Retrieved from http://clapi-univ.lyon2.fr (last accessed March 2014).

Pallaud, Bertille, Stéphane Rauzy, and Philippe Blâche. 2013. "Auto-interruptions et disfluences en français parlé dans quatre corpus du CID [Self-interruptions and disfluencies in spoken French in four corpora of the CID]." *TIPA Travaux Interdisciplinaires sur la Parole et le Langage* 29.

Pander Maat, Henk L. W., and Liesbeth Degand. 2001. "Scaling Causal Relations and Connectives in Terms of Speaker Involvement." *Cognitive Linguistics* 12 (3): 211–245.

Pander Maat, Henk L. W., and Ted J. M. Sanders. 2000. "Domains of Use and Subjectivity. On the Distribution of Three Dutch Causal Connectives." In *Cause, Condition, Concession and Contrast: Cognitive and Discourse Perspectives*, ed. by Bernd Kortmann, and Elizabeth Couper-Kuhlen, 57–82. Berlin: Mouton de Gruyter. doi:10.1515/9783110219043.1.57

Pander Maat, Henk L. W., and Ted J. M. Sanders. 2001. "Subjectivity in Causal Connectives: An Empirical Study of Language in Use." *Cognitive Linguistics* 12 (3): 247–273.

Pascual, Elena, and Ludivine Crible. 2017. "Discourse Markers within (Dis)fluent Constructions in English, French and Spanish Casual Conversations: The Challenges of Contrastive Fluency Research." Paper presented at the International Conference on Fluency and Disfluency across Languages and Language Varieties, February 15–17, Louvain-la-Neuve, Belgium.

Pawley, Andrew, and Frances Syder. 1975. "Sentence Formulation in Spontaneous Speech." *New Zealand Speech Therapists' Journal* 30 (2): 2–11.

Pawley, Andrew, and Frances Syder. 1983. "Two Puzzles for Linguistic Theory: Nativelike Selection and Nativelike Fluency." In *Language and Communication*, ed. by Jack C. Richards, and Richard Schmidt, 191–225. London: Longman.

Pawley, Andrew, and Frances Syder. 2000. "The One-Clause-at-a-Time Hypothesis." In *Perspectives on Fluency*, ed. by Heidi Riggebbach, 163–199. Ann Arbor: The University of Michigan Press.

Pichler, Heike. 2016. "Uncovering Discourse-Pragmatic Innovations: *Innit* in Multicultural London English." In *Discourse-Pragmatic Variation and Change in English: New Methods and Insights*, ed. by Heike Pichler, 59–85. Cambridge: Cambridge University Press.

Pitler, Emily, and Ani Nenkova. 2009. "Using Syntax to Disambiguate Explicit Discourse Connectives in Text." In *Proceedings of the ACL-IJCNLP Conference Short Papers: 13–16*.

Poggi, Isabella, and Emanuela Magno Caldognetto. 1996. "A Score for the Analysis of Gestures in Multimodal Communication." In *Proceedings of the Workshop on the Integration of Gesture and Language in Speech*, ed. by Lynn Messing, 235–244. Newark and Wilmington: Applied Science and Engineering Laboratories, University of Delaware.

Pons Bordería, Salvador. 2008. "La combinación de marcadores del discurso en la conversación coloquial: Interacciones entre posición y función [Combination of discourse markers in casual conversation: Interactions between position and funcion]." *Estudos Linguísticos/ Linguistic Studies* 2: 141–159.

Popescu-Belis, Andrei, Thomas Meyer, Jeevanthi Liyanapathirana, Bruno Cartoni, and Sandrine Zufferey. 2012. "Discourse-Level Annotation over Europarl for Machine Translation: Connectives and Pronouns." In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'8)*.

Postma, Albert, Herman Kolk, and Dirk-Jan Povel. 1990. "On the Relation among Speech Errors, Disfluencies, and Self-Repairs." *Language and Speech* 33 (1): 19–29.

Prasad, Rashmi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. "The Penn Discourse TreeBank 2.0." In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 08), Marrackech, Morroco*: 2961–2968.

Ragan, Sandra L. 1983. "Alignment and Conversational Coherence." In *Conversational Coherence: Form, Structure and Strategy*, ed. by Robert T. Craig, and Karen Tracy, 157–171. Beverly Hills: Sage Publications.

Razgouliaeva, Anna. 2002. "Combinaisons des connecteurs *mais enfin* [Combinations of the connectives *mais enfin* 'but well']." *Cahiers de Linguistique Française* 24: 143–168.

Redeker, Gisela. 1991. "Linguistic Markers of Discourse Structure." *Linguistics* 29: 1139–1172.

Reese, Brian, and Nicholas Asher. 2007. "Prosody and the Interpretation of Tag Questions." In *Proceedings of Sinn und Bedeutung 11*, ed. by Estela Puig-Waldmüller, 448–462. Barcelona: Universitat Pompeu Fabra.

Rendle-Short, Johanna. 2004. "Showing Structure: Using *Um* in the Academic Seminar." In *Pragmatics* 14 (4): 479–498. doi:10.1075/prag.14.4.04ren

Roberts, Benjamin, and Kim Kirsner. 2000. "Temporal Cycles in Speech Production." *Language and Cognitive Processes* 15 (2): 129–157. doi:10.1080/016909600386075

Rossari, Corinne. 1990. "Projet pour une typologie des opérations de reformulation [Project for a typology of reformulative operations]." *Cahiers de Linguistique Française* 11: 345–359.

Rossari, Corinne. 1994. *Les Opérations de reformulation* [*Reformulative operations*]. Bern: Peter Lang.

Roulet, Eddy, Antoine Auchlin, Jacques Moeschler, and Christian Rubattel. 1985. *L'Articulation du discours en français Contemporain* [*Discourse articulation in contemporary French*]. Bern: Peter Lang.

Rühlemann, Christoph, and Matthew B. O'Donnell. 2012. "Introducing a Corpus of Conversational Stories. Construction and Annotation of the Narrative Corpus." *Corpus Linguistics and Linguistic Theory* 8 (2): 313–350. doi:10.1515/cllt-2012-0015

Sanders, Ted J. M. 1997. "Semantic and Pragmatic Sources of Coherence: On the Categorization of Coherence Relations in Context." *Discourse Processes* 24 (1): 119–147. doi:10.1080/01638539709545009

Santorini, Beatrice. 1990. *Part-of-Speech Tagging Guidelines for the Penn Treebank Project (3rd Revision, 2nd Printing)*. Technical Report, Department of Computer and Information Science, University of Pennsylvania.

Schegloff, Emanuel, Gail Jefferson, and Harvey Sacks. 1977. "The Preference for Self-Correction in the Organization of Repair in Conversation." *Language* 53: 361–382. doi:10.1353/lan.1977.0041

Schiffrin, Deborah. 1987. *Discourse Markers*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511611841

Schmid, Hans-Jörg. 2000. *English Abstract Nouns as Conceptual Shells. From Corpus to Cognition*. Berlin: Mouton de Gruyter. doi:10.1515/9783110808704

Schmid, Hans-Jörg. 2012. "Generalizing the Apparently Ungeneralizable. Basic Ingredients of a Cognitive-Pragmatic Approach to the Construal of Meaning-in-Context." In *Cognitive Pragmatics*, ed. by Hans-Jörg Schmid, 3–22. Berlin: Mouton de Gruyter. doi:10.1515/9783110214215.3

Schmidt, Richard. 1992. "Psychological Mechanisms Underlying Language Fluency." *Studies in Second Language Acquisition* 14: 357–385. doi:10.1017/S0272263100011189

Schmidt, Thomas, and Kai Wörner. 2012. "EXMARaLDA." In *Handbook on Corpus Phonology*, ed. by Jacques Durand, Ulrike Gut, and Gjert Kristoffersen, 402–419. Oxford: Oxford University Press.

Schönefeld, Doris. 1999. "Corpus Linguistics and Cognitivism." *International Journal of Corpus Linguistics* 4 (1): 137–171. doi:10.1075/ijcl.4.1.07sch

Schourup, Lawrence. 1999. "Discourse markers." *Lingua* 107: 227–265. doi:10.1016/S0024-3841(96)90026-1

Segalowitz, Norman. 2010. *Cognitive Bases of Second Language Fluency*. New York: Routledge.

Seyfeddinipur, Mandana. 2006. *Disfluency: Interrupting Speech and Gesture*. MP Series in Psycholinguistics.

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

Simon, Anne-Catherine, Antoine Auchlin, Matthieu Avanzi, and Jean-Philippe Goldman. 2010. "Les phonostyles: Une description prosodique des styles de parole en français. [Phonostyles: A prosodic description of speech styles in French]." In *Les Voix des Français. En Parlant, en Ecrivant*, ed. by Michaël Abecassi, and Gudrun Ledegen, 71–88. Bern: Peter Lang.

Sperber, Dan, and Deirdre Wilson. 1986. *Relevance. Communication and Cognition*. Oxford: Blackwell.

Spooren, Wilbert, and Liesbeth Degand. 2010. "Coding Coherence Relations: Reliability and Validity." *Corpus Linguistics and Linguistic Theory* 6 (2): 241–266. doi:10.1515/cllt.2010.009

Strassel, Stephanie. 2003. *Simple Metadata Annotation Specification v.5*. Technical report, Linguistic Data Consortium.

Stubbe, Maria, and Janet Holmes. 1995. "*You Know*, *Eh* and Other 'Exasperating Expressions': An Analysis of Social and Stylistic Variation in the Use of Pragmatic Devices in a Sample of New Zealand English." *Language & Communication* 15 (1): 63–88. doi:10.1016/0271-5309(94)00016-6

Stukker, Ninke, and Ted J. M. Sanders. 2012. "Subjectivity and Prototype Structure in Causal Connectives: A Cross-Linguistic Perspective." *Journal of Pragmatics* 44: 169–190. doi:10.1016/j.pragma.2011.06.011

Sweetser, Eve. 1990. *From Etymology to Pragmatics*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511620904

Swerts, Marc. 1998. "Filled Pauses as Markers of Discourse Structure." *Journal of Pragmatics* 30: 485–496. doi:10.1016/S0378-2166(98)00014-9

Tesnière, Lucien. 1959. *Éléments de syntaxe structurale* [*Elements of structural syntax*]. Paris: Klincksieck.

Tonelli, Sara, Giuseppe Riccardi, Rashmi Prasad, and Aravind Joshi. 2010. "Annotation of Discourse Relations for Conversational Spoken Dialogs." In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 10)*, Valletta, Malta: 2084–2090.

Tottie, Gunnel. 2015. "*Uh* and *Um* in British and American English: Are They Words? Evidence from Co-Occurrence with Pauses." In *Linguistic variation: Confronting Fact and Theory*, ed. by Nathalie Dion, André Lapierre, and Rena Torres Cacoullos, 38–54. New York: Routledge.

Traugott, Elizabeth C. 2007. "(Inter)subjectification and Unidirectionality." *Journal of Historical Pragmatics* 8: 295–309. doi:10.1075/jhp.8.2.07clo

Unger, Christoph. 1996. "The Scope of Discourse Connectives: Implications for Discourse Organization." *Journal of Linguistics* 32: 403–438. doi:10.1017/S0022226700015942

Urgelles-Coll, Miriam. 2012. *The Syntax and Semantics of Discourse Markers*. London: Bloomsbury.

Van Bogaert, Julie. 2011. "*I Think* and Other Complement-Taking Mental Predicates: A Case of and for Constructional Grammaticalization." *Linguistics* 49 (2): 295–332.

Vasilescu, Ioana, Rena Nemoto, and Martine Adda-Decker. 2007. "Vocalic Hesitations vs Vocalic Systems: A Cross-Language Comparison." In *Proceedings of the ICPhS 16th International Congress of Phonetic Science*.

Vinay, Jean-Paul, and Jean Darbelnet. 1995. *Comparative Stylistics of French and English: A Methodology for Translation*. Translated and ed. by Juan C. Sager, and Marie-Josée Hamel. Amsterdam: John Benjamins. doi:10.1075/btl.11

Wagner, Suzanne E., and Ashley Hesson. 2014. "Individual Sensitivity to the Frequency of Socially Meaningful Linguistic Cues Affects Language Attitudes." *Journal of Language and Social Psychology* 33 (6): 651–666. doi:10.1177/0261927X14528713

Waltereit, Richard. 2007. "À propos de la genèse diachronique des combinaisons de marqueurs. L'exemple de *bon ben* et *enfin bref* [About the diachronic genesis of combinations of markers. The example of *bon ben* 'right well' and *enfin bref* 'I mean anyway']." *Langue Française* 154: 94–128.

Waltereit, Richard, and Ulrich Detges. 2007. "Different Functions, Different Histories. Modal Particles and Discourse Markers from a Diachronic Point of View." In *Catalan Journal of Linguistics* 6: 61–80.

Watanabe, Michiko, Keikichi Hirose, Yasuharu Den, and Nobuaki Minematsu. 2008. "Filled Pauses as Cues to the Complexity of Up-Coming Phrases for Native and Non-Native Listeners." *Speech Communications* 50: 81–94. doi:10.1016/j.specom.2007.06.002

Willems, Dominique, and Annemie Demol. 2006. "*Vraiment* and *really* in Contrast: When Truth and Reality Meet." In *Pragmatic Markers in Contrast*, ed. by Karin Aijmer, and Anne-Marie Simon-Vandenbergen, 215–235. Amsterdam: Elsevier.

Zechner, Klaus. 2001. *Automatic Summarization of Spoken Dialogues in Unrestricted Domains*. PhD thesis, Carnegie Mellon University.

Zeileis, Achim, David Meyer, and Kurt Hornik. 2007. "Residual-Based Shadings for Visualizing Conditional Independence." *Journal of Computational and Graphical Statistics* 16 (3): 507–525. doi:10.1198/106186007X237856

Zeyrek, Deniz, Işın Demirşahin, Ayişigi B. Sevdik Çalli, and Ruken Çakici. 2013. "Turkish Discourse Bank: Porting a Discourse Annotation Style to a Morphologically Rich Language." *Dialogue & Discourse* 4 (2): 174–184. doi:10.5087/dad.2013.208

Zhao, Yuan, and Dan Jurafsky. 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.

Zufferey, Sandrine, and Bruno Cartoni. 2012. "English and French Causal Connectives in Contrast." *Languages in Contrast* 12 (2): 232–250. doi:10.1075/lic.12.2.06zuf

Zufferey, Sandrine, and Liesbeth Degand. 2013. "Representing the Meaning of Discourse Connectives for Multilingual Purposes." *Corpus Linguistics and Linguistic Theory* 10.

Zufferey, Sandrine, and Andrei Popescu-Belis. 2004. "Towards Automatic Identification of Discourse Markers in Dialogs: The Case of 'Like'." In *Proceedings of SIGDIAL'04 (5th SIGdial Workshop on Discourse and Dialogue)*, Cambridge, MA: 63–71.

APPENDIX 1

Discourse markers by register

	1st	2nd	3rd	4th	5th
English					
convers.	and (176)	well (132)	but (123)	so (83)	you know (68)
phone	but (91)	well (88)	and (83)	so (83)	I mean (36)
interview	and (343)	so (149)	but (82)	you know (67)	because (46)
radio	and (160)	but (39)	because (38)	I mean (33)	if (23)
classroom	and (105)	so (72)	but (51)	if (36)	I mean (36)
sports	and (174)	but (46)	as (17)	so (15)	well (14)
political	and (69)	but (24)	if (16)	when (9)	indeed (8)
news	and (30)	but (21)	when (6)	if (5)	however (4)
French					
convers.	et (263)	mais (242)	quoi (216)	enfin (94)	ben (70)
phone	donc (66)	alors (60)	hein (39)	parce que (38)	et (36)
interview	et (246)	mais (120)	hein (112)	alors (110)	donc (93)
radio	et (96)	mais (58)	parce que (34)	alors (25)	donc (24)
classroom	et (37)	donc (19)	bon (12)	mais (11)	alors (8)
sports	et (105)	mais (33)	hein (19)	donc (13)	alors que (11)
political	et (44)	si (21)	mais (15)	alors (9)	pour que (8)
news	et (42)	mais (27)	donc (8)	alors que (4)	et puis (4)

APPENDIX 2

List of discourse markers in *DisFrEn* and their functions

English discourse markers (4249)	
and (1140)	*addition* (651), *specification* (180), *consequence* (101), *topic-shift* (41), *temporal* (27), *punctuation* (24), *conclusion* (20), *topic-resuming* (16), *contrast* (13), *opening boundary* (13), *enumeration* (10), *comment* (9), *concession* (8), *emphasis* (6), *quoting* (4), *addition-punctuation* (3), *opposition* (3), *punctuation-conclusion* (2), *topic-resuming-conclusion* (2), *specification-comment* (1), *motivation* (1), *enumeration-topic-resuming* (1), *punctuation-consequence* (1), *topic-shift-specification* (1), *contrast-addition* (1), *topic-resuming-specification* (1)
but (477)	*opposition* (203), *concession* (142), *contrast* (38), *topic-resuming* (22), *topic-resuming-opposition* (10), *closing boundary* (9), *topic-shift* (9), *closing boundary-opposition* (9), *topic-shift-opposition* (6), *punctuation* (6), *opening boundary* (6), *opening boundary-opposition* (4), *topic-resuming-motivation* (1), *addition* (1), *specification* (1), *enumeration-opposition* (1), *addition-opposition* (1), *topic-resuming-conclusion* (1), *cause-topic-resuming* (1), *disagreeing* (1), *cause-topic-shift* (1), *emphasis* (1), *punctuation-opposition* (1), *exception* (1), *reformulation* (1)
so (429)	*conclusion* (198), *consequence* (123), *specification* (25), *topic-shift* (17), *topic-resuming* (16), *topic-resuming-conclusion* (16), *closing boundary-conclusion* (10), *closing boundary* (7), *punctuation* (3), *reformulation* (3), *addition* (2), *opening boundary-conclusion* (2), *consequence-specification* (1), *punctuation-conclusion* (1), *topic-resuming-consequence* (1), *emphasis* (1), *enumeration* (1), *motivation* (1), *opening boundary* (1)
well (304)	*opening boundary* (177), *reformulation* (26), *punctuation* (20), *disagreeing* (15), *quoting* (12), *topic-shift* (10), *disagreeing-opening boundary* (7), *agreeing* (6), *emphasis* (4), *comment* (4), *specification* (3), *conclusion* (3), *punctuation-conclusion* (3), *topic-resuming* (2), *disagreeing-reformulation* (2), *face-saving* (2), *opening boundary-motivation* (2), *motivation-reformulation* (1), *topic-resuming-reformulation* (1), *disagreeing-punctuation* (1), *punctuation-reformulation* (1), *opening boundary-specification* (1), *comment-reformulation* (1)

English discourse markers (4249)	
you know (196)	*monitoring* (180), *quoting* (3), *monitoring-specification* (3), *monitoring-closing boundary* (2), *monitoring-quoting* (2), *monitoring-topic-shift* (1), *reformulation* (1), *monitoring-punctuation* (1), *monitoring-reformulation* (1), *face-saving* (1), *face-saving-monitoring* (1)
if (195)	*condition* (132), *relevance* (55), *motivation* (3), *concession* (2), *temporal* (1), *cause-relevance* (1), *contrast-motivation* (1)
because (190)	*cause* (98), *motivation* (89), *specification* (1), *topic-resuming-motivation* (1), *condition* (1), *motivation-specification* (1)
I mean (174)	*specification* (64), *reformulation* (41), *punctuation* (26), *opening boundary* (11), *conclusion* (5), *comment* (4), *topic-resuming* (4), *motivation* (4), *emphasis* (3), *punctuation-specification* (2), *reformulation-specification* (2), *punctuation-reformulation* (2), *face-saving-reformulation* (1), *motivation-specification* (1), *addition* (1), *topic-shift-reformulation* (1), *face-saving* (1), *punctuation-motivation* (1)
when (129)	*temporal* (120), *relevance* (3), *cause* (3), *condition* (2), *concession* (1)
actually (97)	*specification* (24), *opposition* (20), *comment* (18), *emphasis* (9), *concession* (7), *reformulation* (3), *punctuation* (2), *comment-opposition* (2), *topic-shift* (2), *emphasis-opposition* (2), *opposition-specification* (1), *consequence* (1), *comment-specification* (1), *face-saving* (1), *closing boundary-specification* (1), *disagreeing* (1), *alternative* (1), *disagreeing-specification* (1)
then (94)	*conclusion* (36), *consequence* (25), *enumeration* (6), *topic-shift* (5), *topic-resuming* (5), *temporal* (4), *topic-shift-conclusion* (4), *contrast* (2), *emphasis* (2), *closing boundary-conclusion* (1), *punctuation-consequence* (1), *punctuation-conclusion* (1), *specification* (1), *opposition* (1)
and then (70)	*temporal* (41), *addition* (12), *enumeration* (10), *consequence* (3), *topic-shift* (1), *temporal-concession* (1), *concession* (1), *contrast-enumeration* (1)
or (65)	*alternative* (45), *reformulation* (15), *alternative-enumeration* (2), *alternative-punctuation* (1), *alternative-closing boundary* (1), *alternative-ellipsis* (1)
sort of (60)	*approximation* (53), *punctuation* (2), *emphasis* (2), *punctuation-approximation* (1), *face-saving* (1), *face-saving-approximation* (1)
now (40)	*topic-shift* (12), *addition* (8), *topic-resuming* (7), *opening boundary* (4), *punctuation* (2), *opposition* (2), *conclusion* (1), *comment* (1), *closing boundary* (1), *contrast* (1), *contrast-enumeration* (1)
as (32)	*temporal* (22), *cause* (4), *cause-temporal* (3), *motivation* (2), *condition* (1)
right (31)	*monitoring, agreeing* (6), *closing boundary* (4), *quoting* (2), *opening boundary* (1), *agreeing-punctuation* (1), *monitoring-closing boundary* (1)

English discourse markers (4249)	
kind of (31)	*approximation* (26), *face-saving* (2), *punctuation-approximation* (2), *approximation-specification* (1)
though (30)	*opposition* (13), *concession* (12), *contrast* (4), *topic-shift-opposition* (1)
in fact (29)	*specification* (8), *comment* (7), *reformulation* (3), *opposition* (3), *emphasis* (3), *topic-shift* (1), *comment-motivation* (1), *topic-shift-conclusion* (1), *concession* (1), *addition* (1)
yeah (27)	*agreeing* (13), *monitoring* (3), *topic-resuming* (3), *agreeing-topic-resuming* (3), *agreeing-closing boundary* (2), *agreeing-opening boundary* (2), *closing boundary* (1)
okay (34)	*monitoring* (18), *agreeing* (3), *monitoring-closing boundary* (7), *agreeing-closing boundary* (2), *closing boundary* (2), *opening boundary* (1), *agreeing-topic-resuming* (1)
and so on (19)	*ellipsis* (14), *closing boundary* (4), *ellipsis-closing boundary* (1)
like (16)	*approximation* (9), *specification* (3), *face-saving-approximation* (2), *punctuation* (2)
although (16)	*concession* (15), *opposition* (1)
therefore (16)	*consequence* (12), *conclusion* (4)
for example (16)	*specification* (16)
anyway (15)	*topic-resuming* (8), *closing boundary* (3), *topic-shift* (1), *emphasis* (1), *reformulation* (1), *opposition* (1)
so that (14)	*consequence* (12), *temporal* (1), *conclusion* (1)
indeed (13)	*motivation* (3), *specification* (3), *comment* (2), *opposition* (1), *reformulation* (1), *comment-specification* (1), *agreeing* (1), *emphasis* (1)
yes (13)	*agreeing-topic-resuming* (4), *topic-resuming* (3), *agreeing* (3), *monitoring* (1), *agreeing-opening boundary* (1), *closing boundary* (1)
if you like (12)	*approximation* (7), *face-saving* (2), *monitoring* (2), *reformulation* (1)
since (11)	*temporal* (8), *cause* (3)
before (11)	*temporal* (11)
while (10)	*concession* (6), *temporal* (2), *contrast-temporal* (1), *contrast* (1)
even if (9)	*concession* (5), *relevance* (3), *opposition* (1)
unless (9)	*exception* (8), *alternative* (1)
oh (9)	*quoting* (9)
you see (8)	*monitoring* (7), *monitoring-specification* (1)
for instance (8)	*specification* (8)
after (8)	*temporal* (8)
once (8)	*temporal*
say (8)	*specification* (6), *approximation* (2)
however (7)	*opposition* (3), *contrast* (3), *concession* (1)
until (7)	*temporal* (7)
whereas (7)	*contrast* (6), *opposition* (1)
for (6)	*cause* (5), *motivation* (1)
etcetera (6)	*ellipsis* (4), *ellipsis-approximation* (1), *closing boundary* (1)
meanwhile (6)	*temporal* (3), *temporal-topic-shift* (2), *topic-shift* (1)

English discourse markers (4249)	
in other words (4)	*reformulation* (4)
yet (4)	*concession*
look (4)	*quoting* (2), *face-saving* (2)
as soon as (4)	*temporal* (4)
by the way (4)	*topic-shift* (2), *comment* (2)
alright (4)	*monitoring* (4)
whilst (3)	*contrast* (2), *temporal* (1)
either (3)	*alternative* (2), *alternative-enumeration* (1)
first (3)	*enumeration* (3)
and things (3)	*ellipsis* (3)
as it were (3)	*approximation* (3)
otherwise (3)	*alternative* (3)
nevertheless (3)	*concession* (2), *topic-shift-opposition* (1)
see (3)	*monitoring* (3)
no (3)	*agreeing* (2), *disagreeing-topic-resuming* (1)
listen (2)	*monitoring* (2)
even though (2)	*concession* (2)
as long as (2)	*temporal* (1), *condition* (1)
I suppose (2)	*approximation* (1), *agreeing* (1)
plus (2)	*addition* (2)
provided (2)	*condition* (2)
first of all (2)	*enumeration* (2)
or something (2)	*approximation* (2)
having said that (2)	*opposition* (2)
in addition (1)	*addition* (1)
finally (1)	*enumeration* (1)
where (1)	*contrast* (1)
considering (1)	*motivation* (1)
but then (1)	*opposition* (1)
second (1)	*enumeration* (1)
whenever (1)	*temporal* (1)
and that kind of stuff (1)	*ellipsis* (1)
only (1)	*exception* (1)
secondly (1)	*enumeration* (1)
I don't know (1)	*specification* (1)
insofar as (1)	*motivation* (1)
after all (1)	*specification* (1)
on the other hand (1)	*opposition* (1)
and still (1)	*concession* (1)
albeit (1)	*concession* (1)
till (1)	*temporal* (1)
instead (1)	*alternative* (1)

French discourse markers (4494)	
et (869)	*addition* (498), *topic-shift* (102), *specification* (78), *consequence* (44), *temporal* (24), *punctuation* (23), *contrast* (18), *concession* (15), *topic-resuming* (13), *enumeration* (13), *conclusion* (12), *comment* (10), *opening boundary* (6), *emphasis* (6), *opposition* (3), *quoting* (1), *alternative* (1), *addition-opposition* (1), *ellipsis* (1)
mais (540)	*opposition* (235), *concession* (107), *contrast* (25), *opening boundary* (25), *topic-resuming* (21), *punctuation* (20), *topic-shift* (19), *emphasis* (12), *specification* (9), *quoting* (8), *addition* (7), *topic-resuming-opposition* (6), *opening boundary-opposition* (6), *disagreeing* (5), *opening boundary-emphasis* (5), *closing boundary* (3), *reformulation* (3), *disagreeing-opening boundary* (2), *topic-shift-opposition* (2), *disagreeing-opposition* (2), *addition-opposition* (2), *comment* (2), *opening boundary-specification* (2), *punctuation-opposition* (2), *closing boundary-opposition* (2), *opening boundary-disagreeing* (1), *enumeration-opposition* (1), *motivation-opposition* (1), *quoting-opposition* (1), *motivation* (1), *agreeing* (1), *addition-opening boundary* (1), *opposition-specification* (1)
donc (291)	*conclusion* (146), *consequence* (55), *specification* (25), *topic-resuming* (20), *topic-resuming-conclusion* (6), *closing boundary* (6), *reformulation* (3), *punctuation-specification* (3), *monitoring* (3), *emphasis* (3), *monitoring-conclusion* (3), *consequence-specification* (2), *punctuation* (2), *addition-conclusion* (2), *punctuation-emphasis* (2), *face-saving* (1), *reformulation-specification* (1), *consequence-topic-resuming* (1), *opposition* (1), *enumeration* (1), *topic-shift-conclusion* (1), *opening boundary-specification* (1), *ellipsis-conclusion* (1), *punctuation-topic-resuming* (1), *opening boundary* (1)
alors (271)	*consequence* (62), *specification* (42), *conclusion* (42), *opening boundary* (40), *emphasis* (16), *topic-shift* (12), *addition* (12), *topic-resuming* (7), *temporal* (7), *punctuation* (6), *enumeration* (4), *comment* (2), *opposition* (2), *opening boundary-conclusion* (2), *quoting-comment* (1), *addition-conclusion* (1), *face-saving-opposition* (1), *topic-shift-opposition* (1), *monitoring-conclusion* (1), *closing-conclusion* (1), *reformulation* (1), *opening boundary-specification* (1), *topic-resuming-specification* (1), *contrast* (1), *face-saving-specification* (1), *face-saving* (1), *addition-specification* (1), *punctuation-comment* (1), *opening boundary-consequence* (1)
hein (260)	*monitoring* (256), *face-saving* (2), *disagreeing* (1), *ellipsis* (1)
quoi (239)	*monitoring* (112), *closing boundary* (46), *punctuation* (21), *conclusion* (18), *face-saving* (15), *monitoring-conclusion* (5), *closing boundary-conclusion* (5), *reformulation* (4), *approximation* (3), *disagreeing* (2), *monitoring-closing boundary* (2), *punctuation-motivation* (1), *specification* (1), *comment* (1), *motivation* (1), *monitoring-approximation* (1), *closing boundary-approximation* (1)

French discourse markers (4494)	
parce que (216)	*motivation* (113), *cause* (94), *opening boundary* (1), *opening boundary-specification* (1), *comment-motivation* (1), *opening boundary-motivation* (1), *emphasis* (1), *topic-resuming-motivation* (1), *specification* (1), *addition* (1), *motivation-specification* (1)
ben (183)	*opening boundary* (85), *punctuation* (41), *quoting* (10), *emphasis* (9), *disagreeing* (8), *specification* (6), *consequence* (6), *agreeing* (3), *opposition* (2), *topic-resuming* (2), *disagreeing-opening boundary* (1), *comment* (1), *consequence-quoting* (1), *closing boundary-conclusion* (1), *concession* (1), *reformulation* (1), *opening boundary-specification* (1), *opening boundary-emphasis* (1), *topic-shift* (1), *approximation* (1), *motivation* (1)
enfin (157)	*reformulation* (99), *conclusion* (11), *specification* (9), *opposition* (7), *emphasis* (7), *closing boundary* (5), *face-saving* (3), *topic-resuming* (2), *approximation* (2), *disagreeing* (1), *concession* (1), *topic-shift* (1), *comment* (1), *punctuation* (1), *enumeration* (1), *enumeration-topic-shift* (1), *topic-resuming-conclusion* (1), *ellipsis* (1), *approximation-reformulation* (1), *closing boundary-reformulation* (1), *motivation* (1)
quand (133)	*temporal* (116), *relevance* (8), *condition* (4), *motivation* (2), *specification* (1), *cause* (1), *opposition* (1)
si (119)	*condition* (80), *relevance* (34), *cause* (2), *concession* (1), *temporal* (1), *motivation* (1)
bon (98)	*punctuation* (38), *closing boundary* (14), *face-saving* (7), *topic-resuming* (7), *opening boundary* (6), *opposition* (4), *quoting* (3), *agreeing* (3), *topic-shift* (3), *specification* (2), *agreeing-punctuation* (2), *reformulation* (2), *face-saving-opposition* (1), *agreeing-closing boundary* (1), *opening boundary-opposition* (1), *conclusion* (1), *emphasis* (1), *face-saving-punctuation* (1), *face-saving-specification* (1)
et puis (97)	*temporal* (40), *addition* (32), *topic-shift* (11), *enumeration* (7), *consequence* (2), *specification* (2), *conclusion* (1), *contrast-enumeration* (1), *opposition* (1)
tu vois (58)	*monitoring* (53), *face-saving* (2), *opening boundary* (1), *specification* (1), *monitoring-specification* (1)
voilà (50)	*closing boundary* (39), *agreeing* (2), *punctuation* (2), *emphasis* (2), *quoting* (1), *topic-resuming* (1), *conclusion* (1), *opening boundary* (1), *face-saving* (1)
en fait (45)	*specification* (18), *emphasis* (5), *opposition* (4), *reformulation* (3), *topic-shift* (2), *punctuation-specification* (2), *comment* (2), *concession* (2), *emphasis-specification* (1), *concession-specification* (1), *topic-resuming-specification* (1), *punctuation* (1), *motivation-opposition* (1), *opening boundary* (1), *opposition-specification* (1)
par exemple (43)	*specification* (43), *topic-shift* (2), *closing boundary-specification* (1)
etcetera (41)	*ellipsis* (35), *approximation* (2), *closing boundary* (2), *monitoring* (1), *face-saving* (1)
puisque (32)	*motivation* (20), *cause* (12)

French discourse markers (4494)	
d'ailleurs (32)	*comment* (24), *specification* (3), *topic-shift* (2), *comment-specification* (1), *opposition* (1), *emphasis* (1)
bon ben (31)	*punctuation* (18), *opening boundary* (6), *face-saving* (2), *opposition* (1), *conclusion* (1), *emphasis* (1), *ellipsis-conclusion* (1), *opening boundary-consequence* (1)
eh bien (30)	*punctuation* (16), *opening boundary* (6), *topic-resuming* (2), *consequence* (2), *punctuation-conclusion* (1), *disagreeing* (1), *conclusion* (1), *emphasis* (1)
oui (30)	*agreeing* (28), *monitoring* (1), *agreeing-opening* (1)
puis (26)	*temporal* (12), *addition* (6), *enumeration* (3), *specification* (2), *consequence* (1), *topic-shift* (1), *temporal-addition* (1)
ou (26)	*alternative* (20), *reformulation* (6)
je veux dire (26)	*reformulation* (12), *specification* (8), *approximation* (3), *punctuation* (1), *emphasis* (1), *face-saving* (1)
et tout ça (25)	*ellipsis* (25)
alors que (25)	*temporal* (8), *concession* (6), *concession-temporal* (4), *contrast* (4), *opposition* (2), *consequence* (1)
comme (21)	*cause* (17), *motivation* (4)
eh ben (21)	*opening boundary* (8), *punctuation* (6), *topic-shift* (2), *topic-resuming* (2), *conclusion* (1), *agreeing-opening boundary* (1), *specification* (1)
écoutez (19)	*monitoring* (16), *quoting* (1), *face-saving* (1), *face-saving-quoting* (1)
au fond (17)	*specification* (8), *conclusion* (3), *opposition* (2), *comment* (1), *reformulation* (1), *consequence* (1), *emphasis* (1)&
je dirais (16)	*approximation* (15), *face-saving* (1)
voilà quoi (15)	*closing boundary* (12), *ellipsis* (1), *face-saving* (1), *consequence* (1)
vous savez (15)	*monitoring* (13), *monitoring-topic-shift* (1), *face-saving* (1)
c'est-à-dire (14)	*specification* (10), *conclusion* (2), *reformulation* (2)
car (14)	*motivation* (7), *cause* (7)
pour que (11)	*consequence* (11)
ouais (11)	*agreeing* (10), *monitoring* (1)
je vais dire (11)	*approximation* (10), *approximation-reformulation* (1)
ou bien (11)	*alternative* (11)
tandis que (10)	*contrast* (8), *temporal* (2)
pourtant (10)	*concession* (9), *opposition* (1)
c'est-à-dire que (10)	*specification* (6), *emphasis* (2), *opening boundary-reformulation* (1), *face-saving-emphasis* (1)
enfin bon (9)	*closing boundary* (4), *opposition* (3), *topic-shift-opposition* (1), *conclusion* (1)
dès que (9)	*temporal* (9)
par contre (9)	*opposition* (5), *addition-opposition* (2), *emphasis* (1), *contrast* (1)
disons (9)	*approximation* (4), *topic-resuming* (1), *reformulation* (1), *specification* (1°, *emphasis* (1), *punctuation-approximation* (1)
d'abord (9)	*enumeration* (9)

French discourse markers (4494)	
non (9)	*monitoring* (4), *topic-resuming* (2), *disagreeing-topic-resuming* (2), *agreeing-topic-resuming* (1)
sinon (8)	*alternative* (5), *exception* (2), *topic-shift* (1)
tiens (8)	*quoting* (7), *topic-shift* (1)
même si (8)	*concession* (8)
soit (7)	*alternative* (7)
si tu veux (7)	*monitoring-approximation* (4), *approximation* (2), *monitoring* (1)
lorsque (7)	*temporal* (6), *cause* (1)
okay (8)	*agreeing* (8)
bien (7)	*topic-shift* (2), *opening boundary* (2), *topic-resuming* (1), *closing boundary* (1), *quoting* (1)
tu sais (6)	*monitoring* (6)
et tout (6)	*ellipsis* (5), *ellipsis-approximation* (1)
tout ça (5)	*ellipsis* (5)
à ce moment-là (5)	*consequence* (3), *exception* (1), *conclusion* (1)
en conséquence (5)	*consequence* (5)
maintenant (5)	*opposition* (2), *topic-shift* (1), *concession* (1), *enumeration* (1)
d'accord (5)	*monitoring* (4), *agreeing-closing boundary* (1)
si vous voulez (5)	*approximation* (3), *reformulation* (1), *monitoring-specification* (1)
entre guillemets (5)	*approximation* (4), *emphasis* (1)
en tout cas (5)	*reformulation* (4), *emphasis* (1)
après (4)	*enumeration* (2), *topic-shift* (1), *opposition* (1)
écoute (4)	*monitoring* (3), *face-saving* (1)
autrement (3)	*exception* (2), *alternative* (1)
à ce propos (3)	*comment* (2), *topic-shift* (1)
un (3)	*enumeration* (3)
en effet (3)	*specification* (3)
du coup (3)	*consequence* (3)
en plus (3)	*addition* (2), *enumeration* (1)
vu que (3)	*motivation* (2), *cause* (1)
mais bon (3)	*opposition* (2), *punctuation* (1)
savez (3)	*monitoring* (2), *specification* (1)
vous voyez (3)	*monitoring* (3)
du moins (3)	*emphasis* (2), *reformulation* (1)
ainsi (3)	*specification* (2), *consequence* (1)
or (3)	*concession* (3)
seulement (2)	*opposition* (1), *concession* (1)
quoique (2)	*reformulation* (1), *concession* (1)
bien que (2)	*opposition* (1), *concession* (1)
ah (2)	*quoting* (2)
ou sinon (2)	*exception* (1), *alternative* (1)

French discourse markers (4494)	
d'un autre côté (2)	*opposition* (1), *contrast* (1)
enfin bref (2)	*ellipsis* (1), *closing boundary* (1)
encore que (2)	*reformulation* (2)
ou quoi (2)	*ellipsis* (2)
déjà (2)	*enumeration* (1), *comment* (1)
sauf que (2)	*exception* (2)
ça va (2)	*agreeing* (2)
voyez (2)	*monitoring* (2)
tant que (2)	*temporal* (1), *condition-temporal* (1)
ou alors (2)	*alternative* (1)
m'enfin (2)	*opposition* (2)
comme ça (1)	*approximation* (1)
d'autre part (1)	*topic-shift* (1)
de un (1)	*enumeration* (1)
au contraire (1)	*opposition* (1)
bref (1)	*topic-resuming-conclusion* (1)
du reste (1)	*comment* (1)
dès lors (1)	*temporal* (1)
du temps où (1)	*temporal* (1)
genre (1)	*specification* (1)
maintenant que (1)	*cause-temporal* (1)
de même (1)	*addition* (1)
à part cela (1)	*topic-shift* (1)
par conséquent (1)	*consequence* (1)
dis (1)	*monitoring* (1)
on va dire (1)	*approximation* (1)
quand même (1)	*emphasis* (1)
à propos (1)	*topic-shift* (1)
mais enfin (1)	*opposition* (1)
deuxièment (1)	*enumeration* (1)
depuis que (1)	*temporal* (1)
cependant (1)	*opposition* (1)
de sorte que (1)	*consequence* (1)
bah (1)	*agreeing* (1)
néanmoins (1)	*néanmoins* (1)
ouais ouais (1)	*agreeing* (1)
boh (1)	*opening boundary* (1)
par ailleurs (1)	*topic-shift* (1)
autrement dit (1)	*reformulation* (1)
du moment que (1)	*condition* (1)
effectivement (1)	*comment* (1)

APPENDIX 3

List of functions in *DisFrEn* and their discourse markers

	English DMs	French DMs
Sequential functions (2680)		
Addition (1238)	*and* (651), *and then* (12), *now* (8), *so* (2), *plus* (2), *in fact* (1), *but* (1), *in addition* (1), *I mean* (1)	*et* (498), *et puis* (32), *alors* (12), *mais* (7), *puis* (6), *en plus* (2), *de même* (1), *parce que* (1)
Opening boundary (404)	*well* (177), *and* (13), *I mean* (11), *but* (6), *now* (4), *right* (1), *so* (1), *okay* (1)	*ben* (85), *alors* (40), *mais* (25), *eh ben* (8), *bon* (6), *et* (6), *bon ben* (6), *eh bien* (6), *bien* (2), *parce que* (1), *donc* (1), *tu vois* (1), *voilà* (1), *boh* (1), *en fait* (1)
Punctuation (284)	*I mean* (26), *and* (24), *well* (20), *but* (6), *so* (3), *sort of* (2), *like* (2), *now* (2), *actually* (2)	*ben* (41), *bon* (38), *et* (23), *quoi* (21), *mais* (20), *bon ben* (18), *eh bien* (16), *alors* (6), *eh ben* (6), *voilà* (2), *donc* (2), *mais bon* (1), *je veux dire* (1), *en fait* (1), *enfin* (1)
Topic-shift (271)	*and* (41), *so* (17), *now* (12), *well* (10), *but* (9), *then* (5), *actually* (2), *by the way* (2), *in fact* (1), *and then* (1), *meanwhile* (1), *anyway* (1)	*et* (102), *mais* (19), *alors* (12), *et puis* (11), *bon* (3), *par exemple* (2), *en fait* (2), *bien* (2), *d'ailleurs* (2), *eh ben* (2), *ben* (1), *sinon* (1), *après* (1), *à propos* (1), *à ce propos* (1), *à part cela* (1), *par ailleurs* (1), *puis* (1), *maintenant* (1), *d'autre part* (1), *tiens* (1), *enfin* (1)
Topic-resuming (167)	*but* (22), *and* (16), *so* (16), *anyway* (8), *now* (7), *then* (5), *I mean* (4), *yes* (3), *yeah* (3), *well* (2)	*mais* (21), *donc* (20), *et* (13), *bon* (7), *alors* (7), *ben* (2), *eh ben* (2), *non* (2), *eh bien* (2), *enfin* (2), *disons* (1), *voilà* (1), *bien* (1)
Closing boundary (166)	*but* (9), *so* (7), *and so on* (4), *right* (4), *anyway* (3), *okay* (2), *etcetera* (1), *yeah* (1), *yes* (1), *now* (1)	*quoi* (46), *voilà* (39), *bon* (14), *voilà quoi* (12), *donc* (6), *enfin* (5), *enfin bon* (4), *mais* (3), *bien* (1), *okay* (2), *enfin bref* (1)
Enumeration (83)	*and then* (10), *and* (10), *then* (6), *first* (3), *first of all* (2), *secondly* (1), *finally* (1), *second* (1), *so* (1)	*et* (13), *d'abord* (9), *et puis* (7), *alors* (4), *puis* (3), *un* (3), *après* (2), *donc* (1), *déjà* (1), *deuxièmement* (1), *en plus* (1), *de un* (1), *enfin* (1), *maintenant* (1)
Quoting (66)	*well* (12), *oh* (9), *and* (4), *you know* (3), *right* (2), *look* (2)	*ben* (10), *mais* (8), *tiens* (7), *bon* (3), *ah* (2), *bien* (1), *et* (1), *écoutez* (1), *voilà* (1)
Emphasis	*well* (1)	

	English DMs	French DMs
Rhetorical functions (2650)		
Specification (626)	*and* (180), *I mean* (64), *so* (25), *actually* (24), *for example* (16), *for instance* (8), *in fact* (8), *say* (6), *like* (3), *well* (3), *indeed* (3), *but* (1), *I don't know* (1), *after all* (1), *then* (1), *because* (1)	*et* (78), *alors* (42), *par exemple* (40), *donc* (25), *en fait* (18), *c'est-à-dire* (10), *mais* (9), *enfin* (9), *au fond* (8), *je veux dire* (8), *ben* (6), *c'est-à-dire que* (6), *en effet* (3), *d'ailleurs* (3), *puis* (2), *bon* (2), *ainsi* (2), *et puis* (2), *savez* (1), *tu vois* (1), *eh ben* (1), *genre* (1), *parce que* (1), *disons* (1), *quoi* (1)
Opposition (546)	*but* (203), *actually* (20), *though* (13), *in fact* (3), *and* (3), *however* (3), *now* (2), *having said that* (2), *but then* (1), *then* (1), *anyway* (1), *even if* (1), *whereas* (1), *on the other hand* (1), *although* (1), *indeed* (1)	*mais* (235), *enfin* (7), *par contre* (5), *bon* (4), *en fait* (4), *et* (3), *enfin bon* (3), *mais bon* (2), *maintenant* (2), *alors* (2), *m'enfin* (2), *alors que* (2), *au fond* (2), *ben* (2), *après* (1), *et puis* (1), *quand* (1), *cependant* (1), *au contraire* (1), *d'ailleurs* (1), *néanmoins* (1), *pourtant* (1), *donc* (1), *seulement* (1), *bien que* (1), *bon ben* (1), *d'un autre côté* (1), *mais enfin* (1)
Conclusion (510)	*so* (198), *then* (36), *and* (20), *I mean* (5), *therefore* (4), *well* (3), *so that* (1), *now* (1)	*donc* (146), *alors* (42), *quoi* (18), *et* (12), *enfin* (11), *au fond* (3), *c'est-à-dire* (2), *eh bien* (1), *bon ben* (1), *bon* (1), *à ce moment-là* (1), *et puis* (1), *voilà* (1), *eh ben* (1), *enfin bon* (1)
Motivation (259)	*because* (89), *I mean* (4), *indeed* (3), *if* (3), *as* (2), *and* (1), *insofar as* (1), *so* (1), *for* (1), *considering* (1)	*parce que* (113), *puisque* (20), *car* (7), *comme* (4), *quand* (2), *vu que* (2), *si* (1), *enfin* (1), *quoi* (1), *ben* (1), *mais* (1)
Reformulation (248)	*I mean* (41), *well* (26), *or* (15), *in other words* (4), *so* (3), *in fact* (3), *actually* (3), *but* (1), *indeed* (1), *if you like* (1), *you know* (1), *anyway* (1)	*enfin* (99), *je veux dire* (12), *ou* (6), *en tout cas* (4), *quoi* (4), *en fait* (3), *mais* (3), *donc* (3), *bon* (2), *encore que* (2), *c'est-à-dire* (2), *au fond* (1), *ben* (1), *si vous voulez* (1), *du moins* (1), *quoique* (1), *autrement dit* (1), *alors* (1), *disons* (1)
Approximation (154)	*sort of* (53), *kind of* (26), *like* (9), *if you like* (7), *as it were* (3), *say* (2), *or something* (2), *I suppose* (1)	*je dirais* (15), *je vais dire* (10), *entre guillemets* (4), *disons* (4), *quoi* (3), *si vous voulez* (3), *je veux dire* (3), *si tu veux* (2), *etcetera* (2), *enfin* (2), *on va dire* (1), *ben* (1), *comme ça* (1)
Emphasis (108)	*actually* (9), *and* (6), *well* (3), *I mean* (3), *in fact* (3), *sort of* (2), *then* (2), *so* (1), *anyway* (1), *indeed* (1), *but* (1)	*alors* (16), *mais* (12), *ben* (9), *enfin* (7), *et* (6), *en fait* (5), *donc* (3), *c'est-à-dire que* (2), *voilà* (2), *du moins* (2), *bon ben* (1), *parce que* (1), *disons* (1), *je veux dire* (1), *eh bien* (1), *par contre* (1), *d'ailleurs* (1), *quand même* (1), *au fond* (1), *entre guillemets* (1), *bon* (1), *en tout cas* (1)

	English DMs	French DMs
Relevance (103)	*if* (55), *when* (3), *even if* (3)	*si* (34), *quand* (8)
Comment (96)	*actually* (18), *and* (9), *in fact* (7), *I mean* (4), *well* (4), *by the way* (2), *indeed* (2), *now* (1)	*d'ailleurs* (24), *et* (10), *mais* (2), *à ce propos* (2), *en fait* (2), *alors* (2), *quoi* (1), *au fond* (1), *ben* (1), *déjà* (1), *enfin* (1), *du reste* (1), *effectivement* (1)
Ideational functions (2064)		
Temporal (500)	*when* (120), *and then* (41), *and* (27), *as* (22), *before* (11), *since* (8), *after* (8), *once* (8), *until* (7), *then* (4), *as soon as* (4), *meanwhile* (3), *while* (2), *so that* (1), *if* (1), *whenever* (1), *whilst* (1), *as long as* (1), *till* (1)	*quand* (116), *et puis* (40), *et* (24), *puis* (12), *dès que* (9), *alors que* (8), *alors* (7), *lorsque* (6), *tandis que* (2), *dès lors* (1), *depuis que* (1), *si* (1), *du temps où* (1), *tant que* (1)
Consequence (478)	*so* (123), *and* (101), *then* (25), *so that* (12), *therefore* (12), *and then* (3), *actually* (1),	*alors* (62), *donc* (55), *et* (44), *pour que* (11), *ben* (6), *en conséquence* (5), *du coup* (3), *à ce moment-là* (3), *et puis* (2), *eh bien* (2), *ainsi* (1), *alors que* (1), *par conséquent* (1), *de sorte que* (1), *puis* (1), *voilà quoi* (1), *si* (1), *au fond* (1)
Concession (368)	*but* (142), *although* (15), *though* (12), *and* (8), *actually* (7), *while* (6), *even if* (5), *yet* (4), *nevertheless* (2), *if* (2), *even though* (2), *in fact* (1), *and still* (1), *and then* (1), *when* (1), *albeit* (1), *however* (1)	*mais* (107), *et* (15), *pourtant* (9), *même si* (8), *alors que* (6), *or* (3), *en fait* (2), *seulement* (1), *enfin* (1), *quoique* (1), *si* (1), *ben* (1), *bien que* (1), *maintenant* (1)
Cause (247)	*because* (97), *for* (5), *as* (4), *since* (3), *when* (3)	*parce que* (94), *comme* (17), *puisque* (12), *car* (7), *si* (2), *vu que* (1), *lorsque* (1), *quand* (1)
Condition (223)	*if* (132), *provided* (2), *when* (2), *because* (1), *as long as* (1), *as* (1)	*si* (79), *quand* (4), *du moment que* (1)
Contrast (129)	*but* (38), *and* (13), *whereas* (6), *though* (4), *however* (3), *whilst* (2), *then* (2), *where* (1), *while* (1), *now* (1)	*mais* (25), *et* (18), *tandis que* (8), *alors que* (4), *d'un autre côté* (1), *par contre* (1), *alors* (1)
Alternative (101)	*or* (45), *otherwise* (3), *either* (2), *instead* (1), *unless* (1), *actually* (1)	*ou* (20), *ou bien* (11), *soit* (7), *sinon* (5), *ou alors* (2), *ou sinon* (1), *autrement* (1), *et* (1)
Exception (18)	*unless* (8), *but* (1), *only* (1)	*sauf que* (2), *autrement* (2), *sinon* (2), *ou sinon* (1), *à ce moment-là* (1)
Interpersonal functions (999)		
Monitoring (718)	*you know* (180), *okay* (18), *right* (16), *you see* (7), *alright* (4), *yeah* (3), *see* (3), *if you like* (2), *listen* (2), *yes* (1)	*hein* (256), *quoi* (112), itu vois *(53),* *écoutez* (56), *vous savez* (13), *tu sais* (6), *non* (4), *d'accord* (4), *écoute* (3), *donc* (3), *vous voyez* (3), *voyez* (2), *savez* (2), *si tu veux* (1), *etcetera* (1), *dis* (1), *ouais* (1), *oui* (1)

	English DMs	French DMs
Ellipsis (99)	*and so on* (14), *etcetera* (4), *and things* (3), *and that kind of stuff* (1)	*etcetera* (35), *et tout ça* (25), *tout ça* (5), *et tout* (5), *ou quoi* (2), *hein* (1), *enfin bref* (1), *voilà quoi* (1), *et* (1), *enfin* (1)
Agreeing (94)	*yeah* (13), *well* (6), *right* (6), *yes* (3), *okay* (3), *no* (2), *I suppose* (1), *indeed* (1)	*oui* (28), *ouais* (10), *okay* (8), *bon* (3), *ben* (3), *voilà* (2), *ça va* (2), *ouais ouais* (1), *bah* (1), *mais* (1)
Face-saving (53)	*look* (2), *well* (2), *kind of* (2), *if you like* (2), *sort of* (1), *actually* (1), *you know* (1), *I mean* (1)	*quoi* (15), *bon* (7), *enfin* (3), *tu vois* (2), *bon ben* (2), *hein* (2), *vous savez* (1), *voilà* (1), *je dirais* (1), *etcetera* (1), *je veux dire* (1), *voilà quoi* (1), *écoute* (1), *donc* (1), *écoutez* (1), *alors* (1)
Disagreeing (35)	*well* (15), *actually* (1), *but* (1)	*ben* (8), *mais* (5), *quoi* (2), *hein* (1), *eh bien* (1), *enfin* (1)

This table is restricted to the thirty single-tagged functions in the taxonomy, leaving out 107 different double-tagged functions which only amount to 350 DM occurrences and are the least frequent function types (except for *exception*) overall in *DisFrEn*. Double tags were included in Appendix 2 so that the information is not lost.

APPENDIX 4

Top-five most frequent functions by register in *DisFrEn*

	1st	2nd	3rd	4th	5th
English					
convers.	*addition*	*opening*	*opposition*	*monitor.*	*conclu.*
phone	*opening*	*addition*	*opposition*	*monitor.*	*conclu.*
interview	*specif.*	*addition*	*conseq.*	*conclu.*	*monitor.*
radio	*addition*	*specif.*	*opposition*	*temporal*	*motivation*
classroom	*addition*	*conclu.*	*temporal*	*monitor.*	*opposition*
sports	*addition*	*conseq.*	*temporal*	*concess.*	*opposition*
political	*addition*	*temporal*	*concess.*	*opposition*	*condition*
news	*addition*	*concess.*	*temporal*	*opposition*	*topic-shift*
French					
convers.	*monitor.*	*addition*	*specif.*	*opposition*	*reformul.*
phone	*opening*	*monitor.*	*punctuation*	*conclu.*	*addition*
interview	*addition*	*monitor.*	*specif.*	*opposition*	*temporal*
radio	*addition*	*specif.*	*opposition*	*motivation*	*monitor.*
classroom	*addition*	*conseq.*	*monitor.*	*punctuation*	*closing*
sports	*addition*	*monitor.*	*opposition*	*conseq.*	*concess.*
political	*addition*	*condition*	*conseq.*	*cause*	*temporal*
news	*addition*	*concess.*	*opposition*	*topic-shift*	*specif.*
Total					
convers.	*monitor.*	*addition*	*specif.*	*opposition*	*opening*
phone	*opening*	*monitor.*	*addition*	*conclu.*	*opposition*
interview	*addition*	*specif.*	*monitor.*	*conseq.*	*conclusion*
radio	*addition*	*specif.*	*opposition*	*motivation*	*monitor.*
classroom	*addition*	*conclu.*	*monitor.*	*temporal*	*conseq.*
sports	*addition*	*conseq.*	*temporal*	*concess.*	*opposition*
political	*addition*	*temporal*	*condition*	*conseq.*	*concess.*
news	*addition*	*concess.*	*opposition*	*temporal*	*topic-shift*

Index

Spoken language is characterized by the occurrence of linguistic devices such as discourse markers (e.g. *so*, *well*, *you know*, *I mean*) and other so-called "disfluent" phenomena, which reflect the temporal nature of the cognitive mechanisms underlying speech production and comprehension. The purpose of this book is to distinguish between strategic vs. symptomatic uses of these markers on the basis of their combination, function and distribution across several registers in English and French. Through deep quantitative and qualitative analyses of manually annotated features in the new DisFrEn corpus, this usage-based study provides (i) an exhaustive portrait of discourse markers in English and French and (ii) a scale of (dis)fluency against which different configurations of discourse markers can be diagnosed as rather fluent or disfluent. By bringing together discourse markers and (dis)fluency under one coherent framework, this book is a unique contribution to corpus-based pragmatics, discourse analysis and crosslinguistic fluency research.

"Without any doubt, with DisFrEn, Ludivine Crible developed the largest, most diverse, and most inclusive spoken corpus annotated at the discourse level. This makes it an invaluable resource for further research, both for theoretical work and (computational) applications. The design of the annotation scheme, which is a topic of continuing research, is a crucial contribution to the field. The detailed way in which she motivates all her methodological decisions and procedures leaves the reader with a very transparent piece of work, making it possible to either replicate some of the studies, or to compare it with related work."

Liesbeth Degand, *Université Catholique de Louvain*

"This excellent and innovative book offers a wealth of new data about discourse markers in spoken French and English. It will appeal to all researchers interested in discourse phenomena, corpus linguistics and cross-linguistic studies. I highly recommend it."

Sandrine Zufferey, *Universität Bern*

ISBN 978 90 272 0046 4

John Benjamins Publishing Company