



Secteur des Sciences Humaines
Faculté de Philosophie, Arts et Lettres

Discourse Markers and (Dis)fluency across Registers

A Contrastive Usage-Based Study in English and French

Ludivine CRIBLE

Thèse soutenue en vue de l'obtention du grade
de Docteur en Langues et Lettres

Membres du jury

Prof. Michel Francard (Université catholique de Louvain), président
Prof. Liesbeth Degand (Université catholique de Louvain), promotrice
Prof. Gaëtanelle Gilquin (Université catholique de Louvain), promotrice
Prof. Maria Josep Cuenca (Universitat de València)
Prof. Kerstin Fischer (Syddansk Universitet)
Prof. Anne-Catherine Simon (Université catholique de Louvain)
Prof. Sandrine Zufferey (Université de Berne)

Février 2017

Table of contents

List of Figures	vii
List of Tables.....	ix
List of abbreviations and acronyms.....	xi
Acknowledgments	xiii
Chapter 1: Introduction.....	1
1.1 Fluency in time and space	1
1.2 Background and scope of the study.....	3
1.3 Preview of the thesis.....	6
Chapter 2: Fluency, disfluency and the non-linear processes of speech production.....	9
Introduction to the chapter	9
2.1 The temporal dynamics of spoken language	9
2.1.1 Speaking and writing: a binary divide?	9
2.1.2 Information packaging: the cyclic flow of speech.....	12
2.1.3 Looking back, moving forward: linearity in question	14
2.2 Fluency, disfluency and disfluencies: defining a multi-faceted construct.....	16
2.2.1 Disfluency or repair? Levelt's legacy.....	16
2.2.2 Holistic definitions of fluency	19
2.2.3 Componential approaches to disfluencies	21
2.2.4 Bridging the gap: fluency as non-linearity	31
2.3 A usage-based account of (dis)fluency.....	32
2.3.1 Key notions in usage-based linguistics.....	33
2.3.2 From schemas to sequences of fluencemes	34
2.3.3 Variation in context(s).....	36
2.3.4 Accessing fluency through frequency	38
2.4 Hypotheses: combination and variation of fluencemes	40
Chapter 3: Discourse markers in spoken language.....	43
Introduction to the chapter	43
3.1 What are discourse markers?.....	44
3.1.1 Core features.....	45
3.1.2 Beyond the terminological debate	49
3.1.3 From monolingual case studies to crosslinguistic categories	52
3.1.4 Towards a corpus-based definition.....	57
3.2 The functions of discourse markers in corpora	58
3.2.1 Writing-based models of discourse relations.....	59
3.2.2 The many scopes of DM functions.....	64

3.3 Are discourse markers fluencemes?	69
3.3.1 “Fluent” vs. “disfluent” discourse markers	70
3.3.2 The non-linearity of discourse markers	74
3.4 Hypotheses: the place of DMs in the typology of fluencemes	77
Chapter 4: Corpus and method	83
Introduction to the chapter	83
4.1 Data	83
4.1.1 From research questions to corpus selection	83
4.1.2 Source corpora.....	85
4.1.3 Comparable corpus design	87
4.1.4 Technical treatment	91
4.2 Annotation protocol at DM level.....	92
4.2.1 Corpus-based coding scheme	93
4.2.2 Annotation procedure	103
4.3 Annotation protocol at fluenceme-level	106
4.3.1 Internal structure of fluencemes	106
4.3.2 Simple fluencemes.....	107
4.3.3 Compound fluencemes	108
4.3.4 Related phenomena and diacritics	110
4.3.5 Replicability of the fluenceme-level annotation protocol	112
4.3.6 Scope of the annotation in <i>DisFrEn</i>	114
4.4 Data extraction and post-treatment.....	116
4.4.1 EXAKT concordancer by analytical level.....	116
4.4.2 Macro-labels.....	118
4.4.3 Additional variables in Excel	120
4.4.4 Identification codes	121
4.5 Conclusion of the chapter.....	123
Chapter 5: Portraying the category of discourse markers	125
Introduction to the chapter	125
5.1 General frequency across languages and registers	125
5.2 Positional variables.....	132
5.2.1 DMs across units	132
5.2.2 Formal and cognitive restrictions of less typical positions.....	136
5.3 Functional variables	142
5.3.1 (Non-)relational type	142
5.3.2 Single domains and functions.....	145
5.3.3 Double domains and functions	159

5.3.4 Integrating syntax and pragmatics	162
5.4 Co-occurrence of DMs	167
5.4.1 Modeling the factors of co-occurrence	168
5.4.2 The company they keep: from co-occurring to complex DMs	173
5.5 Summary and interim discussion: the potential of bottom-up research	180
Chapter 6: The (dis)fluency of discourse markers: insights from the clustering of fluencemes	183
Introduction to the chapter	183
6.1 Paradigmatic annotation of fluencemes in interviews	183
6.1.1 Fluenceme rates	184
6.1.2 Sequence length and most frequent clusters	188
6.1.3 Fluency-as-frequency across different degrees of granularity	191
6.2 DM-based sequences across registers	198
6.3 From isolation to combination: focus on DMs and pauses in interviews	206
6.3.1 Independent uses of DMs and FPs	207
6.3.2 DM, FP, UP: the impact of clustering	211
6.3.3 Modeling the clustering of DMs and pauses	215
6.4 The scope of semantic-pragmatic pairs	217
6.5 Potentially Disfluent Functions	222
6.6 Towards a cognitive-functional scale of (dis)fluency?	228
6.7 Summary and interim discussion: the “silence” of corpora	234
Chapter 7: From qualitative repair categories to a formal scale of fluency	239
Introduction to the chapter	239
7.1 Previous approaches to repair	240
7.1.1 Reformulation and its markers: the French classics	240
7.1.2 Contrastive perspectives on reformulation markers	242
7.1.3 From reformulation to repair: Levelt’s (1983) typology of repair	245
7.1.4 Research questions and hypotheses	247
7.2 Identification and coding scheme	249
7.2.1 Selection criteria	249
7.2.2 Repair category	250
7.2.3 Formal variables	254
7.2.4 Relation to annotated fluencemes	258
7.2.5 Other information	259
7.2.6 Procedure and post-treatment	259
7.2.7 Coding consistency and intra-annotator agreement	260
7.3 Results	262
7.3.1 Distribution of repair categories	262

7.3.2 Repair category and formal correlates.....	264
7.3.3 RMs in repair.....	275
7.3.4 DMs in repair.....	279
7.3.5 Other elements in the repair and editing phase.....	288
7.4 Summary and interim discussion: low quantity, high quality?	290
Chapter 8: Conclusion	297
8.1 Summary of the main findings	297
8.2 General discussion.....	300
8.3 Implications and research avenues	302
Bibliography.....	305
Appendices	331
Appendix 1: DM-level annotation protocol (Crible 2014).....	333
1 Introduction	335
2 Overview of the protocol: design and applications	336
3 Annotation tiers and values	343
4 Mapping of pragmatic and non-pragmatic functions	357
5 Guide to frequent polysemous discourse markers	361
References in Appendix 1	365
Appendix 2: Fluenceme-level annotation protocol (Crible et al. 2016)	371
1 Introduction	373
2 Overview of the protocol: design and applications	374
3 Categories of fluencemes covered by the annotation protocol	376
4 Queries.....	392
5 Conclusion.....	392
References in Appendix 2	394
Appendix 3: Top-five most frequent discourse markers by register	397
Appendix 4: List of discourse markers in <i>DisFrEn</i> and their functions.....	398
Appendix 5: Macro-syntactic position of DMs by register	409
Appendix 5.1: Distribution of DMs across macro-syntactic slots by register.....	409
Appendix 5.2: Proportions of macro-syntactic slots by register.....	409
Appendix 6: Mapping of domains and (non-)relational type by register	410
Appendix 7: List of functions in <i>DisFrEn</i> and their discourse markers.....	411
Appendix 8: Top-five most frequent functions by register in <i>DisFrEn</i>	416
Appendix 9: Proportions of sequence types by position and domain.....	417

List of Figures

Figure 2.1: Levelt's (1983) terminology.....	p.17
Figure 2.2: Shriberg's (1994: 57) annotation model.....	p.24
Figure 3.1: Present terminological model of discourse markers and other pragmatic expressions.....	p.51
Figure 3.2: Illustration of the functional flexibility of DMs.....	p.68
Figure 4.1: English and French source corpora in <i>DisFrEn</i> (in minutes)	p.86
Figure 4.2: Distribution of registers (in minutes) per language in <i>DisFrEn</i>	p.87
Figure 4.3: Levels of elicitation in <i>DisFrEn</i> (in minutes).....	p.88
Figure 4.4: Number of speakers in <i>DisFrEn</i> (in minutes).....	p.89
Figure 4.5: Degrees of preparation in <i>DisFrEn</i> (in minutes).....	p.89
Figure 4.6: Degrees of interactivity in <i>DisFrEn</i> (in minutes).....	p.90
Figure 4.7: Media coverage (broadcasting) of the registers in <i>DisFrEn</i> (in minutes).....	p.90
Figure 4.8: (Non-)professional interactions in <i>DisFrEn</i> (in minutes).....	p.91
Figure 4.9: Macro-syntactic segmentation for DM position.....	p.99
Figure 4.10: Partitur Editor annotation interface.....	p.103
Figure 4.11: Annotation interface in Partitur Editor with both protocols in <i>DisFrEn</i>	p.116
Figure 4.12: EXAKT annotation extraction interface at DM level.....	p.117
Figure 4.13: EXAKT annotation extraction interface at sequence level.....	p.117
Figure 5.1: Proportions of part-of-speech tags in news broadcasts.....	p.131
Figure 5.2: Proportions of part-of-speech tags in conversations.....	p.131
Figure 5.3: Proportions of DMs in micro-position (clause-level) by register.....	p.133
Figure 5.4: Macro-position (dependency-level) of DMs.....	p.134
Figure 5.5: Proportions of turn-initial DMs by degree of interactivity.....	p.135
Figure 5.6: Proportions of POS-tags across macro-syntactic positions.....	p.137
Figure 5.7: Proportions of scores of coding complexity by micro-syntactic positions.....	p.140
Figure 5.8: (Non-)relational type by position in the turn.....	p.144
Figure 5.9: Distribution of DM domains across registers.....	p.147
Figure 5.10: Proportions of interpersonal and sequential DMs in each register.....	p.149
Figure 5.11: Balance of domains in the three degrees of preparation.....	p.150
Figure 5.12: Number of function types making up 50% of DMs by register and language.....	p.155
Figure 5.13: Proportions of macro-syntactic slots in each domain.....	p.162
Figure 5.14: Extended association plot of domains and macro-position.....	p.164
Figure 5.15: Pruned classification tree of domains.....	p.166

Figure 6.1: Proportions of sequence type (coarse-grained) by sequence length.....	p.195
Figure 6.2: Proportions of sequence type (fine-grained) by sequence length.....	p.195
Figure 6.3: Conditional inference tree for isolated, clustered and co-occurring DMs.....	p.200
Figure 6.4: Conditional inference tree for sequence category by register.....	p.202
Figure 6.5: Extended association plot between sequence categories and registers.....	p.203
Figure 6.6: Patterns of position and clustering for <i>euh</i> , <i>uh</i> and <i>um</i>	p.210
Figure 6.7: Patterns of position for FPs with and without a DM.....	p.214
Figure 6.8: Multiple correspondence analysis of the clustering of FPs with DMs.....	p.215
Figure 6.9: Pruned classification tree of [DM+pause] configurations.....	p.216
Figure 6.10: Extended association plot of PDFs and non-PDFs across registers.....	p.224
Figure 6.11: Extended association plot of PDFs and non-PDFs across sequence types.....	p.225
Figure 6.12: Length of sequences in fluenceme tokens in PDFs and non-PDFs.....	p.227
Figure 6.13: Extended association plot of functional domains by sequence type.....	p.228
Figure 6.14: DM domains on the scale of (dis)fluency.....	p.230
Figure 6.15: Multiple correspondence analysis of domains, position and sequence type.....	p.232
Figure 7.1: Proportions of interrupted units by repair category.....	p.265
Figure 7.2: Distance (in words) between occasion and interruption by repair category.....	p.267
Figure 7.3: Ways of restarting across repair categories.....	p.269
Figure 7.4: Span of retracing (in words) across E, A and R-repairs.....	p.271
Figure 7.5: Proportions of distance (in words) by moment of interruption.....	p.274

List of Tables

Table 2.1: The oral-written dichotomy, from Horowitz & Samuels (1987: 9).....	p.10
Table 2.2: Conceptual representation of temporality and linearity across speech and writing.....	p.15
Table 2.3: Annotation of self-interruptions in Pallaud et al. (2013a, 2013b).....	p.27
Table 2.4: Götz's (2013) three-fold typology of fluencemes.....	p.29
Table 2.5: Typology of fluencemes investigated in this study.....	p.32
Table 3.1: Characteristic features of DMs in four definitions.....	p.48
Table 3.2: Discourse relations in the PDTB 2.0 with their typical connective (Prasad et al. 2008).....	p.60
Table 3.3: Revised PDTB from Zufferey & Degand (in press).....	p.61
Table 3.4: Summary of analyses.....	p.80
Table 4.1: List of all part-of-speech tags for DMs (with examples).....	p.101
Table 4.2: Overview of the annotation tiers specified by the DM-level protocol.....	p.102
Table 4.3: Confusion matrix for the inter-annotator agreement on the reduced dataset.....	p.113
Table 4.4: Macro-labels for the internal structure of the sequence.....	p.120
Table 5.1: Raw and relative frequency of DMs by language and register.....	p.126
Table 5.2: Distribution of DMs across degrees of preparation and interactivity.....	p.127
Table 5.3: Top five most frequent DMs in English and French.....	p.129
Table 5.4: Type-token ratio of DMs.....	p.130
Table 5.5: Position in the clause (micro-position) by language.....	p.132
Table 5.6: Relative frequency (ptw) of (non-)relational DMs by language and register.....	p.142
Table 5.7: Taxonomy of DM domains and functions.....	p.145
Table 5.8: Distribution of single domains by language.....	p.146
Table 5.9: Relative frequency of domains (ptw) by number of speakers.....	p.148
Table 5.10: Cross-tabulation of domains and part-of-speech tags in English and French.....	p.150
Table 5.11: (Non-)relational type by domain.....	p.152
Table 5.12: Ten most frequent functions and their relative frequency by language.....	p.153
Table 5.13: Standardized DM function ratio by language and register.....	p.156
Table 5.14: Distribution of double domains per language.....	p.160
Table 5.15: Distribution of double tags and overall proportion by register.....	p.161
Table 5.16: Combinations of DMs by decreasing frequency (excluding <i>hapax legomena</i>).....	p.169
Table 5.17: Number and proportion of co-occurring DMs across micro-syntactic positions.....	p.170
Table 6.1: Relative frequency (per thousand words) of fluenceme tokens in each subcorpus.....	p.185
Table 6.2: Sequence length (in number of fluenceme tokens) by register and language.....	p.189

Table 6.3: Relative frequency of sequences (N > 100) ptw by language and register.....	p.190
Table 6.4: Relative frequency (ptw) of sequence structures in each subcorpus.....	p.192
Table 6.5: Relative frequency of DM-based sequences ptw in <i>DisFrEn</i>	p.198
Table 6.6: Proportions of isolated DMs across positions.....	p.208
Table 6.7: Proportions of DMs with unfilled pauses across positions.....	p.209
Table 6.8 Filled pauses (with and without UP) by position and language.....	p.210
Table 6.9: Distribution of [DM+pause] clusters in face-to-face interviews.....	p.212
Table 6.10: Proportions of [DM+FP] clusters across positions.....	p.213
Table 6.11: Proportions of macro-syntactic positions for the semantic-pragmatic relations.....	p.219
Table 6.12: Proportions of sequence types for the semantic-pragmatic relations.....	p.219
Table 6.13: Relative frequency (ptw) of PDFs per language and register.....	p.223
Table 6.14: Proportions of micro-syntactic positions by sequence type.....	p.231
Table 6.15: Significant effects for the multiple logistic regressions by domain.....	p.233
Table 7.1: Revised typology of repair from Levelt (1983).....	p.251
Table 7.2: Definition and examples of boundaries at the interruption point.....	p.255
Table 7.3: Variations of repair according to the way of restarting.....	p.257
Table 7.4: Intra-coder consistency for each repair type.....	p.260
Table 7.5: Final values for repair type and number of (dis)agreements.....	p.262
Table 7.6: Frequency and proportions of repair categories and subtypes by language.....	p.263
Table 7.7: Preferred ways of restarting and their proportion by subtype of repair.....	p.270
Table 7.8: Cross-tabulation of moment of interruption and way of restarting.....	p.272
Table 7.9: Proportions of modified repetitions across repair types.....	p.276
Table 7.10: Distribution of RMs in interrupted units by repair type.....	p.277
Table 7.11: Proportions (and frequency) of RM and FS fluencemes included across repair types.....	p.278
Table 7.12: Distribution of DMs across repair types and positions in the repair (if any).....	p.280
Table 7.13: Functions and frequent lexemes of DMs in the editing phase.....	p.283
Table 7.14: Presence and position of DMs and RMs.....	p.287

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
GloLin	global linearization repair
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
LocLin	local linearity repair
MCA	multiple correspondence analysis
MDMA	Model for Discourse Marker Annotation
MID	middle-field macro-syntactic position
news	subcorpus of news broadcasts
NN	noun phrase
NRDM	non-relational discourse marker
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
RDM	relational discourse marker
RHE	rhetorical domain
RIGHT	right-integrated macro-syntactic position
R-sequence	sequence containing repetitions
RI	identical repetition
RM	modified repetition
RST	Rhetorical Structure Theory
S1	first segment in a discourse relation
S2	second segment containing the DM in a discourse relation
SC	subordinating conjunction (also subord. conj)
SDRT	Segmented Discourse Representation Theory
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

A PhD thesis is always somewhat of a solitary adventure, or rather a love-hate duo between the candidate and their computer(s). In my case, however, this is not entirely true for a number of reasons, and these reasons are wonderful people whom I wish to warmly thank here. 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. We would frequently exchange short updates, jokes and sometimes “slightly” off-topic conversations which kept us in touch with each other, and I will cherish the sound of their fast heels walking up the fourth floor. I did not expect to share so much of myself with my professors, and I can proudly say that I truly felt considered more like a colleague than a student. Liesbeth and Gaëtanelle, thank you for your restless guidance and your friendship!

I was also lucky to be a member of not one but 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 (and the few hybrid ones like me). A special thought goes to the members of the ARC-Fluency team, both in Louvain-la-Neuve and Namur, who taught me the merits (and challenges!) of teamwork and whose contribution to the present work is substantial. 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 DM community here and abroad (Sílvia, Elena and more). One last – although chronologically first – 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: here’s to our continued collaboration!

But 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 and its recent Franco-Swiss addition, 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 supported me, even though they did not always understand me (yes, now I am finally off school). Last but not least, to my loving teammate, who suffered endless blabber about discourse markers (ask him anything!), post-conference debriefs and office gossip with a true stoic class, cared for me, fed me, always believed in me, my most faithful supporter, Kévin. Collector!

Chapter 1: Introduction

1.1 Fluency in time and space

“Spoken language exists in time, not space”

Carter & McCarthy (2006: 193)

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 this proposal shows compelling descriptive value, 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: the spacetime continuum. Beyond the taste for science fiction that it gives away, 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) brings forward the specificity of speech as opposed to writing and, to a lesser extent, sign language: speakers and listeners cannot “rewind” nor move forward but are “stuck” in the linear flow of speech. In this, speech contrasts (i) with written texts, which are not constrained by the same time pressure and where writers and readers are free to navigate along the different graphical parts (Danks & End 1987) and (ii) 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 graphical 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 timely nature of online production. In other words, fluent speech should be linear. The notion of linearity is affiliated to Levelt’s (1989) “linearization” and his “blueprint” model of speech production, whereby speakers have to handle simultaneously macroplanning (i.e. designing the communicative intention), microplanning (i.e. structuring the utterance) and monitoring (i.e. comparing the utterance with the intention and instructing adjustments if necessary) within their

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*.

limited cognitive abilities, especially those of working memory. While linearization is a fact of thought and not of speech (Levelt 1981), linearity, on the other hand, is modality-dependent. It emerges from the inadequacy of the linear phonological channel to render the output of our complex mental processes effectively. Equating fluency with linearity is therefore not true to the cognitive processes of language production and perception, and particularly unrealistic for spontaneous, unplanned speech.

In this thesis, I will strive to show that linguistic expressions and structures triggering and/or reflecting non-linear processing are not systematically problematic (as opposed to what a writing-based standard of fluency would argue) but can actually offer solutions to circumvent the linearization problem and create coherent and efficient discourse. Non-standard structures such as so-called 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 should be viewed as “tricks” that allow speakers to reconstitute a spatial dimension to the temporality of speech by manifesting the directionality and non-linearity of particular discourse moves. In this sense, the spacetime metaphor is related to that of language as motion. By pursuing such a growing line of research, this thesis thus answers Auer’s (2009) call for more research taking the notions of linearity and temporality as central in the study of speech.

This approach puts to the forefront of (dis)fluency markers those expressions 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 will therefore be the central focus of this research, connecting their many forms and functions to a non-linear view of (dis)fluency and studying their combination with other (dis)fluent devices such as pauses or repetitions in different 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 knows* and many *ums* in a 30-minute interview.² By contrast, an American teacher published on her blog an article on “How and why to teach discourse markers” to students: “These markers are important in connecting parts of the discourse as well as contributing to fluency. In addition, they guide the listener or reader in the direction of the discourse”.³ Many similar online articles and videos point to a duality between disruptive (even

² <http://languageblog.ldc.upenn.edu/nll/?p=964>. Last accessed on Dec. 9th, 2016.

³ <http://busyteacher.org/10076-how-and-why-to-teach-discourse-markers.html>. Last accessed on Dec. 9th, 2016.

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 exclude elements which do not appear to signal any inference to be made but merely translate some kind of trouble on the speaker's part. These "symptoms", as opposed to "signals" (Clark & Fox Tree 2002), are two sides of the same coin, and it will be argued throughout this thesis that it is only through a cluster of contextual and linguistic variables that, for a single element, the "symptom" vs. "signal" 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 non-linearity, 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 could 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) and methods (quantitative-qualitative analysis). Another source of information to feed this scale of fluency is to use frequency as a cue 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 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 specificities and commonalities 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 scope of the study

In the previous section, I have introduced the focus of the present study on elements triggering non-linear processes of production and perception, especially those related to discourse structure such as discourse markers, in a quantitative-qualitative usage-based approach. Despite the relative novelty of this approach to fluency as (non-)linearity, it is far from being unique or

standalone and owes many of its conceptual foundations to previous research, especially in its concrete application to corpus data. While this legacy will be presented and discussed at length in the following chapters, some restrictions should already be mentioned in order to situate the scope of the present study in a broader scientific background.

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 theoretical and methodological frameworks. In this section, I provide a very brief literature review over the theoretical and methodological approaches which are quite different from mine and whose influence in the remainder of the thesis is therefore limited. This section thus illustrates the breadth of fluency research with representative examples of some major frameworks, as well as their differences, limitations (if any) and relevant lessons in relation to the present approach.

A first exclusion from the scope of the present study is pathological disfluency (often spelt “dysfluency” in this case), which is concerned with medical conditions such as stuttering, aphasia or dyslexia. These studies (e.g. Wingate 1987; Mahesha & Vinod 2012) seek to distinguish “normal” from pathological disfluencies, such as the difference between repetition and stuttering. These approaches in speech therapy and psychotherapy will not be discussed any further in this thesis, which is concerned with “normal” speech disfluencies.

Another restriction regarding the speakers under investigation is that of second language (henceforth L2) learners. Studies on L2 fluency have occupied center-stage of the field ever since its beginning and are still gaining ground, usually by comparing native and non-native speakers in order to assess the fluency or proficiency of the latter. In other words, these works often take native speech as a reference or target, which in itself calls for further research on L1 fluency in its full array in order to base the comparison on solid grounds. Major references in L2 fluency are, among others: Chambers (1997); Freed (2000); Lennon (2000); O’Connell & Kowal (2004); Segalowitz (2010); Gilquin & De Cock (2011); Osborne (2011); Götz (2013). Some of these references proved relevant beyond the realm of L2 fluency and will therefore be commented on in Chapter 2.

The next group of references takes us once more to the origins of the field, which started with the study of temporal variables of fluency such as speech rate, articulation rate and pause duration. This line of research, also called *pausology*, was initiated by the founders Maclay & Osgood (1959) and Goldman-Eisler (1968), as well as the contrastive milestone Grosjean & Deschamps (1975) for English and French. Many more authors have taken up the study of pauses and their multiple functions (e.g. Lundholm 2015 recently), uncovering the complexity of the phenomenon which also relates to other aspects of prosody such as stress and intonation patterns. Prosody and pauses in particular constitute a whole field of study of their own since they involve more (quantitative) parameters than most other works in fluency, as well as separate tools (e.g. automatic annotation) and methods. Candéa (2000) and Moniz et al. (2009), for instance, illustrate the typical use of experimental paradigms in order to relate production and perception, sometimes in combination with corpus data.

With respect to the present study, pauses will be included in the typology and analysis of fluency phenomena – as opposed to some of the major frameworks where they are completely excluded (e.g. Shriberg 1994) – but only in a basic way. No distinction of pause types based on their duration will be established given that they are essentially related to rate of articulation and thus require a relative baseline specific to each speaker (Little et al. 2013). No other prosodic features will be investigated. It could be argued that studying spoken language without focusing on what makes it spoken, i.e. prosody, is a strong limitation to the validity and outreach of this research. While I acknowledge this caveat, I believe that the integrity of the study is better preserved by excluding rather than treating lightly a complex object of research which requires deep, focused analyses and would not directly answer my research questions.

Related to prosodic studies are experimental studies which are flourishing in the recent literature and generally tend to show positive effects of disfluencies. For instance, Arnold et al. (2003: 32) study the role of English *uh* in reference resolution by means of an eye-tracking experiment, with which they show that “disfluency increases the accessibility of discourse-new objects during reference resolution” (cf. also Barr & Seyfeddinipur 2010; Bosker et al. 2014 for similar experiments). Corley and colleagues (e.g. Corley et al. 2007; MacGregor et al. 2009) use event-related potential (ERP) experiments to measure the cognitive response to disfluencies and their effect on long-term memory. They found that fillers differ from repetitions by enhancing the recognition of words following *er* in surprise memory tests, while repetitions do not seem to ease the processing of subsequent words. Liu & Fox Tree (2012) also work on memorability and the focusing-attention function of hedges such as *maybe*, *I think* or *I dunno*, comparing their effect to that of *like*. They conclude that hedges mark information as less worthy yet more accurately remembered, whereas *like* does not show similar effects probably because of the degree of familiarity between speakers that it seems to require. These are only a few examples of the vast body of experimental research which cannot be fully reviewed here. Although insightful, these studies are usually restricted to one type of disfluency (very often filled pauses like *uh*) in highly controlled (even unnatural) settings, as opposed to the present bottom-up approach to many types of disfluencies and their combination in authentic communication, which is why experimental works will only be punctually referred to when they are relevant to the present corpus findings.

Another methodological distinction with a stronger applied perspective is that of computational linguistics or natural language processing which, when concerned with fluency research, tends to pursue a single endeavor, namely to automatically detect and erase disfluencies from transcriptions (e.g. Zechner 2001; Strassel 2003; Mieskes & Strube 2008). This line of research relies on the assumption that disfluencies need to be “cleaned” or removed, a position which stands in sharp contrast with the present approach. Another computational perspective is that of speech synthesis and human-machine communication. Robots and other speaking devices benefit from the input of disfluencies to interpret human reactions and improve the naturalness and smoothness of the exchange (e.g. Fischer 1999) or to be themselves perceived as more polite or likeable (e.g. Torrey et al. 2013). Overall, computational studies on fluency pursue a different agenda (modeling for removal or modeling for synthesis vs. modeling for theory-building) usually with a more coarse-grained level of precision than what is presently aimed at.

More frameworks and areas of fluency research could probably be identified. Suffice it to say that methods (corpus, experimentation, automatic processing), research topics and agendas keep on developing towards new avenues. Nevertheless, against this background of existing research, a number of gaps in the field need to be filled and motivate the present study. The major gap is probably the quasi-absence of crosslinguistic 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-Vandenberghe 2006), however not in direct relation to fluency and disfluency – many of them actually work on discourse markers in writing.

This thesis aims at addressing both these gaps in the field, namely studying contrastive (L1) fluency in English and French and relating discourse markers to their role in (dis)fluency. By considering discourse markers as full-fledged markers of (dis)fluency, I wish to reconcile mainstream DM studies, which do not investigate their contribution to fluency, with the research community on fluency, which acknowledges the role of discourse markers in speaker's fluency (especially for their role on 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). In this respect, the present study stands as rather innovative against both DM and fluency research, in addition to its large scope over entire categories as opposed to the numerous case-study approaches in each field.

1.3 Preview of the thesis

The twofold goal over DMs and (dis)fluency introduced above is reflected in the structure and method of analysis of the thesis.⁴ Each domain will first be defined and analyzed as two distinct levels (henceforth referred to as DM level and sequence level) in separate chapters before being combined and integrated in synthesizing analyses, as the following overview will now make apparent.

This thesis includes six chapters besides introduction and conclusion: two theoretical, one methodological and three empirical. The next chapter (Chapter 2) develops the present non-linear 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 fluency. The assumptions of usage-based linguistics and their application to the present object of study will

⁴ Each section and subsection includes a reference to the ongoing chapter: for instance, Section 2.3.1 is the first subsection of the third section in Chapter 2. However, the numbering of examples is reset to (1) at the beginning of each chapter, unlike footnotes which are numbered continuously throughout the thesis.

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 specific challenges of a contrastive bottom-up approach to the highly multifunctional 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 (non-)linearity and in relation to the (relatively small) literature combining these two objects of study.

Chapter 4 presents the dataset (corpus design, sampling and technical treatment) and methodology, detailing the annotation schemes for DMs and (dis)fluency markers. Reports on intra- and inter-annotator agreement will also be discussed in order to assess the reliability of the annotation.

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 or connectivity. In particular, the integration of positional and functional variables will uncover interesting form-meaning patterns. Special attention will be paid to the phenomenon of DM co-occurrence in a qualitative-quantitative methodology testing different degrees of fixation on corpus annotations. 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 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? This chapter will first situate DMs within this typology by identifying the rate and strength of association between different markers, focusing in particular on the combination of DMs with pauses. Functional variables will then be integrated in the analysis of clusters in order to identify more or less fluent DM functions and sketch a tentative scale of (dis)fluency on the basis of form-meaning patterns and their distribution across registers.

Chapter 7 will present the findings of a conversation-analytic study of repair combining the annotations of (dis)fluency and discourse markers with a qualitative categorization of repair types and formats strongly inspired by Levelt's (1983) model. This chapter will disentangle the complex interplay between (dis)fluency markers, functions of discourse markers and the formal structure of repair sequences, identifying their respective contribution to fluent vs. disfluent stretches of talk. This last analysis is designed to provide a more direct access to the interpretation of fluency and disfluency, pursuing the same overarching goal to build a scale of fluency based on form-function patterns.

The main findings of the thesis will be summarized and discussed in Chapter 8, along with suggestions for further research avenues and implications of the present study. Each chapter begins with an introduction and Chapters 4 to 7 include their own summary and interim discussion. Chapter 8 ties together the different parts and results of this thesis and suggests an integrated view of the (dis)fluency of discourse markers across registers in English and French.

Chapter 2: Fluency, disfluency and the non-linear processes of speech production

Introduction to the chapter

The aim of this chapter is to define and discuss the theoretical notions, models and frameworks related to the concepts of fluency and disfluency, starting with some core characteristics of spoken language, namely its temporal, cyclic and (non-)linear nature. Different approaches to both definition and annotation of (dis)fluency will be systematically compared, before introducing the approach presently adopted. 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. We then turn to the hypotheses and research questions related to (dis)fluency across registers in English and French. At the end of this chapter, it should appear clearly to the reader what the introductory metaphors of spacetime and linearity entail and what aspects of (dis)fluency are covered in this thesis. No detailed mention of discourse markers will be made in this chapter which targets the general mechanisms and components of (dis)fluency, rather than the specific category of DMs which it includes (see Chapter 3).

2.1 The temporal dynamics of spoken language

While the study of (dis)fluency emerged with the interest for natural spoken language, many definitions are actually based on an idealized written mode of communication, so much so that it is important to situate the specificities of speech as opposed to writing in terms of language production and processing. While many of the studies discussed in this section make direct references to fluency, the goal of this section is to set the scene as to the particular conditions of spoken communication and their impact on (non-)linearity and (dis)fluency as developed later on.

2.1.1 Speaking and writing: a binary divide?

Speaking and writing are two natural modes of communication which have both been the object of extensive linguistic research, although not to the same extent. A considerable – and still growing – part of the literature strives to compare these modalities: beyond their obvious technical differences, authors have identified a number of distinctive features at the levels of production and comprehension of spoken vs. written language. The main comparative findings are laid out in this section, starting with Horowitz & Samuels (1987: 7) who define oral language as “primarily phatic communication” (that is, listener-oriented and anchored in context) and as being “swift, brief, linear, with constructions that are based on a single predication and simplicity”. While they acknowledge the variation (e.g. in registers) within each modality, their account is rather coarse-grained and tends to rank writing as a superior form of language. They identified comparative features for each side of the dichotomy, which are

summarized in Table 2.1. One criticism to this table is that it groups fundamental properties of each medium (such as the “fleeting” vs “permanent” distinction) together with features which are only typical but not necessary and depend on the register rather than the modality (such as “narrativelike” vs. “expositorylike” or “spontaneous” vs. “planned”).

Table 2.1: The oral-written dichotomy, from Horowitz & Samuels (1987: 9)

Oral language – Talk	Written language – Text
Reciprocity between speaker and listener	Limited reciprocity between author and reader
Narrativelike	Expositorylike
Here and now	Future and past
Informal	Formal
Interpersonal	Objective and distanced
Spontaneous	Planned
Sharing of situational context	No common context
Structureless	Highly structured
Repetition	Succinctness
Simple linear structures	Complex hierarchical structures
Fleeting	Permanent
Unconscious	Conscious and restructures consciousness

In the same volume, Halliday (1987: 65) takes a rather opposite – all the while more nuanced – stand by showing that “speech is not, in any general sense, ‘simpler’ than writing; if anything, it is more complex”. Instead of a dichotomy, Halliday talks of a continuum with differences in “texture”, namely lexical density (high in writing, low in speech) and grammatical intricacy (low in writing, high in speech), concluding that both modalities are complex in their own way. He further makes controversial yet relevant observations regarding the fluent flow of unprepared speech:

Speech, we are told, is marked by hesitations, false starts, anacolutha, slips and trips of the tongue, and a formidable paraphernalia of so-called performance errors [...] They are characteristic of the rather self-conscious, closely self-monitored speech that goes, for example, with academic seminars [...]. If you are consciously planning your speech as it goes along and listening to check the outcome, then you naturally tend to lose your way: to hesitate, back up, cross out, and stumble over the words. But these things are not a particular feature of natural spontaneous discourse, which tends to be fluent, highly organized and grammatically well formed. If you are interacting spontaneously and without self-consciousness, then the clause complexes tend to flow smoothly without you falling down or changing direction in the middle, and neither speaker nor listener is at all aware of what is happening. (Halliday 1987: 68)

While this position might be slightly extreme and partly invalidated by more recent corpus-based studies, showing indeed a higher frequency of hesitation phenomena in spontaneous speech (e.g. Beliaio & Lacheret 2013, see Section 3.3.1), it still points out the special relation

between planning, monitoring and fluency and their relative modality-independence: hesitation and revision can also apply to the production of written texts (when erasing and re-organizing different parts of a thesis, for instance), the difference with speech being that in the former, the reader has only access to the final product, whereas in the latter, the listener witnesses the on-going process.

Chafe (1994: 43) also expands on this property of written texts to be “worked over”, as opposed to the spontaneity of conversations. According to him, a corollary to this property is the difference in tempo between written (slow rate) and spoken production (faster pace). He further identifies additional distinctive features such as the richness of prosody, naturalness, copresence and situatedness; however, the main difference seems to lie in the evanescence of speech as opposed to the permanence of texts: *verba volant, scripta manent* or, as he puts it, “[w]ritten language, including transcriptions of spoken language, is not only preservable through time and space, but can be dissected, analyzed and otherwise manipulated” either by readers or linguists (1994: 42).

It is this very affordance of writing to be re-read that Danks & End (1987) draw upon when they study the cognitive demands of the two modalities, in their endeavor to compare mechanisms of processing (and not of production). Like Chafe (1994), Danks & End (1987) refer to time and space to oppose listening and reading:

The auditory system is temporally based. The listener has limited control over speech rate and has no continuing access to it. [...] In contrast, the visual system is spatially oriented. The reader has continuous access to the complete text, at least in naturalistic situations, and also has more-or-less complete control over the rate and order of input. (Danks & End 1987: 273)

This space-time contrast, which is also at the core of the introduction to this thesis (Chapter 1), is a powerful representation of the “limited control” over spoken material, as opposed to the freedom of movement within a written text, both for production and comprehension. Another related consequence is the higher demands on working memory in speech since it lacks “a permanent record to be reexamined by the listener in case of processing difficulty” (Danks & End 1987: 285). Listeners can be helped in this matter by efficient recipient design, that is the speakers’ listener-oriented task of facilitating information retrieval through various strategies (one of them being (dis)fluent elements). While Danks & End (1987) acknowledge that reading and listening share a number of processing mechanisms (such as a common knowledge base), they converge with other works discussed in this section in identifying fundamental differences between speech and writing which potentially impact the fluency of discourse.

The link between fluency and the temporal nature of speech is made explicit by Clark (2002) who combines the notion of planning with that of “synchronization”, that is, the speakers’ attention to the listener’s needs: “they must attend closely to the timing of their own and their partner’s speech and deal with timing when it goes awry. That makes timing central to spontaneous speech, and it leads to the special design of disfluencies” (2002: 5).⁵ Without

⁵ See also Auer (2009) on the temporality of spoken language which he decomposes as transitoriness (or rapid fading), irreversibility and synchronization.

going into the detail of this relation (see Section 2.2), we can already say that Clark's analysis puts forward the spontaneity or lack of planning of speech.

This focus on planning suggests to distinguish spoken language in general, which is always by nature temporal, from unplanned speech, which is only one use of this modality characterized by its spontaneity. Going one step further, Halliday (1987: 66) considers "written" and "spoken" to be indeterminate categories which "may refer to the medium in which a text was originally produced, or the medium for which it was intended, or in which it is performed in a particular instance; or not to the medium at all, but to other properties of a text which are seen as characteristic of the medium", thus going beyond the binary speech-writing divide. This approach is very similar to Koch & Osterreicher's (2001), who distinguish the medial dimension (spoken vs. written) from the conceptual dimension, which is more gradual and covers many parameters such as privacy, intimacy, emotionality, copresence, spontaneity, etc. Such a finer view of communication settings will prove highly relevant to the study of disfluency (see Section 2.4, see also the situational metadata in Chapter 4).

2.1.2 Information packaging: the cyclic flow of speech

The temporality of speech, that is the high situatedness and dependence on the ever-changing context of production and perception, is reflected in many superficial and structural ways, one of them being the regular rhythm of segmentation of the speech signal. Many authors from corpus-based and experimental frameworks have gathered evidence of the cyclic flow of speech, as opposed to the more continuous and linear lay-out of sentences in written texts. An impressive number of studies have in fact shown that speakers' planning skills are constrained by limitations of working memory to produce one conceptually coherent unit of talk at a time, that is, one proposition or idea or clause between major pause boundaries. I will focus here on three notions: Chafe's (1994) "focus of consciousness", Pawley & Syder's (1975/2000) "one-clause-at-a-time constraint" and Greene & Capella's (1986) discourse moves (see also Givón's (1975: 202-204) "one unit per proposition" and Du Bois's (1987: 826) "one new argument constraint").

In his endeavor to study language in the mind, Chafe (1994: 29) develops the notion of "focus of consciousness" to describe the restricted activation of small parts of our experience at once, resulting from our limited cognitive abilities, which is then "reflected linguistically in the brief spurts of language that will be discussed as intonation units" (with a mean length of 4.84 words in his English corpus). He further argues that besides foci of active consciousness, a periphery of semiactive information can also be expressed through the spreading of an intonation unit across several clauses, a feature of spoken language also attested by the Basic Discourse Units model (Degand & Simon 2009). However, these clusterings are, according to him, the exception against his "one new idea hypothesis" stating that "each intonation unit expresses something different from the intonation unit immediately preceding and following it" (1994: 29), posing that "speakers tend to avoid elaborate syntactic complexities" (1994: 143) but rather prefer a dynamic flow of single pieces of information.

Pawley & Syder (1975/2000: 164) directly draw on this notion to build their “one-clause-at-a-time constraint”, which they consider “a by-product of a more fundamental limit on cognitive processing”, all the while acknowledging the possibility of speakers to plan more complex units. They explain this paradox by (i) the knowledge of conventional expressions and (ii) the use of planning pauses, either silent or filled by a variety of elements including “utterance-initial discourse markers”, conjunctions, tags or hedges (2000: 173). These elements not only occur at boundaries but also more disruptively within clauses, especially in complex embedded clauses which are usually not fully planned in advance. The authors further suggest that this “one-clause-at-a-time constraint” can also be seen as a capacity to structure the flow of speech in full independent clauses, considering that chaining simple clauses is the most efficient and fluent planning strategy as opposed to integrating complex clauses, a view which does not fully overlap with Halliday’s (1987) claim on the intricacy of spoken grammar (see Section 2.1.1). Pawley & Syder (1975/2000) thus provide a first tentative account of the functional relation between planning, information packaging and fluency which has been supported by many authors since (e.g. Swerts 1998 on the discourse-structuring role of filled pauses; Osborne 2011 on the relation between temporal fluency, syntactic structure and informational content).

Lastly, Greene & Capella (1986: 141) pursue the same endeavor of finding “the nature of the fundamental units of communicative behavior” but, unlike the two previous proposals, consider a higher-level unit, viz. the “discourse move”, a supra-clausal cycle of an average duration of 20 seconds – although they admit to length variation and suggest a similarity with Chafe’s (1994) “focus of interest”. They support this claim with experimental findings of an increase of disfluency at the boundary of such units, during which speakers plan the next move. This cyclical rhythm is however attenuated in the case of “simple, familiar, and pre-planned material” (1986: 155), which points to an effect of high cognitive load. Roberts & Kirsner (2000) make very similar observations: their basic unit is the “macroplan”, a topic-driven supra-phrasal structure lasting between 10 to 30 seconds and segmented by pauses, a cycle which they explain by the alternation of a preparation and an execution phase in speech production. Their conclusion, which is also underlying in all the studies discussed in this section, is that speaking and planning can hardly be done at the same time because they compete for cognitive resources: speech “does not become fluent until the macroplanning process is complete and the system’s resources are available solely to speech preparation and production processes” (2000: 153). The notion of macroplanning will be further discussed in Section 2.2.1 in relation to Levelt’s (1989) model of speech production, repair and monitoring.

While authors may disagree on the basic unit of discourse segmentation (clausal or supra-clausal), they all converge in finding regular alternation patterns forming a flow of information units articulated by planning phases. This temporal rhythm is due to the competition of cognitive resources while planning for speaking, and is therefore specific to unprepared speech where the demands (and resulting difficulties) are higher.

2.1.3 Looking back, moving forward: linearity in question

In this section, a selective review of studies on linear and non-linear processes of spoken language will uncover the conceptual overlap between the notions of temporality and linearity, in which the definition of fluency will be grounded (see Section 2.2.4). The main reference in this issue is Levelt (1981: 305) and what he terms “the speaker’s linearization problem”: “The channel of speech largely prohibits the simultaneous expression of multiple propositions: the speaker has a linearization problem – that is, a linear order has to be determined over any knowledge structure to be formulated”. In other words, when planning upcoming utterances, speakers have to organize information in an order that best accounts for the intended message (with all its connections and complexities) within the physical restriction of the articulatory apparatus, viz. a unique channel of sound signal.

Levelt (1981) later stresses that this linearization is modality-independent, a fact of thought and not of speech which also affects writing given that it highly depends on our (limited) ability to manage multidimensional information and to keep complex knowledge in working memory. Indeed, there is a shared discrepancy between the linearity of the language product (speech or text) and the non-linearity of the underlying production processes (speaking and writing).⁶ Concretely, written texts are laid out as continuous, linear sequences of sentences, yet the writing process is not bound to this linear nature: for instance the writer can start at the end or re-write a single sentence several times (e.g. Flower & Hayes 1981; Leijten & van Waes 2013). Similarly, the sound signal is a linear stretch of phonemes, whereas the act of speaking involves many non-linear processes such as planning while speaking (pre-articulation), announcing upcoming material with linguistic or para-linguistic cues (during articulation) or monitoring one’s speech to check for incongruities (post-articulation).

This discrepancy is corroborated by Fortescue (2007) at the morphophonological level: he opposes to the traditional view of a step-by-step construction of utterances a more complex model of non-linear dynamics in speech production, whereby “the end [of the speech chain] need not be predetermined when the beginning is already being produced” (2007: 342), thus being more consistent with psychological reality than the metaphor of communication as a linear channel. Fortescue considers his framework to be compatible with Levelt’s (1989) model (detailed in Section 2.2.1) and claims that it “combines linearity of expression with non-linearity of overall processing” (2007: 350).

To sum up, so far, we have established that speech and writing are similar with respect to the non-linearity of their production process on the one hand, and the linearity of their final product (sound signal and text) on the other. Beside this similarity, differences emerge when looking at the comprehension processes, since listening and reading seem to show different degrees of linearity: while the reception of written texts is not bound by the linear nature of the product (the reader can jump over or retrace back to several sentences, or re-read the same sentence several times), the reception of speech is both linear, in that sound is necessarily delivered through the linear phonological channel which cannot be manipulated by the listener (cf. Danks & End’s (1987) idea of limited control over spoken material, Section 2.1.1), and not-

⁶ “Speaking” in this section refers to the pre-articulatory process of planning speech, not the actual audio output itself, which is rather referred to as “speech”.

linear, thanks to the listener’s abilities to make connections between current and previous utterances or to draw inferences on the nature and salience of upcoming elements. While the linearity of spoken comprehension is fairly basic and technical, the non-linear aspects are somewhat more complex (and more interesting), as can be shown by evidence from phonology (Wauquier-Gravelines 1999), syntax (Auer 2015; Du Bois 2014) and discourse (Rohde & Horton 2014).

Starting from the minimal level, Wauquier-Gravelines (1999) argues that non-linear, specifically retroactive processes co-exist with the linear segmentation of articulatory gestures, which allows the listener to retrospectively reconstruct prosodic boundaries from signal portions temporarily stored and awaiting lexical segmentation. In syntax, Auer (2015: 28) develops the notions of projection (i.e. the ability of “recipients to predict – on the basis of what has been said so far – structural slots in the emergent syntactic gestalt more or less accurately”) and latency (i.e. relating the structure of new and preceding utterances). He identifies different types of these synchronizing phenomena such as co-constructions, failed projections, final overlaps or structural resonances. Du Bois (2014) focuses on the latter in an emergentist framework of spoken grammar called dialogic syntax: he studies the “dynamic emergence of structural resonance”, that is “the catalytic activation of affinities across utterances” (2014: 360) in the form of cross-turn parallel structures or “diagraphs”. Lastly, at discourse-level, Rohde & Horton (2014) provide experimental evidence of listeners’ pragmatic expectations about coherence relations, triggered by verbs and connectives (*so* or *because*).

It is challenging to tell apart the non-linear processes which are part of production and those which belong to comprehension, given the high inter-dependence of these aspects. What can be said however is that spoken production (speaking) is entirely non-linear, while spoken comprehension (listening) involves both linear and non-linear processes. It remains now to address the mapping between this set of distinctions, and those drawn from the notion of temporality: is linearity equal to temporality? What aspects of spoken and written language are concerned by these notions? Do they partly overlap, and when? We have established in the previous sections that (i) speech appears to the listener as an evanescent on-going process and that (ii) speech flow is delivered in cyclic patterns which correspond to the alternation between planning and speaking. It would therefore seem that all aspects of spoken language (production, product and reception) are bound by this temporal nature (i.e. constantly fluctuating and unstable). On the contrary, written production is not time-bound at all (texts can be “worked over”), neither is the product (an eternal static object) nor the processing (self-pace reading). These conclusions are summarized and graphically represented in Table 2.2.

Table 2.2: Conceptual representation of temporality and linearity across speech and writing

	Linear	Non-linear	
Temporal	speech	listening	speaking
Spatial	text	writing reading	

We can see a first divide (represented by the red square) between the temporal and spatial dimensions: the former contains all three aspects of spoken language, while the latter those of written language. Across this space-time line, cross-modal similarities exist: both speech and text are linear (blue square) and both speaking and writing are non-linear. However, listening stands out in the middle of the table, thus representing its intermediary status involving both linear and non-linear processes, while reading is entirely non-linear. I hope to have shown convincing evidence (including empirical evidence from a diversity of research frameworks) that the notions of linearity and temporality – although quite abstract and theoretical – adequately describe the specificities of the spoken modality with respect to issues of planning, delivery and overall processing.

In the remainder of this thesis, I will no longer address the spatial/written dimension but focus on temporal/spoken processes, which I repeat here for convenience: spoken production (pre-articulation) is non-linear since it has to handle past, current and future utterances simultaneously (which results in the occurrence of disfluencies); the sound signal is bound by the linear phonological channel which forbids simultaneous articulation; listening is partly bound by the linearity of the sound signal (especially in comparison with the freedom of reading) but can still make use of non-linear cues such as salience, expectations or resonances. The difference between speech and writing thus appears to lie in this degree of freedom of the reading process with respect to the linearity of the written product, as opposed to the partial dependence of listening, in addition to the more fundamental divide between temporality and spatiality.

2.2 Fluency, disfluency and disfluencies: defining a multi-faceted construct

I will now turn to the definition of fluency and disfluency proper, and the presentation of (a selection of) major annotation models which are prevalent in the design of the present approach. 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 (cf. Section 1.2). This section rather focuses on works which are (i) relevant to the present study and (ii) representative of theoretical differences in the terminology (Section 2.2.1) and definition of (dis)fluency, whether holistic (Section 2.2.2) or componential (Section 2.2.3), and within the latter whether qualitative, quantitative or both. Further distinctions will be discussed in order to provide a structured state of the art upon which my own definition of (dis)fluency is built. 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 thesis (see in particular Chapter 7).

2.2.1 Disfluency or repair? Levelt's legacy

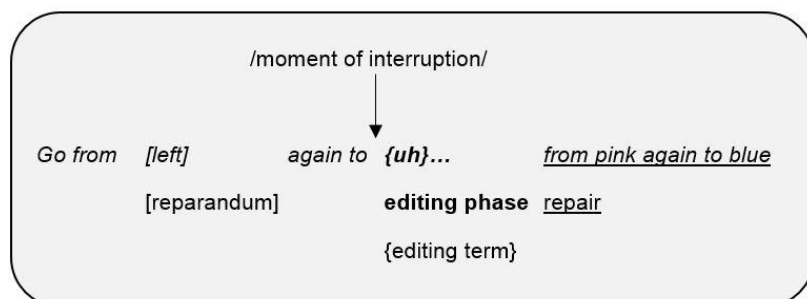
Like other fields in linguistics, fluency studies suffer from a lack of consensus at the level of definition, which is “notoriously difficult” to agree upon (Hasselgren 2002: 147), but also at the lower level of terminology. As mentioned in Section 1.2, research on fluency started with the study of pauses 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 this latter 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 because of two reasons: (i) the notion covers different meanings, which need to be disentangled for the sake of clarity; (ii) it is often associated with Levelt’s (1983, 1989) larger model of speech production, which remains referential 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 large 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”.⁷ In both cases (large 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. (2013a, 2013b) or Dutrey et al. (2014). 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



⁷ See Section 7.1.3 for a detailed review of the relation and partial overlap between repair and reformulation.

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).

As I said, 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 (i) the selection of information relevant to the realization of the communicative intention, and (ii) 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 (cf. Section 2.1.2).

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 (*I went to to London*). Overt repairs correspond to the narrow definition of repairs presented above, while the larger 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). These terms are in tune with the notion of non-linearity developed in Section 2.1.3: the process of monitoring either one’s own or someone else’s speech involves playing with and moving along the linear articulatory channel, either backwards for retracing and reformulating, or forwards by announcing

upcoming material or, from the listener's perspective, anticipating new or complex material based on hesitations or repetitions.

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 7.1.3 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 7.

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 rather tends to disappear from the literature, in spite of the quality of the overall model which connects issues of linearity, temporality and cognitive processing. 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 7 where further details will be provided.

2.2.2 Holistic definitions of fluency

As mentioned in the introduction to Section 2.2, fluency and disfluency have been studied in many frameworks which present major differences, regarding either terminology (see previous section), definition or annotation. The first group of references on fluency is concerned with definition and considers the concept as a holistic assessment or impression over the general production of a speaker, usually focusing on one aspect of language, although the specific aspect may differ from one definition to another. I consider holistic those approaches that 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). I have gathered from the literature four of these central concepts, namely automaticity, flow, efficiency and confluence, which are all reviewed in the following.

As mentioned in the introductory chapter to this thesis (Section 1.2), the study of fluency and disfluency originates from (i) the study of pauses in speech and (ii) the interest in second language learners, which might 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 the automaticity of each production component (see Section 2.2.1) as “a main condition for the generation of uninterrupted fluent speech”. This first group of definitions is therefore strongly cognitive and speaker-oriented.

The second notion, which is also present in 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 found negatively (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 by its metaphorical and descriptive power, which also relates to the temporality of spoken delivery discussed in Section 2.1.2: 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, so much so that definitions such as “steady tempo” might not be accurate in this regard.

The next 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 acquisition (henceforth L2): 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 systematically distributed in efficient strategies seems promising and realistic beyond the realm of L2 fluency.

Lastly, McCarthy (2009: 19) introduces the notion of confluence to refer to the interactive dimension of fluency, which he argues to be lacking in most frameworks: “the co-creation of fluency in a conversation rather than the fluency of an individual speaker. Judgments of fluency which lack such an interactive dimension may therefore be considered as providing only a partial picture of the speech event”. McCarthy (2009) focuses on turn-taking to study dialogic synchronization or confluence, yet such strategies concern many other conversational features including disfluencies (cf. Clark 2002 discussed in Section 2.1.1).

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 (cf. here the rather opaque use of “lucid”). Other synthetic and multidimensional definitions are developed in the following sections where they will be associated to more componential approaches to fluency.

2.2.3 Componential approaches to disfluencies

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. We see that these terms do not fully map with the holistic-componential distinction, which rather refers to the methodological approach, either encompassing undistinguished variables into a global impression, or investigating specific features separately.

2.2.3.1 Qualitative dimensions of perception

If the literature on fluency was ranked on a scale with holistic and componential approaches at each extreme of the continuum, 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 distinct variables of perception.

Fillmore (2000: 51) identifies four dimensions of fluency: (i) 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; (ii) 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”; (iii) 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 (iv) 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 judge 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, only one of which being fully measurable: 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). 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 admits to 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 as very broad, combining aspects of both speaking and listening in a complex but partitioned model.

2.2.3.2 Quantitative dimensions of production

All the authors discussed so far aim at defining the concept of fluency in more or less holistic and/or qualitative ways, with the proposals of the previous section being closer to a componential, partly quantitative approach. I have already hinted above (Section 2.2.2) that these broad definitions 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 fluency are abundant in the literature and reflect a diversity of theoretical approaches and research agendas in many languages and data types. It seems however that these works focus on parts over the whole, thus neglecting the first step of defining the global concept (see Section 2.2.3.3 for an integrated, qualitative-quantitative and holistic-componential proposal). Still, they remain greatly valuable to the methodological approach taken here and are therefore reviewed in detail in the following sections, where they are grouped according to their underlying conception of fluency and disfluency, namely “disfluencies as errors” (Section 2.2.3.2.1) or “disfluencies as ambivalent devices” (Section 2.2.3.2.2).

2.2.3.2.1 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 indebted to the development of computational linguistics. 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 which are often associated with a large reference corpus. 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 worked on three corpora of the Linguistic Data Consortium representing different data types (e.g. human-human and human-computer dialogues) and 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”. Concretely, 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 incompleteness.⁸ 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* (cf. Section 2.2.1 for more details). Labels, types and annotated examples can be found in Figure 2.2, taken from her thesis.

Her corpus analysis showed that (i) disfluency rate is dependent on utterance length, an effect which is itself dependent on the corpus or interaction settings; (ii) disfluencies mostly affect utterance-initial words; (iii) 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 (comparing human-human and human-machine dialogues) 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) and similar across speakers’ age. Moniz

⁸ 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” (see Section 3.3.1).

(2013) integrated an annotation of disfluencies following Shriberg (1994) to her corpus of Portuguese where she also investigated prosody and syntax focusing on interrogatives. She concludes on the crucial role of prosody in fluency rating tasks, although prosodic cues are less discriminating in dialogues than in monologues. Lastly, Christodoulides et al. (2014) based their automatic annotation tool for French corpora on Shriberg’s (1994) original typology to which they added a high-precision morphosyntactic and multi-word-unit tagger.

Figure 2.2: Shriberg’s (1994: 57) annotation model

Symbol	Explanation	Example	Section
Region-delimiting			
[]	onset RM, offset RR	(see all examples below)	4.3.4.1.1
.	IP	(see all examples below)	4.3.4.1.2
Syntactic-word			
r	repeated word	she she liked it [r . r]	4.3.4.2.1
s	word in substituted string	she my wife liked it [s . s s]	4.3.4.2.2
i	inserted word	she liked really liked it [r . i r]	4.3.4.2.3
d	deleted word	it was very she liked it [d d d .]	4.3.4.2.4
Extra-syntactic-word			
f	filled pause	she uh liked it / she uh he liked it [f] [s . f s]	4.3.4.3.1
e	explicit editing term	she sorry he liked it [s . e s]	4.3.4.3.2
p	discourse marker	she liked well she liked it [r r . p r r]	4.3.4.3.3
Inter-sentence-word			
c	coordinating conjunction	she saw it and and she liked it [c . c]	4.3.4.4.1
Diacritics			
-	word fragment	she li- he liked it [s r- . s r]	4.3.4.5.1
~	misarticulated word	shle she liked it [r~ . r]	4.3.4.5.2
^	contracted word	she’d she’ll like it [r^s . r^s]	4.3.4.5.3
=	substituted-string fragment	she thought highly she liked it [r s s= . r s]	4.3.4.5.4

Another largely spread framework is that of the Switchboard corpus of telephone conversations and the annotation model by Meteor et al. (1995): in the perspective of “cleaning” transcriptions, they provide a three-step annotation, covering (i) “non-sentence elements” (filled pause, explicit editing term, discourse marker, coordinating conjunction, aside), (ii) “slash-units” (tagged as complete or incomplete in case of mid-utterance interruptions) and (iii)

“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 pressure. In that sense, Clark & Wasow (1998) are closer to functional accounts of disfluencies (see next section) rather than the “disfluency-as-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: one of the steps in this endeavor is to detect and remove disfluencies, in addition to other annotation layers such as nucleus-satellite relations or topical boundaries. 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, dedicating a specific section to the complex disambiguation of discourse markers such as *like* or *so* and allowing 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, which they found to be improved by lexical cues either alone or in combination with acoustic cues.

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 $\kappa = 0.924$. We can also see that their approach is 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).

2.2.3.2.2 *The functional ambivalence of disfluencies*

Pallaud et al.'s (2013a, 2013b) 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. (2013a: 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, see Section 2.2.1), and further distinguished according to the grammatical class of the item affected by the interruption (word, determinant, phrase, etc.). The categories identified by Pallaud et al. (2013a, 2013b) 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 (see Section 7.1.3), as can be seen in Table 2.3 which reproduces their annotation model.

According to this grid, each disfluency or self-interruption is assigned six labels: for instance, in an utterance such as *you are uh you are tired*, the *reparandum* is tagged as (i) temporary interruption, (ii) phrase *reparandum*, (iii) lexical word; the break type is (iv) filled pause; the *reparans* is (v) phrase restart and (vi) repairing through repeating. Although potentially more fine-grained than others, this model presents the disadvantage of having to combine multiple labels in lengthy, opaque tags (such as *R,P,lw,fp,pr,rp* for the example detailed above) 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 qualify the opposition between the two accounts discussed so far, namely “disfluencies as removable errors” vs. “disfluencies as functional devices”.

Table 2.3: Annotation of self-interruptions in Pallaud et al. (2013a, 2013b)

Reparandum	
Reparandum type	Temporary interruption
	Definitive interruption
Reparandum	Word reparandum
	Phrase reparandum
Lexical type	Tool word
	Lexical word
Break type	
	No interval
	Silent pause > 200ms
	Filled pause
	Discursive connector
	Parenthetical statement
	Truncation repetition
Reparans	
POS of the reparans	No restart
	Word restart
	Determinant restart
	Phrase restart
	Other restart
Reparans type	Continuing the item
	Repairing without change
	Repairing through repeating
	Repair with change in the truncated word
	Repair with multiple change

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 thesis adopts a similar functionally ambivalent perspective.

2.2.3.3 *Qualitative-quantitative approaches to both perception and production*

This section focuses on a recent milestone in the study of first- (L1) and second-language (L2) fluency, namely Götz's (2013) comprehensive approach to both production and perception of English speech. 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.⁹

First, her definition of L1 fluency is rather consensual and synthetic, largely borrowing from Lennon (2000) and other holistic definitions (Section 2.2.2): “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). A first observation at this defining stage is that production and perception are both associated to the speaker's perspective, and therefore potentially more accessible to the analyst, as opposed to 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.

Turning to L2 fluency, Götz (2013) notes two trends of definition: either fluency as communicative effectiveness in a particular situation (Kennedy & Trofimovitch 2008; Bygate 2009) or fluency as nativelike behavior (her approach, cf. also Denke 2009). She warns against the reference to a monolithic native speaker which is “mythical” (Lennon 1990: 392), albeit “useful” (Davies 2003: 214): native speakers are not all equally fluent, and even a single individual can vary depending on contextual, emotional or other external factors. Götz (2013) still follows this comparative L1-L2 line of research whereby corpus data is taken to give access to the ideal native speaker, following Esser's (1993) notion of “abstracted corpus norm” (see also Mukherjee 2005). In this sense, any investigation of L1 fluency, as in the present thesis,

⁹ 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 rather qualifies as componential.

could therefore be connected to L2 fluency as well, provided the data and observed variables are comparable.

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 is less negatively connotated 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 thesis. 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 full list and categorization of fluencemes can be seen in Table 2.4.

Table 2.4: Götz's (2013) three-fold typology of fluencemes

Productive fluency	
Temporal variables	Speech rate
	Mean length of run
	Unfilled pauses
	Phonation/time ratio
Fluency-enhancement	Speech management strategies (repeats, filled pauses)
	Discourse markers
	Smallwords
Formulaic sequences	
Perceptive fluency	
	Accuracy
	Idiomacity
	Intonation
	Accent
	Pragmatic features
	Lexical diversity
	Register
	Sentence structure
Nonverbal fluency	
	Gestures
	Facial expressions
	Body language
	Looks
	Emblems

We can see the diversity of elements under consideration, covering almost every aspect of language (prosody, lexicon, discourse, pragmatics, non-verbal communication). This typology does not fully map with Segalowitz's (2010) own tripartite definition of fluency (cognitive, utterance, perceived, cf. Section 2.2.3.1): 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 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 shown repeatedly 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.

Götz bases her categorization of fluencemes on the results of a pilot study using corpus data, questionnaires and rating tasks, from which she draws the following methodological conclusion: "Productive fluencemes can be analyzed quite easily by way of corpus analyses, but do not seem to be able to be rated by the raters accurately" (2013: 87). She therefore focused her corpus study on these productive fluencemes, while perceptive fluencemes were investigated experimentally (leaving out nonverbal fluencemes). This combination of corpus and experimentation has been promoted since De Mönnink (1997) (see also Arppe & Järviö 2007; Gilquin & Gries 2009) and is now a major trend in cognitive linguistics and in the study of fluency in particular. This multi-method approach proves especially fruitful in the study of pauses as carried out by Candéa (2000) and Lundholm (2015).

To sum up, Götz's proposal is valuable for many reasons: she combines definition and typology in a single model; she accounts for a wide range of well-defined phenomena; 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 connotated "disfluencies". One limitation which prevents direct use of this typology in the thesis is the lack of technical guidelines in the perspective of corpus annotation: Götz extracted a selection of investigated features (semi-)automatically without directly annotating the data, a methodological choice which is time-saving but perhaps questionable on aspects of replicability (the extraction process remains opaque), exhaustivity and granularity (some unexpected forms or structures may be only identifiable manually). 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.

2.2.4 Bridging the gap: fluency as non-linearity

It now remains to situate the present approach to fluency and disfluency within the vast body of research discussed in the previous sections. Instead of proposing a whole new definition or directly borrowing an existing one, I suggest to combine all these frameworks and keep only the elements that are theoretically and methodologically valid, innovative and relevant to the present research purposes. Concretely, my definition takes up (i) the forward-backward (or covert-overt) distinction from Levelt's (1983) notion of repair, (ii) the notions of flow and efficiency from the holistic definitions, (iii) the interest for quantitative measures of production and their functional ambivalence from componential approaches to disfluency and (iv) the notion of fluenceme in particular from Götz (2013). I therefore consider fluencemes to be discrete devices which function as **signals of non-linear processes of speech production and perception**, following 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. In other words, the occurrence of a fluenceme reflects some non-linear production process on the speaker's part (e.g. the need to edit a previous utterance) while triggering a non-linear interpretation process on the listener's end (e.g. retrieving a previous utterance stored in working memory). **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.

Overall, the main conceptual influence is the comparison of speech and writing along the axes of temporality and linearity: while many frameworks and definitions consider fluencemes as disruptions to be removed in order to obtain a cleaned and continuous (in other words, written-like) stretch of talk, the present approach rather strives to define spoken fluency by taking into account the temporal and non-linear nature of spoken language, instead of imposing a written-like standard. Coming back to the spacetime metaphor developed in the introduction to the thesis, I hope to have made it clear that **speech is entirely temporal and partly bound by the linearity of its channel, although fluencemes can be seen as introducing some spatiality in the form of non-linear moves**.

While Segalowitz (2010) already called for the use of operational definitions instead of metaphorical thinking, I would like to suggest that broad metaphorical definitions can still be useful, provided they are combined with an annotation model for corpus-based analysis, so that theory and data can feed each other. Such a mapping between definition and annotation is at the core of the present approach, which is why I conclude this section on definition with an overview of the typology of fluencemes (Crible et al. 2016) used as the analytical grid for the present analyses.¹⁰ Table 2.5 indeed shows the ten fluencemes, three related elements and three “diacritics” (the term is taken from Shriberg 1994) which will be investigated, with their abbreviated tags and examples. All these elements will be defined at length in Chapter 4, and analyzed in Chapter 6.

¹⁰ The elaboration of this typology (Crible et al. 2016) is the product of joint work with my colleagues A. Dumont, I. Grosman and I. Notarrigo, whom I wish to thank here.

Table 2.5: Typology of fluencemes investigated in this study

Tags	Fluencemes	Examples
UP	unfilled pause (sec.)	(0.380)
FP	filled pause	<i>uhm, uh, euh</i>
DM	discourse marker	<i>so, because, well, I mean...</i>
ET	explicit editing term	<i>oops, what is it?...</i>
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”

We can already say that this typology is indebted to componential and quantitative approaches to (dis)fluency, mainly Shriberg (1994) and Götz (2013), and that it does not involve *a priori* judgments of the relative fluency of its components, in line with the claim of functional ambivalence. It remains to be established how this conceptual framework (definition and typology) can be related to interpretations of fluent vs. disfluent uses of the fluencemes and thus answer the research questions of this study (see Section 2.4), which is the topic of the next section in this chapter.

2.3 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.

In the following sections, key notions of this framework will be defined (Section 2.3.1) and their application to the study of (dis)fluency will be presented, focusing especially on schemas (Section 2.3.2), context (Section 2.3.3) and frequency (Section 2.3.4).

2.3.1 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. In other words, language is not independent of other cognitive systems and, within the language system itself, components (e.g. grammar, lexicon, phonology) are inter-related. 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 tenants of the usage-based framework, especially the central roles of frequency and context, lie at the core of the present study of (dis)fluency, as will be developed in the following sections.

Many theories adhere to the usage-based conception of language, such as Construction Grammar (Goldberg 2006) mainly applied to grammatical structure or the Competition Model (Bates & MacWhinney 1989) for sentence processing and acquisition. In the remainder of this section, I will illustrate the key notions of usage-based linguistics through one of its major representative framework, viz. Cognitive Grammar. Cognitive Grammar was developed by Langacker (1987, 1998a, 1988b) in reaction to the generative tradition, starting with the primary assumption that “language evokes other cognitive systems and must be described as an integral facet of overall psychological organization” (1988a: 4). This integrated view echoes the present conception of fluency in that grammaticality judgments are seen as gradual and reflecting “the subtle interplay of semantic and contextual factors” (1988a: 5). Langacker further argues that the distinction between grammar and lexicon or semantics is arbitrary since all units are symbolic and can therefore integrate aspects of other linguistic levels (phonological,

morphological, lexical, etc.). These symbolic units of language are characterized by their schematicity, that is, they are represented in schemas which are “the commonality inherent in multiple experiences to arrive at a conception representing a higher level of abstraction” (Langacker 2013: 17). For instance, the [consonant-vowel-consonant] canon is instantiated by *tip*; the morphological derivation [process –er] is instantiated by *sharpen**er*, etc.

Schemas and lower-order units are structured in complex networks: the same unit can be categorized as (i) the instantiation of a more abstract schema or (ii) a schema itself in relation to more specific units. These schemas, like all units of language, are at least partly conventional and (recognized as) shared in a particular linguistic community. They 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). The resulting routines only contain the common base between all usage events and are therefore less detailed: “A unit corresponds to just selected aspects of the source events, and the commonality it reflects is only apparent at a certain level of abstraction” (2013: 220). In sum, the central notion of schema is characterized by its flexible degrees of abstraction and of entrenchment, and constitutes the basic unit for the acquisition, use and innovation of language. As a usage-based model of language, Langacker’s (1987) Cognitive Grammar starts from the observation of actual patterns, their full potential of realizations and the factors that influence them in order to define more abstract units and only then build general principles.

Before concluding on this brief presentation of usage-based linguistics (illustrated by Cognitive Grammar), I would like to raise a number of potential issues regarding the ambitions and methods of the framework, based on Divjak (2015). She identifies four challenges which question the methodological assumptions of usage-based linguistics in its current form, namely (i) the variability of corpus annotation, (ii) the (over-)reliance on frequentist statistical techniques, (iii) the lack of cognitive realism in statistical models and (iv) the artificiality of experimental data. The first two are of particular relevance in the present method and analyses which involve fine-grained discourse annotation and frequency-based interpretations of results. While precautions have been taken (see Sections 4.2.2.3 and 4.3.5 for an assessment of annotation reliability), these limitations need to be borne in mind while reading the remainder of this thesis.

2.3.2 From schemas to sequences of fluencemes

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 a network with particular instantiations and other related schemas. In the perspective of fluenceme 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 (see Section 4.3.6 for the operationalization of this notion to corpus annotation).

The similarity between sequences and schemas mainly concerns 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: (i) the identical repetition of *the*, (ii) the filled pause *uh* and (iii) 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 determinant 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 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 account for this tendency, Shriberg (1994) integrated in her model a "Type Classification Algorithm" where patterns or sequences of fluencemes can be grouped in eight smaller classes such as the REP (for repetition) class corresponding to a repetition with optional coordinating conjunction and filled pause, or the HYB (for hybrid) class which has to contain more than one element among substitutions, insertions and deletions.

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. Unlike Shriberg (1994) who established pre-defined classes or patterns to summarize her data, I will suggest a more bottom-up approach: corpus annotation follows the typology introduced in Section 2.2.4 and, once extracted, co-occurring fluencemes are grouped in

sequences or “macro-labels” of different degrees of abstraction (see Sections 4.3.6 and 4.4.2 for further details on the annotation, extraction and macro-labels of sequences).

2.3.3 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 and non-linguistic patterns are processed and acquired in an integrated way, from phonetics to pragmatics, and undergo the influence of linguistic co-text and extra-linguistic context. Following this claim, items and structures integrate to a certain extent information from their local (i.e. linguistic co-text) and global (i.e. communicative context) environment in their meaning and use, so much so that the same pattern in different contexts could be used and perceived differently.¹¹ 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 will be studied in relation to their co-text and context, that is, register variation. In this section, I will define the aspects of context relevant to this study and situate the present approach to fluencemes in previous variationist research.

In my analysis, I will only take into consideration some aspects of local and global context, leaving out background knowledge and any other information which is not available in the metadata or annotations. The relevant aspects of the linguistic context (e.g. combination patterns, position, meaning-in-context) will be presented in the methodology (Chapter 4), along with the operationalization of the register variable. Still, 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 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 refining hypotheses on casual

¹¹ Cutting (2008) refers to these types of information as “co-textual context” and “situational context”, respectively.

vs. formal registers: while extreme settings might be expected to be sharply contrasted in the production of fluencemes (e.g. more frequent in impromptu conversation than scripted news broadcast), intermediary registers such as interviews or professional encounters may actually 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 (cf. Section 2.1.1). De Jong et al. (2012: 136-137) seem to confirm such an effect of heightened attention in elicited monologues of varying complexity (general vs. abstract topic, formal vs. informal, descriptive vs. persuasive) where they found that “cognitively more demanding tasks lead to more repairs” but also to a higher functional performance (i.e. communicative success) and a more diverse vocabulary.

The effect of task complexity, in particular topic familiarity, on fluency was investigated at length by Merlo & Mansur (2004), who tested the hypotheses that (i) familiar discourse should be more fluent than unfamiliar discourse due to a greater memory strength of the topics in the former and that (ii) the distribution of fluencemes should vary across the different referent categories in descriptive discourse (e.g. whole, parts, attributes). Covering a broad range of phenomena (fillers, pauses, prolongations, some discourse markers, repetitions, false-starts), the authors found that different types of descriptive discourse were more prone to certain types of fluencemes than others (e.g. more fillers in description of parts, fewer repetitions in comments). Merlo & Mansur (2004) also observed that the overall frequency of fluencemes is not influenced by topic familiarity, contrary to their expectation, but rather by the on-going step of the discourse: for instance, descriptions of parts and attributes increase the occurrence of fillers (which points to lexical difficulty), but decrease that of repetitions, which means that these two types of fluencemes may be functionally different. A final result of particular relevance for the present thesis is the lack of statistical significance of what they term “lexical pauses” (i.e. discourse markers such as *well*, *look*, *for example*, *that is*), found frequently and indiscriminately in all discourse parts, which they explain by their role as introducers of new topics or new steps in an on-going topic. Overall, their study does not confirm an effect of memory strength on fluency but rather suggests an effect of discourse tasks (e.g. objective description of attributes, subjective evaluation) and their respective psycholinguistic demands and difficulties.

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. 2013a, 2013b), 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 exception 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. 2017a in English-French), revealing a great variety of forms, from vocalizations (English *uh*, French *eu*) 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 (Section 3.1.3).

Overall, this state of the art seems to call for more research tackling the crosslinguistic 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 behavior within and across particular texts. The typology by Crible et al. (2016) introduced in Section 2.2.4 (and detailed in Section 4.3) was elaborated and tested on multilingual (English, French) and multimodal data (spoken and sign language) across various registers, bearing in mind the variationist perspective of this research.

2.3.4 Accessing fluency through frequency

As I said before, it is not a realistic ambition of this thesis to provide a relative measure of fluency for each occurrence of a fluenceme or even each speaker in the corpus, because of the inability of corpus analysis to access perceptive information in addition to the high variability and uncontrolled factors that need to be taken into consideration in such a rating task on authentic data. Rather, a more feasible endeavor would be to relate corpus observations to fluent and disfluent tendencies in different registers. Such a program implies relying on the significant association between linguistic and contextual variables on the one hand, and on frequency information on the other. In this section, I will question the relation between fluency and frequency and present the approach adopted in this regard.

We have seen (cf. Section 2.2.2) that 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 should be less cognitively demanding for both production and reception and trigger limited hesitations given their high accessibility for the participants. On the other hand, rare patterns should strike as less automatic in production and unexpected in reception, 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* should be perceived more markedly (possibly more negatively) in a news broadcast than in a casual conversation. Such observations of frequency across registers can be refined by taking into account a large array of variables, integrating in the schema not only the particular sequence but also its syntactic position, co-occurrence tendencies and pragmatic interpretation, at various degrees of abstraction, thus matching the present functional, relative and flexible definition of (dis)fluency.

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.¹²

This principle of cognitive corpus linguistics is especially interesting if we consider not only textual frequency but also “conceptual frequency”, a distinction proposed by Hoffmann (2004) who defines the latter as the 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. I will therefore pursue such an inclusive approach to the linguistic categories under scrutiny in order to provide an exhaustive account of all fluencemes in the typology, their relative position in the typology (paradigmatic) and their combinatory patterns (syntagmatic).

This use of corpus distribution as mirroring not only language in use but also language in the mind raises questions about the link between frequency and prototypicality. A prototype is the most central and representative member of its category, that is, the one that is more directly activated and attracts one’s attention (Rosch 1975). Such a definition equates prototypicality with cognitive salience (e.g. Radden 1992, Janda 2010), which can in turn be related to frequency of use through the mechanisms of entrenchment: the more frequent a particular unit for a given community or individual, the more easily this unit will be activated

¹² This view of language reinforces the choice of sequences of fluencemes as the basic unit of analysis, as opposed to individual fluencemes (cf. Section 2.3.2).

in context. In this view, prototypes should be produced and perceived quite automatically – therefore more fluently – thanks to their high frequency and degree of entrenchment.

However, Gilquin (2006, 2008), Shortall (2007) and Glynn (2010), among others, provide evidence of the possible mismatch between corpus frequency and prototypicality or salience in syntax and semantics, where they found that the most frequent constructions are not necessarily the most salient. Glynn (2010: 14), for instance, argues that “[i]n terms of semantic content, most frequent often equates least semantically important, where rarity, or marked usage, indicates greater semantic importance”. Similarly, while Gilquin (2006) acknowledges some connection between prototypicality and frequency, she considers their equation to be a “methodological shortcut” (2006: 168) and concludes that frequency is not the whole story of prototypicality, which is a more complex concept combining neurological principles, linguistic usage, abstract thinking and other aspects of salience.¹³

Another 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 sum up, frequency has been linked to entrenchment, salience and prototypicality, although not consensually. 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 frequency and prototypicality, although authors are still struggling to clarify the situation. I intend to treat frequency information 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, thus contributing to the theoretical and methodological debate on frequency effects by converging multiple kinds of evidence across the different chapters. While I will strive to keep mind and language apart, notably by avoiding any reference to prototypicality in my analyses, I will still pursue the investigation of fluency as frequency, among other hypotheses which are laid out in the next section.

2.4 Hypotheses: combination and variation of fluencemes

In the previous sections, the definition and approach to (dis)fluency was developed in relation to the speech-writing continuum, to previous research on fluency and disfluency and to the usage-based framework. It was made clear that sequences of fluencemes constitute the first

¹³ Geeraerts (1998: 222) considers frequency of occurrence in a more nuanced way as a “heuristic tool in the pinpointing of prototypes”, thus acknowledging both the imperfection of the method and its useful value for lack of better measurements.

level of analysis in this thesis, investigating in particular their combinatory patterns and contextual variation (in language and register) in order to uncover frequency-based tendencies which could be tentatively related to different ends on the fluency-disfluency scale. At this sequence level, a number of hypotheses emerge from each of the three main notions taken from Section 2.3, namely combination, variation and frequency. They will be stated in the following.

Isolation vs. combination: 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 (cf. Section 2.3.2) tends to suggest that fluencemes do occur more frequently in clusters than in isolation, which is in part due to the high frequency and pervasiveness of unfilled pauses (Section 6.1).

Variation across registers: the rate and combination of fluencemes will be systematically compared across the different subcorpora (i.e. eight registers in two languages), following the hypothesis that unplanned discourse is cognitively more demanding on the speaker's production processes and should therefore lead to more frequent and more varied fluencemes than planned speech. Types and sequences of fluencemes which are specific to informal registers (unplanned interactive dialogues) and rare or absent from more formal registers could be considered as typically and relatively "disfluent", and vice versa (sequences specific to formal registers are relatively "fluent"), while fluencemes showing no significant difference across registers should be more ambivalent, and in need of further investigation with additional sources of information. It should be noted that such interpretation of the fluency of sequences can only be (i) relative to other sequences extracted from the data and (ii) 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 and paradigmatic 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 (Section 6.2).

Variation across situational features: the explanatory power of register variation can be improved by a more fine-grained categorization of the speaking tasks in the corpus according to situational features characterizing each register along six gradual dimensions such as degree of planning or number of participants (see Section 4.1.3 for more details). In this refined approach to contextual factors, different speaking tasks showing all or almost all of the same dimensions can be expected to behave similarly with respect to fluenceme rate and distribution. This method also makes it possible to pinpoint the impact of specific features of communication: for instance, the effect of physical co-presence on fluency can be identified by comparing conversations with private phone calls. Planning and interactivity are expected to be particularly influential because of their relation to cognitive demands (higher in unplanned discourse), time pressure (partly for turn-holding in interactive situations) and participants' synchronization (fewer cognitive resources for recipient design in unplanned speech, more synchronizing devices available in interactive situations) (Section 6.2).

The disfluency of intermediary registers: we saw in Broen & Siegel (1972), Halliday (1987) and De Jong et al. (2012) that the situation might be more complex than a planned-unplanned

divide, namely 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, as opposed to casual and unplanned situations where fluencemes might be equally frequent but less marked and generally unnoticed (cf. Section 2.3.3). 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 (Section 6.2).

Fluency as frequency: the fluency-as-frequency hypothesis will be pursued as a methodological research question. I will be looking for converging evidence that would support the proposed heuristic equivalence between high frequency and fluency (and its negative equivalent, low frequency with disfluency). This underlying goal will be fed by the analyses throughout Chapter 6.

Variation across languages: no specific evidence in the literature motivates any major expectation of crosslinguistic variation between English and French. The two systems will therefore be compared in a more exploratory manner, looking for any language-specific patterns and uncovering potential “universals” of (dis)fluency.

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 striving 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, which is why this thesis takes a special interest in one type of fluenceme, namely discourse markers, which provides a more thorough and qualitative level of analysis bringing us closer to the targeted cognitive-functional scale of fluency. The category of discourse markers and its relation to (dis)fluency are the topic of the next chapter.

Chapter 3: Discourse markers in spoken language

Introduction to the chapter

The present research takes the notion of (non-)linearity as central to a definition and model of (dis)fluency which incorporates the specificities of spoken language. In the previous chapter, (non-)linearity was associated with backward- and forward-looking moves during speech production and comprehension, such as editing previous utterances or anticipating new or complex material. This view directly connects (dis)fluency with discourse structure in that fluencemes are considered to mark a number of directional operations relating segments and utterances together, either for local (i.e. contiguous unit pairs) or more global organization (i.e. long-distance relations, higher-level discourse cycles). Consequently, discourse markers (DMs), as expressions dedicated to the management of “local and global content and structure” (Fischer 2000: 20), appear highly central and relevant to the study of (dis)fluency as (non-)linearity. As a result, the thesis focuses on discourse markers among the typology of fluencemes.

Another 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, although it bears functional similarities with DMs (Swerts 1998), conveys a much vaguer meaning. This semantic-pragmatic and discourse-functional layer of analysis is all the more welcome insofar as the present corpus-based approach does not include any perceptive validation (in the form of experiments or ratings) nor a comparison with control groups (as is the case for L1-L2 comparative studies). Instead, 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. What it will not tackle, however, are aspects and dimensions of DMs which are largely studied and very interesting but which only bear an indirect relation to fluency or are not relevant to the view of (dis)fluency as (non-)linearity. These exclusions concern, among others, the link between DMs and emotionality (Romano & Cuenca 2013), common ground (Fetzer & Fischer 2007), persuasion (Blankenship & Holtgraves 2005; Hosman & Siltanen 2011), sociolinguistic variation across language varieties (Dostie 2009) or speakers (Beeching 2007), grammaticalization and diachronic studies (Traugott 1995; Waltireit & Detges 2007). The variety of approaches to DMs is as wide as that of fluency research (perhaps even wider), so much so that these restrictions in scope are unfortunate but necessary.

DMs share with fluency a lack of consensus regarding the definition of the category, 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 (Section 3.1). Section 3.2 will focus on (corpus-based) models designed to capture the multifunctionality of DMs, thus discussing the merits and differences of each framework before introducing the selected taxonomy and how it handles the specificities of spoken language. The relation of DMs to (dis)fluency (and its relative absence from the literature) will be developed and connected to (non-)linearity in Section 3.3. Finally, Section 3.4 will lay out the research questions and hypotheses specifically related to the variation of DMs (analyzed in Chapter 5) and the interrelation between DMs and other fluencemes (Chapter 6).

3.1 What are discourse markers?

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. Words tend to acquire new uses, especially pragmatic, expressive and intersubjective meanings created and shared for the needs of spoken conversation (Traugott 1995: 2).¹⁴ Another reason for the lack of consensus concerns the many different frameworks which have investigated the category 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. It is nevertheless possible to identify a common core of features usually shared among authors. These general characteristics will be presented in Section 3.1.1 in order to grasp a first understanding of the expressions under scrutiny.

I will then proceed to a theoretical discussion of the terminological debate opposing in particular *markers* to *connectives* and *discourse* to *pragmatic*, and will motivate the present choice for *discourse markers* (Section 3.1.2). Section 3.1.3 addresses the present state of crosslinguistic research on the DM category, focusing on the English-French pair, and the requirements of contrastive analysis, namely the concept of *tertium comparationis*. Lastly, an integrated view of the category of DMs will be developed in Section 3.1.4, introducing the functional corpus-based definition and grounding it in the ambition of cognitive pragmatics (Schmid 2012). This chapter will however not cover in detail the explicit instructions on how to apply the theoretical definition to corpus annotation, which requires a lengthy methodology and will rather be developed in Chapter 4 (Section 4.2.1.1).

¹⁴ In this respect, Traugott (1995), Sweetser (1988) and others stand against a more traditional view of language change whereby grammaticalization involves a loss of semantic meaning, also called bleaching or attrition (e.g. Lehmann 1985). Regarding DMs, it could be said that expressions such as *you know* lose propositional meaning (*to know*) but gain in return an interpersonal pragmatic function (see Brinton 1996: 54f.).

3.1.1 Core features

The study of discourse markers arose in the 1970s with the increasing interest for discourse, pragmatics and expressions functioning beyond the sentence. Schifffrin's (1987) seminal monograph, while focusing on particular expressions (e.g. *oh, well, now, I mean*¹⁵), is usually cited as the first attempt to categorize under one term elements coming from distinct grammatical classes and to define the general function of DMs as "sequentially dependent elements which bracket units of talk" (1987: 31). In her view, DMs contribute to coherence by punctuating units of talk and giving instructions on how to interpret the current utterance with respect to (i) previous or upcoming material, (ii) the speaker-hearer relationship and (iii) so-called "planes of talk", that is, generic functional domains such as content (ideational structure) or attitude (participation framework).

Schifffrin (1987) brings to the forefront of her definition the connectivity of DMs and acknowledges their multiple, non-linear scope over the linear string of connected units ("brackets look simultaneously forward and backward", 1987: 37), thus accounting for their possibility to occur at the end of utterances. Connectivity is indeed a recurrent feature of many DM definitions, as Schourup (1999: 230) points out in his exhaustive and well-structured state-of-the-art. He specifies that most authors agree on the necessary condition of connectivity for DMs (e.g. Fraser 1996; Hansen 1997), although the connection can be more or less strictly defined and can apply to different types of units (verbally expressed or not, single or multiple, contiguous or distant; see Hansen 2006 and Section 3.2.2). Thus, the primary role of DMs seems to be a connecting one, related to interpretation processes and building up the coherence between past, current and future utterances. Such a definition stands in sharp contrast with more outdated accounts of DMs which associate them with "unskilful speakers" and "powerlessness" (O'Donnell & Todd 1980: 67; Ragan 1983: 166), in a similar dichotomic treatment as was found in the field of fluency research (cf. Clark & Fox Tree 2002).

Schourup (1999) further mentions two consensual core features of DMs, namely syntactic optionality and non-truth-conditionality. Optionality refers to the possibility to virtually remove a DM from its host unit without altering the grammaticality of the utterance. One related area of controversy is whether this optionality is only grammatical, or if it also affects the semantic-pragmatic interpretation of the utterance. DMs are either seen as completely redundant (a causal relation would be perceived with or without a *because*) or pragmatically necessary to constrain the interpretation of discourse. Hansen (2006: 26) argues for the latter position, which she exemplifies by the following example:

- (1) Max forgot to go to the meeting. **In any case**, the committee decided to adjourn the meeting.

She argues that without the DM, a causal relation could be inferred when it should not (the meeting was adjourned because Max forgot about it), since "in any case" rather introduces a background information with an implicit concessive nuance: although Max forgot the meeting,

¹⁵ In this thesis, expressions in italics refer to the generic lexeme and not particular instances, which are rather signaled by double quotations marks as in "in any case" below. When the DM is in another language than English, it is always followed by a suggestive translation equivalent between single quotation marks, as in: *donc* 'so'. All translations (of examples, quotes and DMs) are mine.

he will not get in trouble because the meeting was adjourned anyway. The optionality of DMs should therefore be restricted to the grammatical or syntactic sense (i.e. removal does not alter the grammaticality of the utterance) while, at the pragmatic level, they are useful (sometimes necessary) cues for interpretation.

Turning to non-truth-conditionality, authors tend to agree on the distinction between DMs and “content” words, although they may use different semantic notions to do so. Fraser (1996) argues that DMs do not contribute to or affect the truth-conditions of sentences, which means that the utterance remains true whether or not the DM is present. This feature sets apart DMs from “content words, including manner adverbial uses of words like *sadly*, and from disjunctive forms which do affect truth-conditions, such as evidential and hearsay sentence adverbials” (Schourup 1999: 232). Non-truth-conditionality is sometimes taken as an equivalent to procedural meaning (e.g. Hansen 2006), in which case the stress is on their contribution to the interpretation (rather than absence thereof). Procedural expressions such as DMs do not refer to an entity or concept but rather give instructions that constrain the interpretation process. Procedurality is to be understood in opposition with conceptuality as defined in Relevance Theory (Blakemore 1987; Wilson & Sperber 1993; Wilson 2011): what is at stake is not absence vs. presence of encoded meaning, but rather the type of encoded meaning, either a concept (as for content words) or a procedure (typically for DMs).

Another set of terms commonly found in this matter is “propositional” vs. “non-propositional”, as in Jucker & Ziv (1998: 3), who say that DMs have “little or no propositional meaning”. Wilson (2011) points out some discrepancies between all these notions and how they only partially overlap, for instance in the case of deictics (pronouns such as *he*, temporal or spatial adverbs such as *now*), which are truth-conditional but where only some properties of the concept are encoded (*he* is a single male referent), thus requiring a procedure of contextual disambiguation to be fully identified. Without going into the detail of this semantic debate, we can safely say that DMs do have meaning, but not a stable one: they require contextual disambiguation and trigger interpretation processes which do not affect the propositional, truth-conditional or conceptual content of the host unit. In this sense, non-truth-conditionality is related to optionality. However, I would rather opt for the notion of procedurality (following, e.g., Bolly et al. 2015, in press) since it offers a positive characterization, as opposed to the negative “absence of effect on truth conditions” and “little or no propositional meaning”.

Another semantic aspect often mentioned in the definition of DMs is their multifunctionality, which applies at three levels: (i) the category includes items that perform a wide range of discourse functions, such as causal relation, reformulation or topic-shift; (ii) a single DM can be polysemous¹⁶ and perform different functions in different contexts; (iii) a particular instance of DM can express several simultaneous meanings in one context. These shades of multifunctionality for a single DM are illustrated in Examples (2)-(4).

¹⁶ While the multifunctionality of DMs is generally agreed upon, authors sometimes differ on their treatment of this phenomenon either as monosemy (one core meaning enriched by context into several interpretations) or polysemy (the expression possesses different yet related interpretations), a third option being homonymy whereby the different interpretations would not be related at all, as in *bass* (a type of fish or a type of music instrument). The contributions in Fischer’s (2006a) edited volume are representative of each approach. In this thesis, I will rather use the term multifunctionality to avoid such considerations, although my approach is closer to polysemy.

- (2) The meal was cold **so** I ate something else
- (3) I spent my Saturday gardening **so** I watered the plants, cleared the weeds and mowed the lawn
- (4) That was all for the nineteenth century **so** we now turn to the industrial era

In (2), “so” signals a relation of objective consequence between the fact that “the meal was cold” and that the (invented) speaker chose not to eat it. In (3), however, the meaning of consequence is no longer expressed: “so” rather expands on the first segment and specifies what the activity of gardening comprises. Examples (2) and (3) thus illustrate multifunctionality in the second sense developed above (different functions in different contexts). Example (4) illustrates the third sense of multifunctionality with an occurrence of “so” which expresses both a consequence (we can turn to the next period because we finished the previous one) and a structuring function similar to topic-shift whereby the DM indicates a new step in the (invented) lesson. Brinton (1996) and Schiffrin (1987) both mention this feature of DMs and further specify that, in the case of simultaneous meanings, it is not always possible nor relevant to identify which one is prevalent, if any. While multifunctionality does not seem to be a controversial aspect of DMs, as attested by the many studies and taxonomies available, it does however complexify the task of definition by blurring the commonality between all DMs under the profusion of their (semantic and other) differences.

Schourup (1999: 232) further identifies additional features which are “less consistently regarded as criterial for DM status” and rather constitute central tendencies of the most representative DMs. The first is “weak clause association” and is somewhat related to syntactic optionality: DMs are detached from the clause elements and the syntactic structure, in other words they are not bound by grammatical relations, unlike subjects or complements. Schourup (1999) relates this loose structure to prosodic independence, that is, the presence of a boundary (pause or intonation) separating the DM from the rest of the tone unit.

In addition, DMs tend to be utterance-initial, typically introducing their host unit. The initial position is also a corollary to the aspect of DM scope. In Degand et al.’s (2013) edited volume on the differences between discourse markers and modal particles, Cuenca (2013) and Valdmets (2013) specify that DMs tend to function at discourse level, inter-sententially and have scope over larger segments than modal particles which are sentence-internal modifiers. The inter-sentential function of DMs thus motivates their initial position, at the boundary between two utterances. Lastly, DMs are further said to be typical of spoken language, as a result of the informality and lack of planning in this modality (Brinton 1996: 33).

Counter examples on all these aspects are numerous. For instance, utterance-medial and utterance-final DMs have long been identified in the literature (e.g. *sort of*, *you know*; Mulder & Thompson 2008; Haselow 2011; Buysse 2014) and DMs largely occur in writing as well (although perhaps not with the same frequency nor the same forms and functions, e.g. Fox Tree 2014). These features are still useful in that they provide a portrait of typical DMs and reflect the complexity of the DM category, which should be considered as a continuum with “fuzzy

boundaries” (Cuenca 2013) rather than a clear-cut category similar to grammatical classes.¹⁷ These optional features, as well as the criterial ones presented above, are summarized in Table 3.1 and mapped onto a selection of widely spread definitions. No symbol means that the feature is not mentioned by the author, whereas the ✕ indicates that the author explicitly denies the feature; “procedural meaning” is taken as an approximate equivalent of non-truth-conditionality and absence of propositional content.

Table 3.1: Characteristic features of DMs in four definitions

	Schiffrin 1987	Brinton 1996	Fraser 1996	Hansen 2006
connectivity	✓		✓	✓
optionality	✓	✓	✓	✕
procedural meaning	✓	✓	✓	✓
multifunctionality	✓	✓	✓	✓
syntactic mobility	✓			
tendency of initiality	✓	✓	✓	
prosodic	✓	✓		✓
orality		✓		
grammatical diversity		✓	✓	✓

We see that most features are shared across at least three definitions, with the notable exceptions of syntactic mobility (i.e. not being restricted to particular slots) and orality, which are specific to one author. Only Hansen (2006) directly challenges one of these features, viz. the optionality of DMs, as discussed above. Such a table could be used as the basis for an integrated definition, including all features which are quasi-consensual (mentioned in at least three out of four definitions). DMs would then be connective, optional, procedural, multifunctional, utterance-initial, prosodically independent and grammatically diverse. Such a definition fits the following English examples of typical DMs: *and, so, but, actually, anyway, I mean, in fact, well*. Although this proposal would exclude some well-established – but less typical – members of the DM category (e.g. *though, you know*), it seems fairly inclusive and allows DMs to be distinguished from other conceptually close categories (such as modal particles).

In Bolly et al. (2015, in press), we proposed an empirical method to test the degree of fit between a number of these features and actual DM instances extracted from corpus data, thus investigating their usability for semi-automatic disambiguation purposes. Our multivariate analysis showed that prototypical DMs are loosely attached to the syntactic structure, which confirms the criterial role of syntactic features of DMs. However, we found no significant effect between DM status and the contiguous presence of pauses, which tends to qualify the importance of prosodic independence. One limitation of Bolly et al.’s (2015, in press) study, which is addressed in the present thesis, is the absence of functional considerations beyond the basic semantics of the marker. It will become apparent in the following sections that definitions

¹⁷ Grammatical classes are not always easily distinguished either, as for instance the adverb-adjective or determiner-pronoun distinctions (Croft 1991).

relying solely on formal parameters cannot fully account for the breadth of the category nor distinguish it from its pragmatic rivals, as already stated by Pons Bordería (2006) and Fischer (2016).

Despite the consensus that Table 3.1 seems to suggest, the picture is actually quite chaotic when it comes to establishing robust definitions for a bottom-up approach to the full DM category, as opposed to the bulk of case studies. Very few features of DMs are unanimous while some are rather scalar or optional, a flexibility which might be considered as slightly incompatible with the rigor and systematicity of empirical (corpus-based) research. Moreover, a number of other features, not discussed so far, are clearly controversial. I will now explore in more detail some reasons for the lack of consensus, starting with terminology and the theoretical choices that each term involves.

3.1.2 Beyond the terminological debate

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-Vandenberg 2006: 2). Beyond idiosyncratic preferences, it appears that terminology involves deeper theoretical disagreements on the definition and delineation of the DM category. I will abide by the tradition and briefly discuss a selection of frequent labels: discourse markers (e.g. Schiffrin 1987); pragmatic markers (e.g. Brinton 1996); discourse particles (e.g. Fischer 2000); connectives (e.g. Fraser 1996).¹⁸ I will develop the benefits and drawbacks of each proposal, suggest some explanations for their differences and motivate my present choice.

I will start with the dichotomy between “discourse markers” (DMs) and “pragmatic markers” (PMs), following Hansen (2006: 28) who 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 “any procedural element contributing to the interpretation of context by other means than semantic decoding” (Crible in press) 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-Vandenberg 2011: 10). The main issue with this broad category concerns the heterogeneity of its members, since they

¹⁸ Less common terms include “discourse operators” (Redeker 1991), “cue phrases” (Knott & Dale 1994), “discourse-relational devices” (proposed by the ISCH 1312 COST Action “TextLink: Structuring Discourse in Multilingual Europe”, e.g. Šliogerienė et al. (2016)) and variants thereof.

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. As a consequence, such a broad class of expressions would hardly be analyzable or interpretable coherently. Supporters of the PM term argue that “discourse marker” is too restricted to connective or structuring functions, as opposed to “pragmatic marker” which also includes elements which “mark illocutionary force or have an interactional function, for instance taking the turn or yielding it” (Aijmer & Simon-Vandenberghe 2011: 9). Overall, PMs are quite consensually seen as including DMs along with a great variety of other expressions, which motivates my choice of the latter for reasons of coherence in the category and feasibility of the analysis, in addition to the “majority rule” argument (DMs being more frequently used than PMs).

Another term commonly found is that of “particles”, either “pragmatic particles” (Östman 1995) or “discourse particles” (Fischer 2000; Aijmer 2002). I will focus on Fischer’s (2000, 2006b) approach to the opposition between “discourse particles” and “discourse markers”. The former seems to convey distinctive formal properties which are useful to draw the boundaries with other classes, whereas the latter is more functional, although formally vague and potentially too inclusive. Müller (2005) suggests that the term “particle” stems from Germanic linguistics where they form a fully fledged grammatical class, which explains the types of formal restrictions mentioned by Fischer (2006b: 4): “The term *discourse particle* suggests a focus on small, uninflected words that are only loosely integrated into the sentence structure, if at all. The term *particle* is used in contrast to clitics, full words, and bound morphemes”. Indeed, if we look at the German expressions investigated in Fischer (2000), we can see that they roughly correspond to (or originate from) short interjections, such as *ja* ‘yes’, *oh* ‘oh’ or *äh/ähm* ‘uhm’, in specific formal and functional configurations. Fischer (2006b) admits that, while helpful to narrow down the range of expressions under scrutiny, such formal criteria might be too restrictive and exclude functionally similar elements, especially when working crosslinguistically.

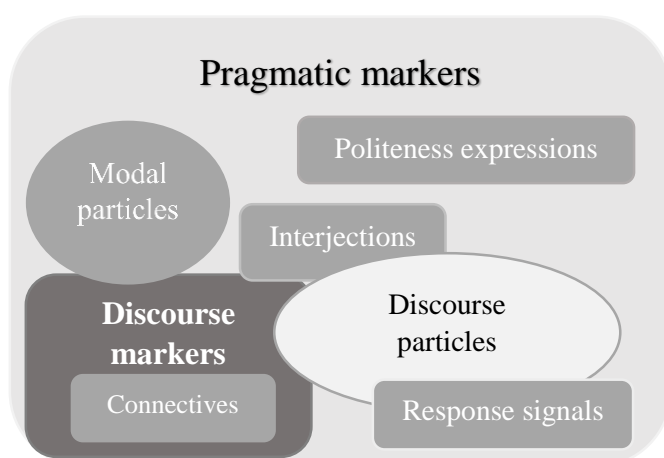
The Germanic influence behind the term “particle” is also explained by the category of modal particles (MPs), which are recognized as a specific word class in these languages (e.g. Diewald 2006, 2013). Interestingly, the function of discourse particles as defined by Fischer (“to mark an utterance as non-initial”, 2000: 14) is the same as that of modal particles (“the MP marks the utterance which contains it as non-initial”, Diewald 2006: 407), which points to yet another source of confusion between these categories. It could also be said that non-initiality is similar (if not equivalent) to the connectivity criterion for DMs mentioned in Section 3.1.1: the DM relates the unit it introduces to previous context (verbally expressed or not). In sum, discourse markers, discourse particles and modal particles all seem to be functionally close and would potentially overlap without the addition of formal criteria.

Cuenca (2013: 195) further indicates that modal particles and discourse markers can be (functionally) distinguished but that there exists some middle ground between two extremes where “a number of units [share] some of the defining characteristics of both DMs and MPs”, which she coins “pragmatic connectives”, such as *well*. With this last term, Cuenca (2013) refines the continuum between modality and discourse, which I have summarized in Crible (in press): “connectives and MPs are formally similar (syntactically integrated, occupying a

prototypical position, easily identifiable grammatical classes), MPs and DMs are functionally close (having an attitudinal and metadiscursive function, respectively, with scope over one unit), and connectives and DMs share a discourse-connecting or structural function”. Connectives, in turn, also benefit from a long trend of studies focusing on their relational and cohesive function. Some variants of this term include “pragmatic connectives” (Van Dijk 1979; Fraser 1988) and “discourse connectives” (e.g. Unger 1996). However, they seem to be restricted to conjunctions and other relational markers (*and*, *so*, *although*, *however*, etc.) and would therefore be considered as a subcategory within a broader view of DMs, alongside other non-connective expressions such as *you know* or *like*, which are otherwise considered as typical of the category.

This complex situation is summarized in Figure 3.1, which represents the theoretical position adopted here regarding this terminological debate. We can see that DMs are a subtype of PMs along with interjections or modal particles. Discourse particles are here considered to cover different subtypes such as interjections or response signals, following Fischer (2000: 15). The fragmented view of pragmatic markers that this figure seems to suggest should actually be nuanced, in light of the research discussed above which showed formal and functional overlap between these subtypes in some borderline contexts of use. In addition, this model is not meant to be exhaustive, and further subtypes of PMs could be identified.

Figure 3.1: Present terminological model of discourse markers and other pragmatic expressions



It may have occurred to the reader that some of these subtypes are specific to spoken language (e.g. response signals, interjections) while others are more typical of writing (e.g. connectives such as *on the other hand* or *therefore*). I have discussed at length some of the differences between these two modalities (cf. Section 2.1.1) and called for a more gradual approach to the apparent binary divide. This applies to the present issue as well, given that some DMs or other PMs might be more or less associated with different kinds of writing and speaking activities. For instance, you would find politeness expressions in corresponding letters, expressive language such as interjections or modal particles in text messages and elaborate connectives in formal (e.g. political) speeches. While the boundaries between subtypes of PMs and even between speech and writing might be porous and overlapping, it remains that the high frequency of particular markers in one data type or the other might influence the terminological choice, as

appears to be the case for DMs (mostly studied in speech) vs. connectives (mostly in writing). A related (but not completely equivalent) divide is between frameworks investigating *markers* and those investigating *relations*, where we find a similar association with speech and writing, respectively (I will address this issue in more detail in Section 3.2.1).

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, so much so that 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. 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.

3.1.3 From monolingual case studies to crosslinguistic categories

After having established more or less consensual core features and situated DMs within other pragmatic categories (at least on a terminological level), I will now proceed to a review of contrastive DM research on the one hand, and onomasiological DM research on the other, which will in turn pave the way for the definition of the DM category in English and French adopted in the present study. 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, especially not 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 discursifs*, 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. In English, the majority of case studies on DMs is interested in speech-specific expressions, as illustrated by Erman’s (1987) conversational analysis of *you know*, *you see* and *I mean*, for which she identifies a number of positional, phonological, syntactic, functional and perceptual features. She concludes that these “pragmatic expressions” (as she calls them) “fulfil a variety of functions in the spoken language, but also that the language user cannot do without them, although s/he may not always be aware of this” (1987: 217). The same DMs are the focus of Fox Tree & Schrock (2002), who investigate the impact of “core meaning” on the multifunctionality and surface behavior of *you know* and *I mean*. The authors relate their functions and uses to the “naturalistic, unplanned, unrehearsed, collaborative nature of spontaneous talk” (2002: 745) and make recommendations on when to avoid using them (e.g.

restrain from saying *you know* when talking to higher-status participants or to a large group of participants).

Other popular DMs in English research are *okay* (Condon 1986, 2001; Beach 1993; Gaines 2011), *well* (Schiffrin 1987; Jucker 1993; Schourup 2001; Lam 2006), *now* (Schiffrin 1987; Aijmer 1988, 2002; Schourup 2011), *like* (Underhill 1988; Romaine & Lange 1991; Andersen 2000; Fuller 2003; Zufferey & Popescu-Belis 2004) and variants of so-called “general extenders” such as *and stuff* (Cheshire 2007; Buysse 2014) or *and that sort of thing* (Aijmer 2002). Some of these studies are descriptive and strive to delineate the functional spectrum of particular DMs, while others adopt a more sociolinguistic approach to their variation. For instance, Andersen (1997) studied the use of *like* in London adolescents and found a correlation with socioeconomic class (more frequent in higher social groups), a result which might contradict the experimental study by Hesson & Shellgren (2015), who rather showed that the same DM has a negative effect on ratings of friendliness and intelligence. Torgersen et al. (2011) pursued a similar endeavour as Andersen (1997) in their corpus-based study of nine “pragmatic markers” (invariant tags such as *innit*; *you know* and longer variants; *you get me*) for which they identified a number of effects pointing to the role of dialect contact in language change.

Apart from these conversational DMs, more register- and modality-neutral expressions have also been investigated, typically conjunctions such as *and* (Dorgeloh 2004; Blakemore & Carston 2005) or *so* (Aijmer 2002; Bolden 2007, 2008; Buysse 2012). Schiffrin (1987) groups *and*, *but* and *or* in one section of her monograph, *so* and *because* in another, yet she acknowledges that all five of them “are used in discourse in ways which reflect their linguistic properties” (1987: 227), as opposed to other DMs such as *well* or *oh*, which are not influenced by their original grammatical nature.

The situation is rather similar in French, where studies on both written and spoken DMs adopt a wide array of approaches:

- theoretical accounts (Hansen 1997 on *alors* ‘well’ and *donc* ‘so’; de Saussure & Sthioul 2002 on *et* ‘and’);
- diachronic studies (Hansen 2005 on *enfin* ‘at last/I mean’; Degand & Fagard 2011 on *alors* ‘so/well’);
- corpus-based studies (Chanet 2001 on *quoi* ‘you know’; Bolly 2009 on *tu vois* ‘you see’; Bolly & Degand 2009 on *donc* ‘so’);
- sociolinguistic investigations (Beeching 2007 on *bon* ‘well’, *c’est-à-dire* ‘that is to say’, *enfin* ‘I mean’, *hein* ‘right’, *quand même* ‘still’, *quoi* ‘you know’ and *si vous voulez* ‘if you will’).

A specificity of French studies as compared to English might be the more frequent inclusion of prosodic features in the analyses of DMs, as in Bertrand & Chanet (2005) on the relation between the pragmatics and prosody of *enfin* ‘I mean’, Anzai (2009) on the morphosyntactic and intonational properties of *tu vois* ‘you see’, or Avanzi (2009), who found that the erasure of intonational boundaries between parenthetical verbs (e.g. *je crois* ‘I believe’) and the rest of

the utterance is a sign of DM use – a study which echoes Dehé & Wichmann's (2010) prosodic analysis of English epistemic parentheticals (*I think, I believe*).

Some French DMs are studied crosslinguistically, although not necessarily with English, such as *alors* 'so/well', which is contrasted to Italian *allora* in Bazzanella et al. (2007), *en fait* 'in fact', contrasted to Dutch *eigenlijk* in Mortier & Degand (2009) or *donc* 'so', contrasted to Norwegian *altså, så* and *derfor* in Nome & Haff (2011). As for the French-English pair, 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 (in press) on *just* vs. French *juste*. Given the paradigmatic scope of the present thesis, these references will not be reviewed in any more detail, since their results are not generalizable to the full category of DMs and are therefore only distantly relevant to the elaboration of hypotheses.

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.

The next group of references takes an even more inclusive perspective on the DM category by proposing theoretical definitions or lexicons of a larger number of markers – although the resulting categories would still be considered as subsets of the whole DM category in the present paradigmatic definition. It remains that these studies aim at covering full (monolingual) categories in their own terms. In French, Auchlin (1981) suggests an inventory of *marqueurs de structuration de la conversation* ('markers of conversation structuring') including, among others, *au fait* 'by the way', *ouais mais* 'yeah but', *quoi* 'you know', *alors* 'well' or *bon ben* 'well', for which he provides a classification in terms of levels of text structure (*niveaux de textualisation*) as well as some positional information (beginning of subclauses, beginning of turns or exchanges). Another seminal reference is Vincent (1983) and her notion of *ponctuants*, which she defines as "verbal signs that, beyond the sentence, contribute to establishing coherence between utterances and cohesion between speakers" (1983: 43, my translation). However, in her study, Vincent only reports data for ten forms which she found to meet this definition in her corpus. Another well-known framework is that of Roulet et al. (1985) and their *connecteurs argumentatifs* which are divided into three classes according to the type of segment they introduce (either an argument, a conclusion or a counter-argument). The last

general reference in French is LEXCONN, a lexicon of connectives elaborated by Roze et al. (2012) where 328 markers have been identified and classified according to their syntactic class and associated discourse function(s), in a natural language processing perspective.

English DMs benefit from a similar interest for connectives as a whole (in writing). Rouchota (1996), for instance, compared two theoretical approaches, namely coherence theory and Relevance theory, in terms of their adequacy to render the behavior of connectives (understood in a broad sense, including e.g. *too*, *indeed*, *even*, etc.). The most widespread theoretical and/or empirical frameworks in English research have largely outgrown the limits of the language (and modality) they were originally designed for, but present one major difference with the references reviewed so far: they target taxonomies of discourse relations (either explicitly marked or not) instead of lexicons of DMs or connectives. Among them, the Penn Discourse TreeBank 2.0 (Prasad et al. 2008) adopts a more lexically-based approach to discourse relations and provides a mapping of connectives and their associated relations as extracted from their corpus of Wall Street Journal.¹⁹ Overall, onomasiological, inclusive studies in the English literature target a slightly different object of study (the relation and not its marker), which is why I will not expand here any further on these works, and leave their detailed review in Section 3.2.1, where the content of the taxonomies will be discussed and compared to speech-based and other models.

Coming back to contrastive approaches, 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. (2014) who used translation spotting techniques to build a multilingual lexicon of DMs from the Europarl²⁰ corpus of parliamentary debates (i.e. cleaned transcriptions of written-to-be-spoken data) in English, Portuguese, French, German and Italian. Using the same parallel corpus, Zufferey & Degand (in press) carried out a multilingual annotation experiment in English, French, German, Dutch and Italian, disambiguating the discourse relations expressed by connectives as varied as *after*, *and*, *despite*, *meanwhile* or *thus* and their translations. While the transcriptions in the Europarl corpus are closer to a written style than to natural speech, other crosslinguistic works have been working with spoken data as well. Kunz & Lapshinova-Koltunski (2015) 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. She develops an insightful taxonomy of DM functions in four groups which will prove crucial to the present approach and which will be developed in Section 3.2.1.

To sum up, so far, I have briefly reviewed a selection of works illustrating monolingual case studies in English and French focusing on speech-specific expressions (e.g. Erman 1987), contrastive case studies (including for the English-French pair, e.g. Lewis 2006a), contrastive

¹⁹ The other two English frameworks are Rhetorical Structure Theory (Mann & Thompson 1988) and Segmented Discourse Representation Theory (Asher & Lascarides 2003), see Section 3.2.1.

²⁰ Available at <http://www.statmt.org/europarl/>.

studies of larger groups of DMs (e.g. Dupont 2015), proposals of monolingual lexicons and categories (e.g. Roze et al. 2012) and multilingual large-scale corpus-based research (e.g. González 2005). It appears that 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 insofar as 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 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 (i) they depend on the particular languages targeted in the study and (ii) they are relative to the agenda and objectives of the study. In the present case, a comparable crosslinguistic definition of DMs would reflect the researcher's interest and focus on individual aspects such as discourse structure, attitude or intersubjectivity, thus potentially rejecting forms and uses which attend to one aspect and not the other. For instance, it is likely that forms as different as *although* (typically used for discourse structure) and *you know* (typically used for intersubjectivity) would not be grouped in the same category, unless the *tertium comparationis* is explicitly designed to encompass these different roles.

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 operationalization phase will be developed in Section 4.2.1.1). 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 large 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). This gap in the field motivates the present study, in addition to the rationale described in the previous chapters (Sections 1.2 and 2.3.4).

3.1.4 Towards a corpus-based definition

After the review of the major trends, differences and limitations of current DM research, let us now turn to the precise definition of the DM category proposed here and specify its theoretical and methodological background. I will first put forward the same argument as in the previous chapter on (dis)fluency and call for, on the one hand, a combination of a broad abstract definition with, on the other, its matching operational annotation model. This endeavor is thought to bridge the gap between two major kinds of definitions available in the literature, “either a theoretical, usually quite abstract account of variables that might affect the behavior of DMs, or more in-depth case studies that specify a method but only for a certain type of elements or data. The first type is rarely operationalized, while the second is hardly reproducible on a large scale” (Crible in press). The ambition to reconcile these two extremes is also in line with recommendations from the field of cognitive pragmatics which advocates the combination (perhaps even the prevalence) of “psychologically plausible [...] ‘realistic’ models of the construal of meaning-in-context” with (or over) the more practical criteria of “parsimoniousness, elegance and descriptive and explanatory power of a theory” (Schmid 2012: 4-5).

Indeed, a definition which is inclusive enough to account for the diversity of DMs in speech, and still provides “the necessary criteria to isolate the specificities of DMs against other pragmatic categories” (Crible in press) is not easy to find. Some definitions target mostly formal features in order to maximize the similarity between members of the category. For instance, Diewald (2006: 408) considers that DMs “relate non-propositional discourse elements which are not textually expressed”, are “syntactically non-integrated” and have “no constituent value”. She further specifies that “DMs are prosodically, syntactically, and semantically independent” (Diewald 2013: 23). Other definitions rather target the role or meaning of the category, e.g. “sequentially dependent elements which bracket units of talk” (Schiffrin 1987: 31); “fulfilling structuring functions with respect to local and global content and structure of discourse” (Fischer 2000: 20); “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” (Hansen 2006: 25); “any type of linguistic expression whose primary function lies at the discourse level, i.e. relating their host utterance to the discourse situation” (Degand 2014: 151).

There is some obvious overlap within each group of definitions (formal or functional), but rarely between them, which motivates the need for a new proposal encompassing all relevant information in a compact phrasing. In addition, existing definitions are only rarely – if

ever – designed in the perspective of large-scale and bottom-up identification of DMs in corpus annotation, except in written data (with the restrictions in scope mentioned in the previous section). My proposal is the following:

DMs are a **grammatically heterogeneous, syntactically optional, multifunctional type of pragmatic markers**. Their specificity is to function on a metadiscursive²¹ 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, expliciting the **structural sequencing** of discourse segments, expressing the speaker's **meta-comment** on their phrasing, or contributing to **the speaker-hearer relationship**.

We see that this definition is (i) relative to other pragmatic categories, (ii) formal-functional, with a primary pragmatic role constrained by syntactic filters (the four types of use will be developed in Section 3.2.2) and (iii) very much indebted to all the proposals discussed so far. It is also broader than most frameworks specifically designed for corpus annotation, where connectivity is usually understood in a strict sense (i.e. scope over two abstract objects), as opposed to this more encompassing definition including non-connective functions (meta-comments, interpersonal collaboration). The inclusion of structural functions (e.g. topic-shift, turn-taking) is also rather innovative compared to writing-based frameworks where they are generally considered as a different level of analysis (e.g. Sanders et al. 1992; see Section 3.2.1). In sum, this definition is thought to overcome the divide between formal and functional features and assumes a broader view of coherence and connectivity.

The combination of syntactic and pragmatic features is the result of the confrontation 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 in press). Similarly to the flunceme typology, I will leave the operationalization of the definition and precise criteria to the description of the methodology (Section 4.2.2.1). In the following sections, I will turn from definitions to annotation models, and develop in more detail the functional spectrum of DMs central to their definition.

3.2 The functions of discourse markers in corpora

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 multifunctionality of DMs further require to dig into detailed taxonomies, in order to better grasp what the general definition presented in the previous section 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 large

²¹ “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.

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 3.2.1) and moving on to more inclusive views of the functions and scope of DMs (Section 3.2.2).

3.2.1 Writing-based models of discourse relations

In this section, I take up the three major English frameworks mentioned in Section 3.1.3 (cf. Footnote 19), namely the Penn Discourse TreeBank 2.0 (henceforth PDTB, Prasad et al. 2008), Rhetorical Structure Theory (henceforth RST, Mann & Thompson 1988) and Segmented Discourse Representation Theory (henceforth SDRT, Asher & Lascarides 2003). All of these taxonomies are writing-based (i.e. originally designed for writing, although not restricted to writing) and include different discourse relations such as cause, concession or condition in more or less fine-grained distinctions of meaning. They have all been adapted to several languages (e.g. Oza et al. 2009 or Zeyrek et al. 2013 for Hindi and Turkish versions of the PDTB, respectively), and recent endeavors have started to transfer these taxonomies to spoken corpora (e.g. Tonelli et al. 2010 for spoken Italian). A last general feature is their applied perspective: all three frameworks aim at high replicability and automatization, in order to be used for summarization or full-text segmentation purposes.

Behind these general similarities in their research agenda, some theoretical and practical differences emerge from a careful review of the three theories. The first divide, as noted by Benamara & Taboada (2015), opposes the PDTB and its lexically-based approach (i.e. start from connectives and annotate the related segments) to both RST and SDRT which rather aim at full-text segmentation and representation: “RST proposes a tree-based representation, with relations between adjacent segments, and emphasizes a differential status for discourse components (the nucleus vs. satellite distinction). [...] Captured in a graph-based representation, with long-distance attachments, SDRT proposes relations between abstract objects using a relatively small set of relations” (Benamara & Taboada 2015: online). The differences between the PDTB, RST and SDRT include:

- coverage of annotated units: connective-based in the PDTB, full-text in RST and SDRT;
- distance between annotated units: adjacent in the PDTB and RST, long-distance in SDRT;
- whether or not a specific annotation format is integrated in the model: none in the PDTB, trees in RST, graphs in SDRT;
- number of discourse relations in the taxonomy: from 20 to 78 in different versions of RST, 43 in the PDTB, only 12 in SDRT (cf. Sanders et al. forthc.).

Beyond the number of discourse relations, a more important difference concerns the types of relations included as well as the granularity of the taxonomy. The three frameworks disagree on the structure of the taxonomy (hierarchical or not) and the particular labels they include. A single relation can be labeled differently across frameworks (e.g. *disjunction* in RST and *alternation* in SDRT) and a single label can cover different relations (e.g. PDTB and RST

distinguish *contrast* from *concession* whereas SDRT subsumes the two relations under *contrast*). I will hereafter focus on the PDTB model, since it was a crucial influence to my approach. Table 3.2 shows its structure and most frequent connectives by relation.

Table 3.2: Discourse relations in the PDTB 2.0 with their typical connective (Prasad et al. 2008)

Level 1	Level 2	Level 3	Typical connective
Temporal	Synchronous		<i>as</i>
	Asynchronous	precedence succession	<i>before</i> <i>after</i>
Contingency	Cause	reason	<i>because</i>
		result	<i>so</i>
	Pragmatic cause Condition	justification	<i>indeed</i>
		hypothetical	<i>if</i>
		general	<i>if</i>
		factual past	<i>if</i>
		factual present	<i>if</i>
		unreal past	<i>if</i>
		unreal present	<i>if</i>
		relevance	<i>if</i>
		implicit assertion	<i>if</i>
Comparison	Contrast	opposition	<i>but</i>
		juxtaposition	<i>but</i>
	Pragmatic contrast Concession		<i>but</i>
		expectation	<i>although</i>
	Pragmatic concession	contra-expectation	<i>but</i>
			<i>nevertheless</i>
Expansion	Conjunction		<i>and</i>
	Instantiation		<i>for example</i>
	Restatement	specification	<i>in fact</i>
		equivalence	<i>in other words</i>
		generalization	<i>in short</i>
	Alternative	conjunctive	<i>or</i>
		disjunctive	<i>alternatively</i>
	Exception	chosen alternative list	<i>instead</i> <i>except</i>

In their revised version of the PDTB model, Zufferey & Degand (in press) 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 3.3, was used as reference for the present design of relational functions of DMs (presented in Section 3.2.2 and detailed in Section 4.2.1.2).

Table 3.3: Revised PDTB from Zufferey & Degand (in press)

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

Another interesting feature of the (original and revised) PDTB, which is also present in RST but not in SDRT, 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 RST, Mann & Thompson 1988), “content” vs. “epistemic” vs. “speech-act” (Sweetser 1990), “ideational” vs. “rhetorical” (Redeker 1991) 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 (5)-(7) borrowed from Sweetser (1990: 77-78).

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

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

(7) **Since** you are so smart, when was George Washington born?

In (5), “because” relates the fact that “John came back” to its external/objective/content explanation (“he loved her”). In (6), 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 (7), “since” introduces a justification (“you’re so smart”, i.e. you should know this) for the upcoming speech-act, here a question (“when was George Washington born”), 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 (terms and taxonomy will be introduced in the next section).

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. 2012 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 specificity of spoken connectives (i.e. more polysemous, 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 (in the strict sense). 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. Degand & Simon-Vandenberg (2011) talk of a scale of relationality, with non-relational DMs such as epistemic parentheticals (e.g. *I think*) at one extreme, and purely relational DMs such as *because* at the other.²² The authors argue that “this [relational] function may, but need not, be the primary one of discourse markers” (Degand & Simon-Vandenberg 2011: 288). Examples (8) and (9) illustrate two more or less relational uses of the same DM, viz. French *alors* ‘well’.

(8) 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

²² Epistemic parentheticals are not included in the present definition of the DM category (they are rather another type of pragmatic markers), yet the argument remains.

*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)*

- (9) 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** ‘well’ uh yes France will be ready for the upcoming century (FR-poli-02)*

The two “alors” in Example (8) 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. 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 (8) 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 (9) 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 range of DMs in the category and in the annotation procedure (see Section 4.2.1.4).

The inclusion or exclusion of non-relational uses of DMs is partly due to (i) whether the taxonomy is originally designed for speech or writing and (ii) whether the taxonomy targets discourse relations or discourse markers. Regarding the former (i), while both relational and non-relational uses are possible and frequent in spoken data, non-relational uses are, for the most part, absent from written texts, since writers aim at maximizing connectivity and cohesion between different segments, in order to ease the reading process in the absence of physical co-presence. Studies on written corpora would be required to assert that non-relational DMs are restricted to speech, so much so that we cannot fully explain the divide by this reason alone. The second explanation (ii), however, seems much more relevant, even beyond the issue of relationality. Whether a taxonomy targets discourse relations or discourse markers implies differences in research interests, scopes and challenges. Discourse *relations* are often studied in the perspective of automatic classification, identification of implicit relations and machine translation (e.g. Meyer & Popescu-Belis 2012), usually on written data – although recent endeavours are moving towards inter-operability with spoken corpora as well. Discourse *markers* involve the issues of delimiting the category, describing their syntactic position and coming to terms with their multifunctionality for sense disambiguation. DMs are often studied in spoken corpora given their special attraction to this modality (Brinton 1996). The issue of relationality is not solely a matter of definition (what are connectives or DMs) but mostly one of research agenda. Whether or not a taxonomy includes non-relational meanings of DMs should therefore not be viewed as a sign of greater or lesser exhaustivity and validity but as one of coherence with a particular research tradition, each with its own merits and limitations.

Finally, another feature of writing-based models of discourse relations, somewhat related to the exclusion of non-relational DMs, is the divide that they see between discourse or

coherence relations on the one hand, and topic or structural functions on the other. Sanders et al. (1992: 2) in particular claim that these two sets of relations constitute different approaches or levels of coherence, one being concerned with relations between segments, the other with the content of segments and topic continuity. The authors further exclude temporal relations from their basic types of relations (additive vs. causal) on the same grounds, i.e. that they mostly rely on the content of the segments, and should therefore be subsumed under additive relations instead of forming their own category, as opposed to the PDTB, for instance, where temporal relations constitute a separate class. Overall, such exclusions based on semantic categorizations (segments content vs. discourse coherence) seem to vary greatly from framework to framework and are very much intertwined with their focus on either relations or markers. A marker-based approach, such as the present one, would tend to include the full range of functions of a single expression, provided each use meets the function-based definition, beyond restrictions of semantic type (temporal or topic relations) and scope (non-relational meanings).

In conclusion, a number of differences and limitations have emerged from this review of writing-based models of discourse relations, which suggests to opt for approaches which are more compatible with the present endeavor, namely speech-based, marker-based, bottom-up and inclusive models of DM functions. Such proposals will be discussed in the next section.

3.2.2 The many scopes of DM functions

The previous review of relational models of DMs has revealed some limitations in their coverage of the full spectrum of DM functions, namely the existence of non-relational uses and the many levels that they affect (micro-structure of adjacent discourse segments vs. macro-structure of topics). 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 a higher level 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. In Crible & Cuenca (under review), we discuss a complex case of *so* which illustrates such multiple interpretations. I report it here:

- (10) <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 Liverpoolian 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)

We argue that "so" introducing the segment "I do find the accent is a bit harsh" can either be interpreted as (i) connecting it to the immediate co-text ("I'm a bit old-fashioned in that way"), thus signaling a relation of objective consequence; (ii) acting as a conclusion to the anecdote about the nephew; (iii) referring back to the previous evaluative segment ("I'm not saying I disapprove of it but I think it's a lazy speech"); or iv) introducing an answer to the interviewer's (<BB1>) original question. In examples like (10) 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 multifunctionality of DMs: "it thus appears that spoken DMs not only tend to take scope over large and distant segments (as in writing), but can also combine local and global scope simultaneously, making the annotation process quite complex" (Crible & Cuenca under review).

Lenk (1998: 208) terms this variability as "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. Moreover, according to Lenk, DMs not only vary in scope but also in directionality, with typically retrospective (such as *anyway* or *however*) and prospective DMs (e.g. *actually*, *incidentally*, *what else*). This duality clearly echoes the difference between backward- and forward-looking disfluencies developed in Section 2.2.1.

Such a perspective on the variation of DM functions also echoes and extends the semantic-pragmatic (also objective-subjective or content-epistemic) divide discussed in the previous section. If we admit that DMs can and do function at these two levels of discourse, we can transfer the same cognitive hypotheses regarding the differences between semantic vs. pragmatic relations to local vs. global scope. These hypotheses stem from experimental literature showing differences in processing load, in particular a disadvantage or higher load for subjective relations (Traxler et al. 1997; Canestrelli et al. 2013). Zufferey (2010) also provides

convincing evidence with her longitudinal corpus study on the acquisition of objective and subjective causal relations in French. She showed in particular that epistemic relations are acquired later by children and that the typically subjective connective *puisque* ‘since’ is not acquired until children are four years old.

Whether the opposition between local and global scope triggers similar differences in cognitive processing as the semantic-pragmatic divide remains to be experimentally tested, and it is certainly one goal of this thesis to provide tentative corpus-based evidence for it (see the hypotheses in Section 3.4). Were such differences to be observed, they could be explained by an effect of working memory: long-distance relations and DMs working at a higher-level of discourse organization rely on an active and shared representation of discourse, which needs to be stored and retrieved efficiently through minimal processing effort. As Meyer et al. (2007: xii) put it:

Producing and understanding utterances necessarily involves holding various types of information in working memory. For instance, speakers must remember whom they are talking to; keep track of what they have said already; and, according to many theories, temporarily buffer utterance fragments before producing them. Similarly, listeners must remember what they have been told already, and keep utterance fragments in a buffer, minimally until semantic and syntactic processing is possible.

One approach is to consider that our resources of memory are limited and can therefore be exceeded by on-line demands, thus leading to a processing slowdown or even complete failure to recollect the relevant information. Martin & Slevc (2014: 441) explicitly relate memory to fluency with experiments which showed evidence for “a role for working memory in discourse fluency, and for a role of simple short-term memory in discourse coherence, but the relative contributions of working memory and short-term memory have not been directly assessed”. They further specify that discourse coherence, i.e. “how well utterances are linked with what has come before” (2014: 440), is an “appropriate measure of fluency at the message level”. This notion of working memory will not be developed any further given that, in its current state, it might still require further research and empirical testing. Nevertheless, working memory appears to be involved in the link between long-distance relations, discourse coherence and fluency, and will thus prove relevant to the present study (see especially Sections 6.4 and 7.3.2.2). Such literature from neuroscience points out the specific limitations of our speech processing abilities for coherence and fluency given the simultaneity of the production and comprehension tasks and the typically low amount of planning available.

Coming back to the functional variation of DMs, multiple levels of discourse coherence are particularly relevant in the case of co-occurring DMs, which we argue to be a strong tendency of unplanned spoken discourse in Crible & Cuenca (under review). 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). I will however focus on Cuenca & Marín’s (2009) tripartite model since it is scalar and involves considerations of functions and scope. According to the authors, 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).²³ I report here one example for each type of co-occurrence (borrowed from Crible & Cuenca under review):

- (11) so it's actually a proper increasing function (2.833) ok (1.730) **so for example if** you wanted to supposing you're looking at sine x (1.500) and you wanted to define (0.920) an inverse to sine x (1.680) I mean how are you going to do that (EN-clas-04)
- (12) he's the guy who is supposed to have left and he had my papers **and so** that was the problem over the party (EN-conv-06)
- (13) and she's feeding a baby (0.220) so uhm (0.333) **and then** yes of course this is a some sort of love scene going on (EN-clas-05)

In (11), two distinct functions across three DMs can be identified and do not take scope over the same segments: “so” and “for example” connect the previous segment with the example that they introduce, in a relation of exemplification or instantiation, while “if” relates the interrupted segment “if you wanted to” with the question “how are you going to do that”, in a (pragmatic) conditional relation. In other words, “so for example” constitute a case of “addition” of two DMs (similar function and scope) and are retrospective, while “if” is juxtaposed and prospective, in Lenk’s (1998) terms. In Example (12), the additive meaning of “and” is combined with the consecutive sense of “so”, thus marking both continuity and conclusion to the previous context. Lastly in (13), the meaning of “and then” cannot be decomposed but rather functions as one unit indicating descriptive continuity. For the present purposes, Example (11) is particularly relevant to illustrate how DMs are not only multifunctional and polysemous but also vary in scope, sometimes in complex combinations. 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.

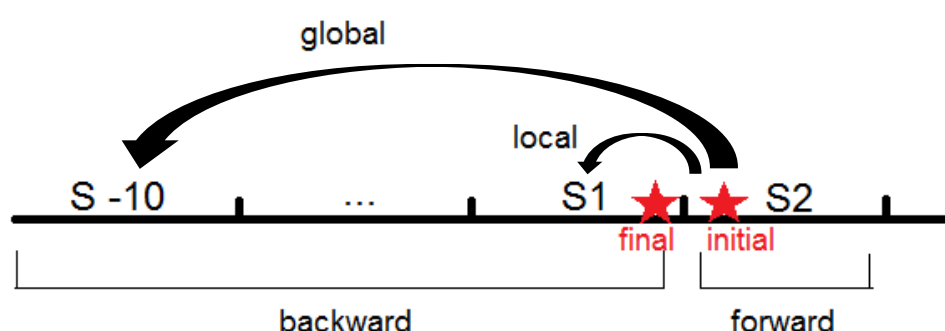
Besides the tendency of spoken discourse to produce clusters of DMs, another explanation for the use of combined DMs is the ability of DMs to occur in final position of the utterance, especially in spoken language, which can in turn be related to the temporality of this modality (cf. Section 2.1.1). 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 (14).

²³ 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).

- (14) 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 (“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), so much so that there is (again) no one-to-one mapping between DM function, position, relationality and scope. This complex picture is represented in Figure 3.2.

Figure 3.2: Illustration of the functional flexibility of DMs



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) models presented in the previous section. 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 a specificity of 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.

The present approach to the functions of DMs is strongly rooted in this last reference, 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 rather exhaustive, this proposal leaves some room for improvement, in particular regarding the operational definition of the functions and their classification in coherent categories. Instead of going into the detail of the final model and its revisions from González's (2005) original taxonomy (which will be laid out in Section 4.2.1.2), I will only announce the four domains as they are presently defined, before I conclude on this section:

- *ideational* domain: relations between real-world events;
- *rhetorical* domain: relations between epistemic and speech-act events and metadiscursive functions;
- *sequential* domain: structuring of discourse segments;
- *interpersonal* domain: interactive management of the speaker-hearer relationship.

To conclude, 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 potential and multifunctionality. González (2005), among others, offers such a proposal, which was thus adopted for the present research (with a number of adaptations detailed in the description of the methodology). We have seen how relational connectives are but one subtype of DMs, and that any accurate and exhaustive account of DMs in spoken language should strive towards inclusiveness and accommodate a number of (non-relational, global-coherence) functions, which are otherwise absent from writing-based models of discourse relations. This endeavor to work with models specifically designed for the type of data and modality at stake is in concordance with the approach already undertaken in Chapter 2 (cf. the non-linear definition of (dis)fluency and the functional ambivalence of fluencemes).

3.3 Are discourse markers fluencemes?

Fluency has been mentioned in passing in the two previous sections, either regarding the definition of the DM category (Section 3.1) or the models for its functional analysis (Section 3.2). However, the link between DMs and fluency remains to be addressed in order to (i) describe the specific contribution of DMs to fluency and disfluency, with particular reference to the notion of (non-)linearity and (ii) situate them within the typology of fluencemes. I will first discuss the rare studies which tackled the relation between DMs and (dis)fluency and try to explain their relative absence from the literature (Section 3.3.1). Then I will develop the particular features of DMs which are directly connected to non-linear processes of production and perception, thus highlighting their relevance in a temporal and (non-)linear view of (dis)fluency (Section 3.3.2). Hypotheses concerning the general behavior of DMs and their inter-relation with fluencemes in particular will be laid out in Section 3.4, synthesizing the theories and models discussed throughout this chapter.

3.3.1 “Fluent” vs. “disfluent” discourse markers

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 (see Section 2.2.3.2.1), where DMs are discarded on the grounds that their optionality is uncertain. While some uses of DMs are disruptive and removable, others are more clearly fluent and useful. 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).²⁴

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 2015a) 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 multifunctionality. 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 multifunctionality of DMs is compatible with the present approach, 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.

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

²⁴ A similar restriction can be found in Pallaud et al. (2013a: 10), who found DMs to be included in 10% of “disfluent interruptions”.

method, while the latter chooses 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, as in 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. In the studies reviewed so far, DMs are either not included at all (e.g. Shriberg 1994), or only partially (e.g. Strassel 2003).

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), and 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, crushing voice) 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 multifunctionality, 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 large 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, beyond the three general types which they identified and which seem to be highly based on the semantics of the expression.

The last 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. She also identified a higher multifunctionality 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), a flexibility in levels of analysis which will prove relevant to the present study as well. However, 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.

In the last three references, DMs and disfluencies are studied jointly and considered to be related yet distinct phenomena. Some features of DMs could indeed be argued to set them apart from the rest of fluencemes and thus partly explain their exclusion, restriction or treatment as a separate category in other studies. I will now develop the two main features that I find relevant in this matter and stress how such features do not prevent a DM-as-fluenceme approach (but even support it). The first feature concerns the lexical-pragmatic content of DMs, which is null or weak in most other fluencemes. Unlike unfilled pauses, truncations or identical repetitions which do not add any sort of content to the utterance but merely serve to halt or interrupt the unfolding of the utterance, DMs (usually) stem from function-words (conjunctions, adverbs) with a core meaning – albeit procedural. For instance, traces of the perception verb *to see* can still be found in the DM use of *you see* and partly explain its attraction to comprehension-ensuring functions (cf. Denke 2009).

However, DMs are not the only fluencemes that have an impact on the interpretation of utterances: substitutions, by changing one propositional referent with another, obviously target the content of the message; filled pauses have also been claimed to perform similar signaling functions as DMs (Swerts 1998; Clark & Fox Tree 2002; Tottie 2015a), although their contribution to utterance content is less clear. In addition, the pragmatic meaning of DMs is available to the analyst and can be used as a qualitative gateway to on-going cognitive

processes, thus helping disambiguate the source and degree of the disfluency. For instance, an occurrence of *I mean* embedded in a substitution could possibly suggest to the hearer that the relation between the substituted and substituting segments is one of reformulation or correction, and not one of enumeration (see the analyses in Chapter 7). Overall, thanks to their pragmatic meaning, richer than most other fluencemes in the typology, DMs appear as privileged windows onto (dis)fluency and cognitive processes in the making.

The second feature of DMs is their ambivalence. Although postulated for all fluencemes in the typology, the ambivalence of DMs is quite extreme considering the broad range of their functions at various levels of discourse, as acknowledged by all frameworks in the field. DMs are, along with pauses, probably one of the most non-optional fluencemes in the typology, since most (if not all) of their uses contribute to the interpretation of utterances in useful, meaningful ways, which prevents their removal from transcription and from analysis without a significant loss of information (cf. the exclusion of DMs from disfluencies for this very reason). DMs cannot and should not be “cleaned out”. This stands in rather sharp contrast with some accounts of DMs as disfluencies, where they are considered to be “secondary variables” of fluency, as in Grosjean & Deschamps (1975) or Götz (2013), on the grounds that, unlike temporal variables, they do not necessarily occur in all utterances. While the syntactic optionality of DMs is indeed criterial to their status (cf. Section 3.1.1), they are definitely not as “removable” as other fluencemes, as attested by their exclusion from many typologies. Overall, the ambivalence (and resulting optionality) of DMs is not viewed as contradictory with their categorization as fluencemes, but rather in support of it, in the present approach to fluencemes as potentially fluent or disfluent devices depending on context (cf. Chapter 2). I would therefore argue that DMs are full-fledged markers of (dis)fluency, in keeping with the assumption of functional ambivalence (symptom vs. signal) underlying the present research.

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. Beyond the limitations of generalizability and comparability of these works, I do agree that the DM category is highly heterogeneous and complex, and that it is not highly informative – nor advisable – to include a great variety of DMs (connectives and others alike) without a framework or method that allows further distinctions to be reliably made between DM types. We can expect very different tendencies regarding frequency, register variation and combination patterns depending on the position and/or function(s) of the DMs, so that results would be particularly hard to interpret on the whole unfiltered category. 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.

In the next section, I will point out what specific features and functions of DMs are particularly relevant to fluency, in order to further justify the inclusion of DMs as one type of fluenceme and to generate hypotheses against the theoretical background laid out so far in this thesis.

3.3.2 The non-linearity of discourse markers

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 “powerlessness”, 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” (cf. their theory of information packaging, Section 2.1.2). 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. Given the many directions that such a program could take, I will henceforth focus on the aspects of DMs behavior which are related to (non-)linearity, in keeping with the present definition of (dis)fluency.

As a first general observation, I take up Levelt’s (1983) notion of linearization and his distinction between macroplanning and microplanning. Speakers have to handle mental organization of complex information into the linear channel of speech (macroplanning), all the while managing the actual form and style of the linguistic output (microplanning). Within these on-line simultaneous processes, DMs function retrospectively and prospectively for both hearer and speaker as frames or signals to orient the production and comprehension:

- For the speaker, DMs are routinized expressions whose automaticity leaves some time and cognitive resources for planning the rest of the utterance. They set up the structure of pairs of utterances and thus guide production (e.g. an *if*-clause is often followed by a *then*-clause). They allow to modify, specify or correct already-pronounced utterances by adding backward-looking connections such as a post-posed concessive comment (e.g. *though*), a reformulation marker (e.g. *at least*) or a mitigator (e.g. *sort of*).
- For the hearer, DMs announce the nature of the relation between the upcoming utterance and the previous one (e.g. causal, contrastive, etc.), which is especially crucial when the relation is reformulative (e.g. *I mean*), in which case the DM cancels past inferences.

DMs signal changes in the context and frame of reference against which the hearer has to build a coherent representation of discourse.²⁵

One of the most striking non-linear effects of DMs is probably their use in organizing higher-order events, i.e. either temporally-related facts or discursively-hierarchical topics. Examples (15) and (16) illustrate these two types of linearization at the content- and discourse-level, respectively.

(15) and with nine minutes to go to half time the soviet union have a penalty kick (0.320) **before** the kick will be taken attention is going to be paid (0.760) to Kolianov who was brought down quite ruthlessly just a few yards outside the penalty area (EN-spor-04)

(16) the funny thing is that none of the sort of Nancy Mitford stuff (0.220) do I mean Nancy (0.260) I can never remember which Mitford is which (0.253) **but anyway** none of the (0.280) u and non-u stuff seems to have washed off on your mother at all (EN-conv-01)

In (15), the sports commentator follows the events of the game as they unfold: the penalty kick is announced, one player is down, and only then will the penalty kick be taken. We see that the “before” allows the speaker to retract from the linear narrative which could be expected from the previous utterance: “the soviet union have a penalty kick”, we will now see the penalty kick, but “before” that “attention is going to be paid” etc. In (16), the speaker inserts a parenthetical aside where he questions the correctness of his choice of words (“Nancy Mitford”), before he takes up his original utterance and modifies the problematic segment. In this second example, the DM virtually cancels the impact of the aside by resuming the main previous topic (“none of this stuff seems to have washed off on your mother”).

This use of DMs corresponds to Auer’s (2005) analysis of delayed self-repairs as structuring devices for complex information. While he focuses on the recycling of syntactic constructions to correct or continue previous talk, DMs are here considered to fulfill a similar structuring function. Auer (2005: 79) directly connects this affordance of spoken language with Levelt’s (1981) linearization problem (“how to translate complex, hierarchically structured information into the linearity of speech”) as well as to the temporal nature of speech: “speakers are caught in a permanent cognitive conflict between, on the one hand, the tendency to formulate first what to them appears to be the most important information [...], and, on the other hand, the necessity to establish common ground on which this information can be processed” (2005: 80). Such a reading of delayed self-repairs seems highly compatible with uses of DMs as illustrated by the two previous examples.

The non-linear role of DMs is also put forward by Pawley & Syder (2000) and their “one-clause-at-a-time hypothesis”, according to which speakers are only able to encode the full content of independent clauses in a single planning act, although they also appear to occasionally plan more complex units (cf. Section 2.1.2). According to them, connecting

²⁵ This function of DMs evokes Van Dijk’s (1997: 194) notion of “cognitive context models”: “participants need to constantly monitor the other participant(s) as well as the other elements of the context and adapt their context models accordingly in order to be able to participate appropriately and competently”. In this perspective, DMs ease the switch from one context model to another.

several clauses into a complex unit of meaning constitutes the basis of fluency. Within connection, they distinguish chaining (weak integration, such as coordination) and integrating (high integration, such as embedded clauses). Although these notions only apply to relational DMs (connectives), we can transfer this scale of integration to the use of coordinating (e.g. *and*) vs. subordinating DMs (e.g. *although*). Subordination reflects a higher level of planning (both the main and embedded clauses are planned in advance) which is more cognitively demanding. Pawley & Syder (2000) found that embedded clauses (integrating mode) are less frequently preceded by a disfluency than in the chaining mode; however they present more utterance-internal disfluencies. This could mean that the overall plan is ready before the onset of the embedded clause, but not its full content. However, this theory is challenged by non-linear views of speech production (cf. Fortescue 2007; Section 2.1.3), according to which “the end need not be predetermined when the beginning is already being produced” (Fortescue 2007: 342). It remains that DMs are generally connected to our planning system, helping it and signaling its troubles.

The final aspect of DMs that intrinsically ties them to temporality and non-linearity is their distribution in the speech string. As developed in Section 2.1.2, the rhythm of spoken language is cyclic but not continuous, and speakers segment their talk by alternating between planning phases and production phases. As a consequence, some slots in the flow of words are more prominent than others, namely the boundaries between major units or discourse moves. It is typically at these unit boundaries that DMs tend to occur, that is, in initial position of utterances. We have already seen how the initiality of DMs is related to their inter-sentential scope (see Section 3.1.1) and we can now also associate it with the beginning of planning cycles, as Greene & Capella (1986) who found more hesitation around cycle boundaries. Speakers initiate new units with DMs to signal that planning is either on-going or completed and to further specify the nature of the relation with previous context. Although DMs are possible in medial and final positions as well, their high attraction to initiality confirms their sensitivity to the temporal rhythm of language and their role in planning and coherence, in addition to their many non-linear functions.

Overall, DMs seem to be the by-products of cognitive processes and are, therefore, bound by the temporal and non-linear nature of speech production. On-line production is constrained by the limitations of our mental systems – especially those of our working memory – to handle past, current and upcoming information (cf. Section 3.2.2). Fehrer & Fry (2007) found indeed that speakers with a lower memory ability produce more “hesitation phenomena”, which include “automatisms” such as *sort of* mostly functioning as strategic devices for planning management. They also showed that speakers use more of these time-buying devices in their L2 than in their native language, which indicates their connection to a high cognitive load.²⁶

In sum, DMs reflect and support non-linear processes of production and comprehension. Figuratively, it could be said that their multiple scopes and crucial role in planning processes

²⁶ Cognitive load (Sweller 1988) is quite a complex notion which can be defined and measured in various ways. It will only be used here in its basic understanding as the amount of pressure and complexity imposed on the production processes by the context and speaking task.

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 linear string of words, in a relative freedom of movement without which communication cannot seem to be performed efficiently.

3.4 Hypotheses: the place of DMs in the typology of fluencemes

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, in addition to those developed in Section 2.4 on the variation and combination of fluencemes in general. Two 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 (Chapter 6). These two chapters also correspond to a difference in analytical levels, viz. DM-based and sequence-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, while Chapter 6 strives towards more theoretical explanations of the observed patterns, in light of the usage-based framework developed in Section 2.3. The hypotheses governing the analyses in Chapter 7 will be developed in their own section (Section 7.1.4) given that they are somewhat independent from the other two chapters, although pertaining to the same general approach and integrating results from Chapters 5 and 6.

The rationale behind this dual structure of analyses is that there is, in principle, no one-to-one mapping between DM features and their relative fluency (e.g. the relational vs. non-relational divide does not *a priori* correspond to different degrees of fluency). Therefore, any tentative scale of fluency should rely on fine-grained analyses of the functions, position and combination patterns of DMs and their inter-relation with the others members in the fluenceme typology. This flexibility in analytical levels is also consistent with the endeavor to study language in use through lenses of varying granularity, from actual occurrences to increasingly abstract categories, in line with the usage-based notion of schema abstraction. Concretely, DMs will be studied (i) individually, making use of all the information available regarding their behavior in context (Chapter 5), (ii) in combination with fluencemes, at a generic sequence-level focusing on formal configurations (Chapter 6, Sections 6.1 – 6.3) and (iii) integratively, synthesizing both levels of analysis (Chapter 6, Sections 6.4 – 6.6).

First, at DM level, corpus observations will be analyzed along six dimensions, integrating more and more variables from monofactorial to multifactorial analyses:

Language variation: as mentioned before, previous contrastive research does not suggest any expectation of differences between French and English at the basic level of frequency. Similarly, I expect the most frequent DM expressions to be semantically equivalent. The French

top-three DMs will be compared to Boula de Mareüil et al.'s (2013) corpus findings (*et, alors, mais*, cf. Section 3.3.1).

Register variation: 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. Similarly, in light of the hearer-oriented uses of DMs and their role in turn-taking and turn-holding, interactive registers (i.e. dialogal, free exchange) should show a higher frequency and variety of DMs.

Syntax: the feature of initiality will be tested in order to find out in what proportion of all DMs it actually applies. Given the consensus in the literature, I expect to find more initial DMs, although not all types of units (clause, dependency structure or turn) should be affected by initiality in the same proportion. I further suggest to look into the relation between grammatical class (part-of-speech) and position to uncover whether the less typical positions (i.e. medial and final) are occupied by specific forms of DMs.

Functions: as suggested by some definitions of the DM category (e.g. Fischer 2000) and previous corpus studies (e.g. Denke 2009), I expect DMs to function most frequently as structuring devices, which corresponds to the sequential domain in the present taxonomy. I further intend to test the extent to which relationality is co-dependent with specific domains or functions, looking in particular for functions that can be used in both relational and non-relational contexts. This analysis will also uncover any DM expressions which are specific to particular functions, domains or (non-)relational types.

Co-occurrence of DMs: I expect DMs to frequently co-occur with one another, following the general tendency of fluencemes to cluster and combine. Degrees of fixation from Cuenca & Marín's (2009) tripartite system will be confronted to corpus annotations.

Integration of DM-level variables: any meaningful co-variation of features will be statistically identified. In particular, the following hypotheses gathered from the literature will be tested:

- **position by function:** interpersonal DMs should be attracted to the final position of utterances (Degand 2014); sequential DMs should occur primarily in initial positions given their role in higher-level discourse planning; medial position should be mainly associated with metadiscursive (rhetorical) DMs used by the speaker to comment on the on-going utterance;
- **co-occurrence by function:** based on Cuenca & Marín's (2009) system, I expect the most frequent combinations to contain DMs expressing similar functions in similar directions or scopes; this analysis should uncover potential "complex" candidates, i.e. combinations of DMs which are on the verge of becoming new fixed units;
- **co-occurrence by position:** again, given the tendency of fluencemes to cluster at discourse boundaries, I expect co-occurring DMs to show higher proportions in initial position;
- **function by register:** Denke (2009) suggests that text-structuring (sequential) functions are favored in monologic situations; interpersonal functions should, by definition, be more frequent in interactive dialogues; ideational DMs (objective relations) should be

prevalent in factual discourse types such as news broadcasts or political speeches; I expect intermediary registers (e.g. interviews) to show a greater variety of functions, as opposed to more extreme speaking tasks, viz. very informal and very formal, which should be respectively associated to interpersonal and ideational functions;

- **any variable by language:** since no specific expectations have been drawn for differences between English and French, I will rather strive to identify any language-specific pattern of co-variation for any of the variables studied so far.

To summarize the analyses at DM-level, I intend to verify in what proportions the core features mentioned in most frameworks (initiality, relationality) actually apply to the full category of DMs as presently defined, 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.

Secondly, at sequence level, the patterns previously identified will be specified by the presence and configuration of co-occurring fluencemes in the direct co-text of DMs. Starting with their syntagmatic behavior, I will look for recurrent combinations and test the following hypotheses:

- **high frequency** of particular sequences should be associated with a low degree of intrusiveness (or disfluency) of the fluencemes in the cluster: in other words, rare sequences of fluencemes should be more marked and less ambivalent than pervasive sequences, following the usage-based assumption that frequency enhances cognitive entrenchment (cf. the fluency-as-frequency hypothesis, Section 2.4);
- the prominent role of **unfilled pauses** within the fluenceme typology (e.g. Grosjean & Deschamps 1975; see Section 2.3.2) should be reflected by their frequent clustering with DMs;
- a more general hypothesis on **combination patterns** suggested by Beliao & Lacheret (2013) poses that DMs very often cluster with other fluencemes, whereas the opposite is not necessarily true in the majority of cases (i.e. sequences of fluencemes without a DM); the position of DMs with respect to other fluencemes in a sequence will also be compared to Boula de Mareüil et al.'s (2013) finding, i.e. DMs should more frequently precede than follow other disfluencies.

These analyses only take into consideration the frequency and syntagmatic configurations of the sequences, and are thus limited in their interpretation of relative (dis)fluency. Multifactorial models are therefore added to integrate the variables from the previous two sets of results:

- **domain by sequence type:** 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. Given the difference in cognitive processing identified in the experimental literature between semantic and pragmatic discourse relations, ideational functions of DMs should be less often clustered with

fluencemes than their rhetorical counterparts; any register-specific or language-specific association of domain and sequence type will be identified at various degrees of abstraction;

- **position by sequence type:** as a follow-up on the previous hypothesis, clusters of text-structuring DMs with pauses should occur utterance-initially; medial positions, given their rather intrusive effect, should be occupied by fluencemes signaling more “urgent” cognitive processes (e.g. inference-canceling); final positions, which I have previously hypothesized to favor the occurrence of interpersonal DMs, should be associated with hearer-oriented or trouble-signaling fluencemes through which the speaker is being collaborative;
- **Potentially Disfluent Functions:** 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 Appendix 1 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;
- associations of DM functions and fluenceme clusters will be qualitatively interpreted in terms of a **cognitive-functional scale of (dis)fluency**; it is assumed that the resulting schemas are not equally relevant or coherent along the scale across all degrees of abstraction, but rather that such interpretation of relative fluency can only apply at a certain level of generalization, beyond which particular occurrences start to diverge; in other words, I will try and identify at what level of abstraction the scale of (dis)fluency starts or stops making sense with the data.

The analyses proposed throughout this section are summarized in Table 3.4.

Table 3.4: Summary of analyses

	Monofactorial	Multifactorial
DM	language	position by function
	register	co-occurrence by function
	syntax	co-occurrence by position
	function	function by register
	co-occurrence	any variable by language
Sequence	fluency-as-frequency	function by sequence
	unfilled pauses + DMs	position by sequence
	independence of DMs	PDFs
		schemas on the fluency scale

At the end of Chapters 5 and 6, general tendencies will be revealed, including tentative interpretations of fluency. It is beyond the scope of this research to rank each and every occurrence of DMs and other fluencemes on the fluency scale, given the complexity and variability of such a task. Chapters 5 and 6 will serve as test beds for the types of conclusions that can be drawn from corpus-based analyses, by converging evidence from several independent layers of annotation (syntax and pragmatics, formal and functional). By contrast, the methodology and research questions of Chapter 7 are more qualitative, with a conversation-analytic approach to the data which should complement the results from the other chapters, in combination with formal variables and quantitative methods.

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

Introduction to the chapter

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 (Sections 2.4 and 3.4), 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 side 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* (Krzyszowski 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, and explain the rationale behind its design. 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). In the last section, I will present the different steps of post-treatment that were conducted to obtain the final enriched format of the dataset with additional variables and macro-labels across the two analytical levels, viz. DMs and sequences of fluencemes.

4.1 Data

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 thesis (Section 4.1.1). They were gathered from available source corpora in French and English (Section 4.1.2) in a comparable corpus design (Section 4.1.3) and underwent a uniform technical formatting (Section 4.1.4). The contribution of this dataset lies in the rich annotations that were manually added to the original texts, following the procedures that are described in Sections 4.2 and 4.3. This first section covers all the preliminary steps related to corpus compilation.

4.1.1 From research questions to corpus selection

The research questions and hypotheses pursued in this thesis, detailed in the two previous chapters (Sections 2.4 and 3.4), call for specific requirements in terms of data type and corpus size. Four aspects are particularly relevant to the corpus design presented in this chapter, and will therefore be briefly discussed in the following.

The first aspect concerns the quantity of data adequate for the (manual) annotation and analysis of discourse markers and other fluencemes. While it is generally acknowledged that DMs are quite frequent in speech (e.g. Brinton 1996), the fact remains that they are syntactically

optional (cf. the definition in Section 3.1.4) and therefore not as pervasive as pauses, for instance. Moreover, this thesis adopts a bottom-up approach to the whole category of DMs, rather than a case-study approach to particular lexemes, so much so that the corpus has to be large enough to show sufficient occurrences of the different forms and functions of the DM category. On the other hand, the interpretive nature and high precision of the annotations considered in this research require manual encoding by a trained annotator, which is not reasonably feasible on too large a corpus, as opposed to the current “big data” trend to automatically process (very) large corpora. To sum up, the corpus should be large enough to cover as much variation as possible, while remaining manageable for manual annotation.

Secondly, regarding the types of texts to include in the corpus, the crosslinguistic and variationist hypotheses of this thesis have to be tested on a variety of registers in English and French so that the frequency and use of different forms and functions of DMs and other fluencemes can be contrasted. In order to represent a broad range of communication settings with scalar differences, a rather large number of registers (eight) was selected. While this choice has led to complications in terms of the availability of source corpora, I believe that it vouches for a more fine-grained analysis of the contextual effects on DM and fluenceme use, revealing the impact of specific communicative features (see the refined metadata system in Section 4.1.3). To cope with the rarity of some data types, priority was given to the balance (in word count and duration) between the two languages for each register (e.g. French conversations and English conversations), rather than between registers within one language (e.g. French conversations and French interviews), as will be detailed in Section 4.1.3. Furthermore, DMs are particularly frequent and varied in natural impromptu conversation (e.g. Brinton 1996; Georgakopoulous & Goutsos 1998), hence the larger proportion of such registers compared to more constrained speaking tasks, like classroom lessons, which are also less easily available.

The third aspect concerns metadata: the present research does not target sociolinguistic variation in age, gender or language variety, and therefore does not require this type of speaker information in the corpus. While very interesting, sociolinguistic studies usually investigate a restricted number of DM expressions (e.g. Beeching 2007 and Dostie 2009 for French; Andersen 2001 and Pichler & Hesson 2016 for English), as opposed to the present paradigmatic approach. Given the relatively large quantity of data in *DisFrEn* and the variation in the internal structure of the subcorpora (see Section 4.1.3), it can reasonably be assumed that idiosyncratic tendencies due to sociolinguistic characteristics do not introduce a bias for one type of profile or another, so much so that no reference to speaker profiles will be made in the remainder of the thesis. Such a line of investigation remains, however, an interesting avenue for future research.

Finally, the fourth requisite based on the research objectives is related to the format of the corpus: when studying spoken data, it is paramount to have access to the audio context, especially for the analysis of fluencemes and the functional interpretation of DMs (see Section 4.2.2.2). Therefore, one major constraint for the corpus design was to select texts from source corpora which make the audio files available in addition to the written transcripts. However, as explained in the Introduction (Section 1.2), this thesis does not tackle any sort of prosodic analysis apart from the basic identification of unfilled pauses (see Section 4.3.2), given that the prosody of DMs and other fluencemes constitutes an independent object of study, as attested

by the numerous works focusing on this phenomenon (e.g. Moniz et al. 2009; Lundholm 2015), which is outside the scope of the present research. As a result, high audio quality was not required since no (automatic or other) phonetic analyses were carried out. These theoretical and practical prerequisites led to the compilation and design of *DisFrEn* as explained in the following sections.

4.1.2 Source corpora

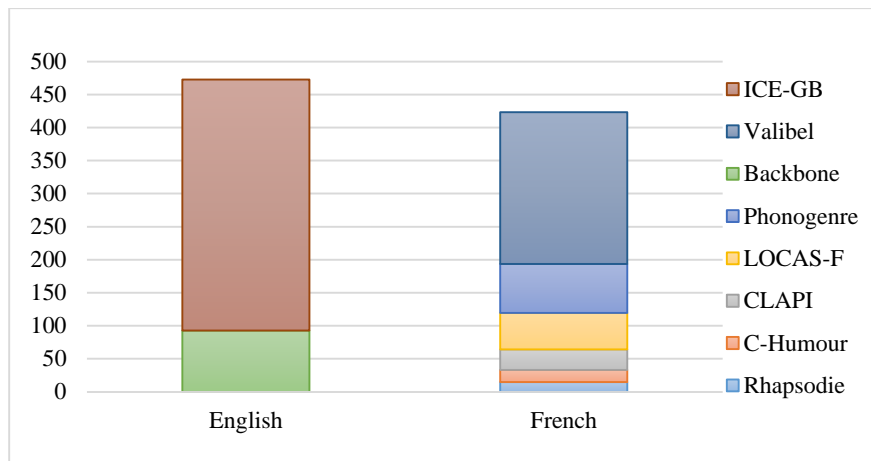
Spoken corpora and databases do not benefit from the same practical convenience as their written counterparts in terms of size and availability of the material, for obvious reasons related to the more intrusive technique to collect it (cf. “observer’s paradox”, Labov 1972) and the human- and time-cost 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/>) or the C-PROM corpus for prosodic analysis (Avanzi et al. 2010). 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 that of databases (e.g. Corpus de langues parlées 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 *DisFrEn*, as can be seen in Figure 4.1. As we can see, 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 (number of speakers, degree of preparation, see next section), thus allowing me to select the desired number and types of transcriptions. Although speakers' age is a well-known relevant variable in discourse markers use (e.g. Andersen 1997), it is not available in the metadata of this corpus; however, as mentioned in Section 4.1.1, sociolinguistic variables are not the focus of this research, so much so that this shortcoming has a limited impact for the present purposes.

Figure 4.1: English and French source corpora in *DisFrEn* (in minutes)



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 (see below Section 4.2): 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.

As for 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, comprising 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 certain 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

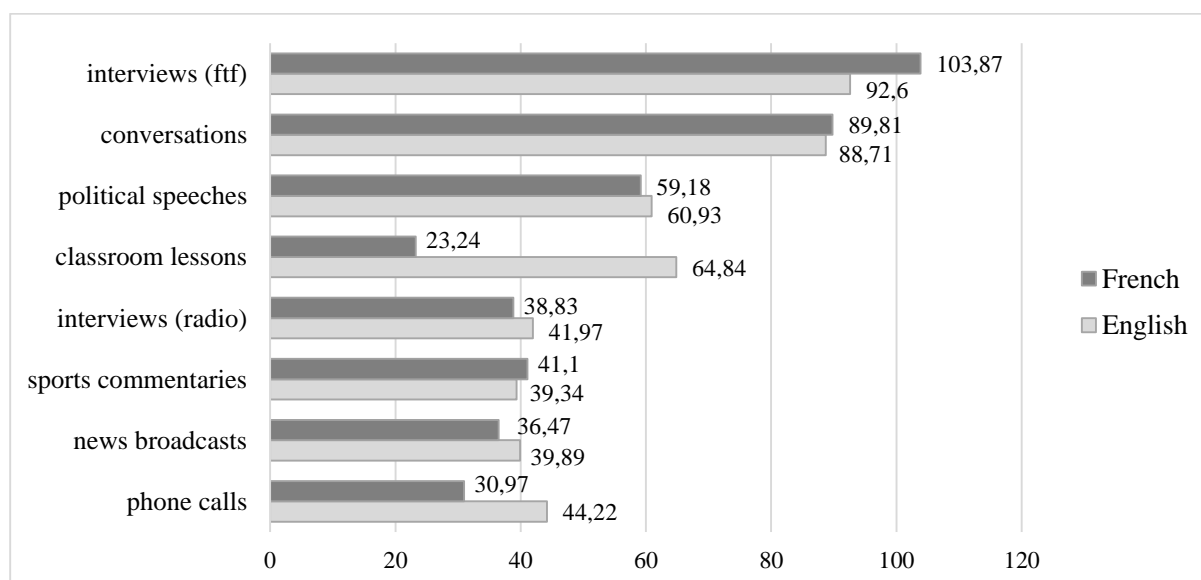
such as age or language variety, with productions from both French and Belgian speakers. It is rather the content of these resources in terms of registers and speaking tasks which was considered, as detailed in the following section.

4.1.3 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. As mentioned above, 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) as well as the theoretical interest and focus of the present research (see Section 4.1.1).

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. The goal was to reach one hour of language per register per language on average: it was successfully met with 896 minutes in total (about 15 hours of recordings), giving an average of 56 minutes for each register in each language. However, the internal structure diverges from this ideal, both at language- and register-level, as can be seen in Figure 4.2.

Figure 4.2: Distribution of registers (in minutes) per language in *DisFrEn*



It clearly appears that two registers emerge as the most represented in the corpus, namely face-to-face interviews and conversations with respectively 196 and 179 minutes for both languages, which corresponds to 35,098 and 34,911 words (out of 161,700 words overall in *DisFrEn*). This difference with the other registers is voluntary and reflects an interest for spontaneous language

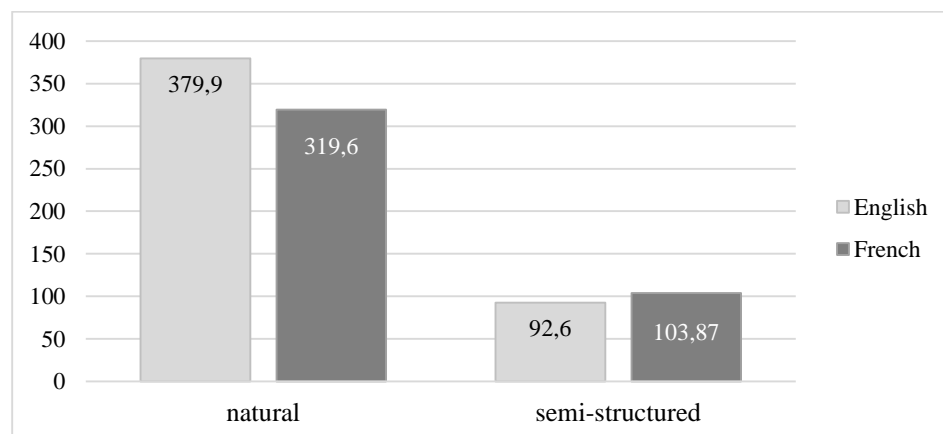
use, as opposed to more formal registers with a more conventional and fixed setting such as news broadcasts. Most other settings comprise around 40 minutes in each language, except for political speeches and English classroom lessons which are slightly larger with 60 minutes.

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, this French subcorpus will not be considered for the contrastive and register analysis in raw frequencies but only when relative frequencies per thousand words are computed for all subcorpora. In fact, this precaution will be taken for any quantitative results discussed in this thesis to ensure the comparability of the different subcorpora, even those which share a similar number of words.

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 2.4 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 jointly with the other project members (A. Dumont, I. Grosman and I. Notarrigo).

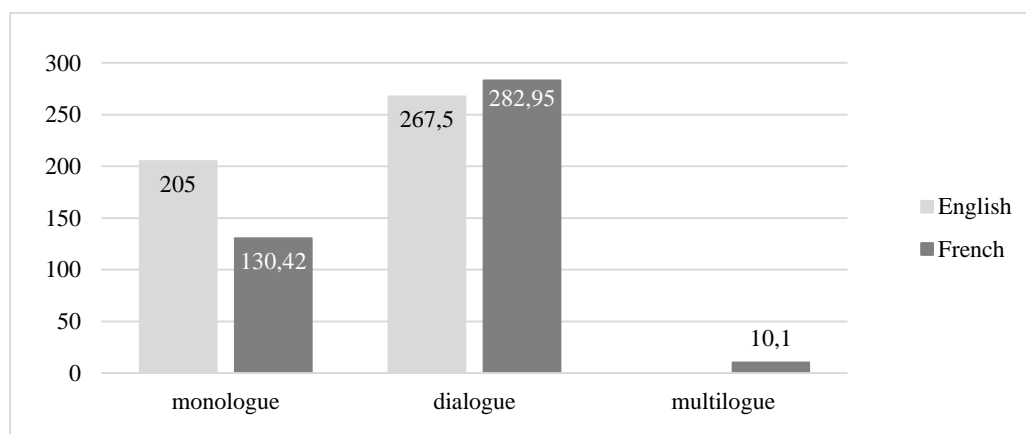
In the remainder of this section, each situational feature will be defined, and the content of *DisFrEn* will be graphically represented according to this second, more refined system. Not surprisingly, the same contrasts and gaps as in the other system are still in order, namely the preference for natural and spontaneous data and the rarity of certain data types. 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). We can see in Figure 4.3 that natural registers are much more frequent in the corpus, restricting semi-structured data to the face-to-face interviews.

Figure 4.3: Levels of elicitation in *DisFrEn* (in minutes)



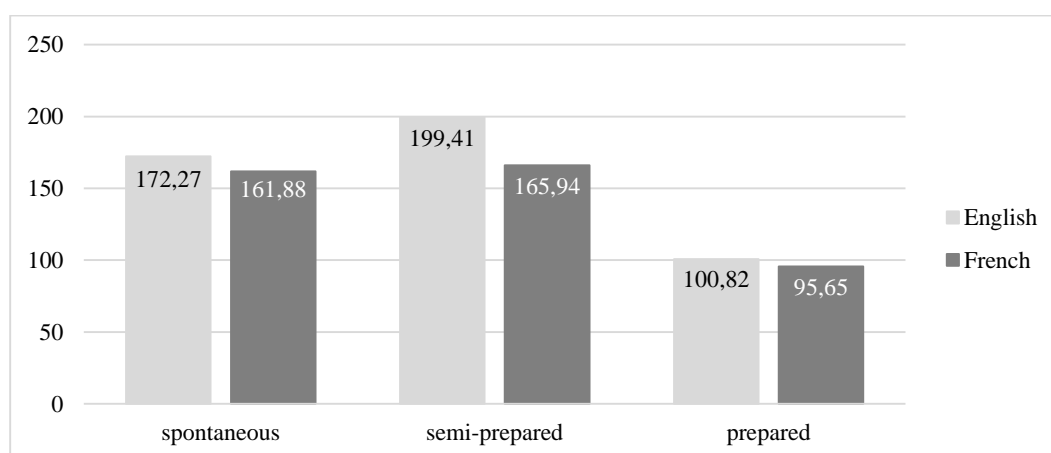
The second feature is the number of speakers actively taking part in the interaction, thus excluding by-standers and silent participants. It is fairly basic and distinguishes monologues, dialogues and multilogues (anecdotal in the corpus), with the distribution represented in Figure 4.4. We can see that while dialogues are equally represented across English and French, there is a seventy-minute difference in monologues which mainly corresponds to the quasi-absence of classroom lessons in French.

Figure 4.4: Number of speakers in *DisFrEn* (in minutes)



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 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, as can be seen in Figure 4.5.²⁷

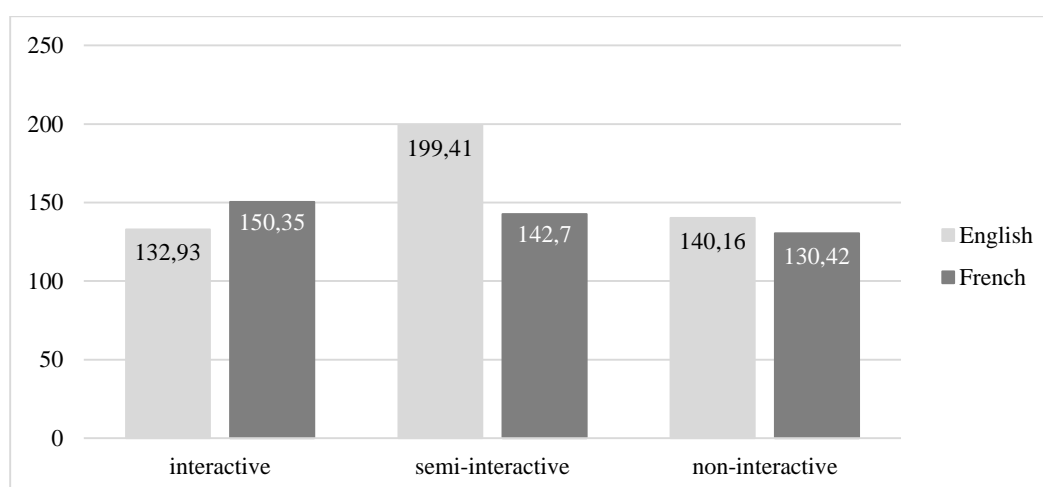
Figure 4.5: Degrees of preparation in *DisFrEn* (in minutes)



²⁷ Face-to-face interviews are considered as semi-prepared settings since a general topic has been pre-established and the interviewer follows a set line of questioning (displayed in writing to the interviewees in the French interviews). Interviews are therefore not as prepared as a classroom lesson, for instance, yet there is considerably less freedom of topic and more preparation (at least for one of the participants) than in conversations.

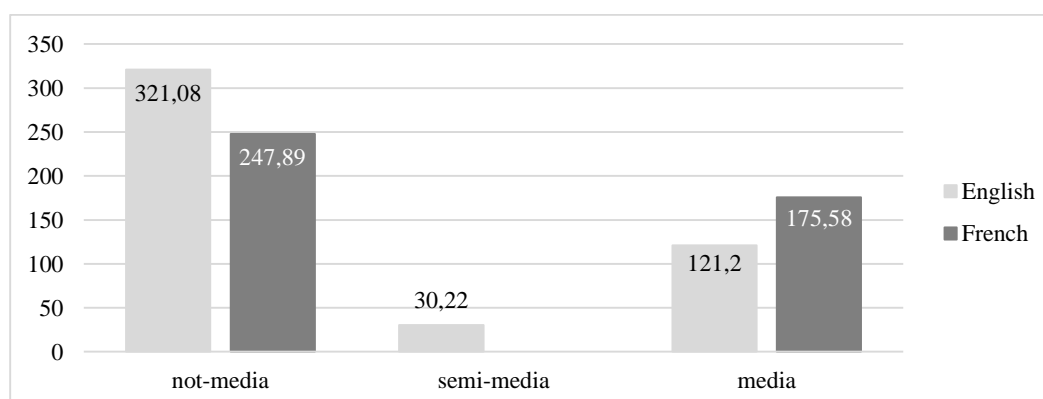
Another feature closely related to the situational hypotheses on fluent and disfluent speech is that of interactivity, i.e. the speaker's ability to adapt their speaking behavior to the other speaker's with respect to what is expected from their status in the interaction. Interactive registers are characterized 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 punctual 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). As can be seen in Figure 4.6, all three levels are represented in similar proportions in *DisFrEn*.

Figure 4.6: Degrees of interactivity in *DisFrEn* (in minutes)



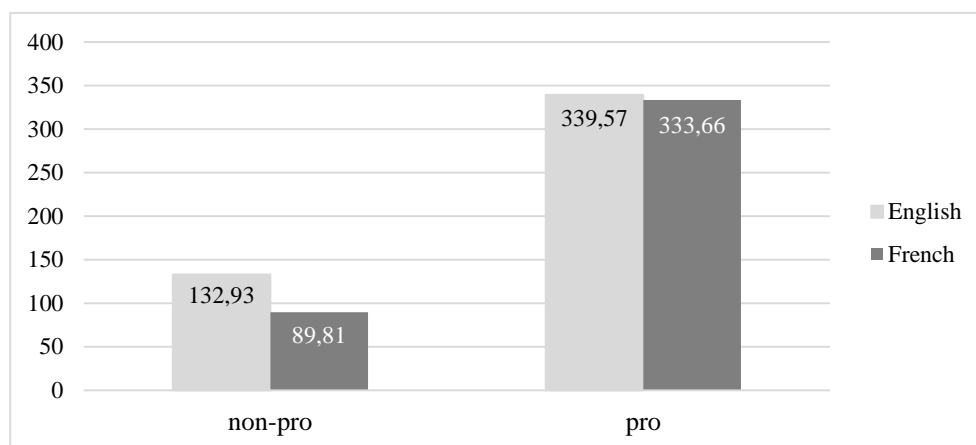
Then, the feature of media coverage defines the extent to which broadcasting is the main goal of the interaction, with three levels again: broadcasting is the main aim of the interaction (“media”); the interaction is broadcast but would have taken place even without broadcasting (“semi-media”); the interaction is not broadcast (“non-media”). In *DisFrEn* (see Figure 4.7), the intermediary level is only represented by English political speeches which consist in parliamentary debates, thus differing from their French counterparts which are recorded to be TV-broadcast.

Figure 4.7: Media coverage (broadcasting) of the registers in *DisFrEn* (in minutes)



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. We see in Figure 4.8 that in *DisFrEn*, for reasons of material availability, professional settings are much more frequent than non-professional settings.²⁸

Figure 4.8: (Non-)professional interactions in *DisFrEn* (in minutes)



All these graphs show that there is no perfect balance between the different levels of all these situational features. Instead, the design of *DisFrEn* was based on task labels and subjected to theory-motivated preferences, as well as practical limitations. 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 do not provide a more balanced picture of the corpus design of *DisFrEn* but are mostly referred 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.

4.1.4 Technical treatment

A natural consequence of the high number of source corpora in *DisFrEn* is the heterogeneity of their input format, be it in terms of extension types, transcription conventions, type of segmentation and audio alignment (if any) or treatment of overlaps. This diversity is not only problematic for practical reasons but also involves more theoretical issues such as the accuracy of the transcription compared to the original speech, a problem found in texts from the CLAPI corpus for instance where transcribers excessively transcribed the language as strongly colloquial, while a more standardized version was actually uttered, thus making it necessary to

²⁸ 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.

transcribe each text again following the Valibel conventions (Bachy et al. 2004). Since the whole corpus was annotated within the same software, namely the EXMARaLDA suite (Schmidt & Wörner 2012), the input format of all these texts needed to be homogenized through several technical modifications in order to meet our methodological requirements as well as the specific configuration of the EXMARaLDA interface.

The first and most time-consuming task was text-to-sound alignment at word-level, which was either not provided by the source corpus or only at turn-level. The automatic aligner EasyAlign (Brognaux et al. 2012; Roekhaut et al. 2014) was used for this process after manual re-transcription and approximate segmentation when necessary. All texts were then part-of-speech-tagged with TreeTagger (Schmid 1994) and converted to the final .exb format (HTML compatible). Further standardization concerned the transcription of unfilled pauses, indicating the duration in seconds in parentheses by an automatic transformation of various symbols used in the different corpora to represent pauses (a hyphen, a backslash, etc.), although the duration conversion is often imperfect and cannot be relied upon for prosodic analysis. Finally, since the EXMARaLDA annotation and extraction tool requires that all annotation tiers apply to one and the same transcription tier, a merged transcription of all speakers was automatically generated (except for monologues), with some adjustments in the case of overlapping speech, namely overwriting the secondary intervention (e.g. backchannelling) by the main on-going turn.

In the end, each text in *DisFrEn* was sound-aligned, segmented at word-level following uniform rules and transcription conventions, and in line with the requirements of the EXMARaLDA interface. Some limitations remain with respect to the technical format of this dataset, especially for already aligned texts which did not undergo the full treatment described above and which therefore kept some of their irregularities. For instance, certain phrases like *parce que* were segmented either as one or two units, sometimes within a single text. However, any time such irregularities were found to concern the transcription of discourse markers or other fluencemes, they were directly corrected in the annotation file, so that the excerpts that were annotated in *DisFrEn* always follow the same conventions.

4.2 Annotation protocol at DM level

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 thesis 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 (cf. Section 3.2.1).

In this section, I will report on the corpus-based methodology used to elaborate the categorical definition of DMs already provided in Section 3.1.4, as well as the coding scheme and annotation procedure. For reasons of space, complete details of each variable accounted for in the coding scheme will not be repeated here (see the guidelines reported in Appendix 1, Crible 2014) but rather the decision-making process and the main criteria that were used during the annotation of *DisFrEn*.

4.2.1 Corpus-based coding scheme

This section presents the eighth version of the coding scheme, which is a revision of previous versions that revealed several flaws after a number of testing phases on a pilot corpus of interviews (cf. Section 4.1.2). It was primarily designed for spoken French and English, although it has also 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 (Gabarro-López under review), by both expert and naïve coders (Crible & Degand under review). The lessons from all these tests and applications to different data vouch for a robust annotation protocol with valid *tertia comparationis* across different languages and modalities.

4.2.1.1 Identification of DM tokens

As mentioned before, many different definitions of what is to be included in the category of DMs are conflicting in the literature. 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 3.1.4:

DMs are a grammatically heterogeneous, syntactically optional, multifunctional type of pragmatic markers. 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, expliciting the structural sequencing of discourse segments, expressing the speaker's meta-comment on their phrasing, or contributing to the speaker-hearer relationship.

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 (cf. Section 3.1.1). Particular occurrences of DMs can show some of these features and not others, as in Examples (1) and (2). In (1), the French DM “dis” (‘tell me’) is monosyllabic yet not frequent in *DisFrEn* with only one occurrence. On the other hand, “c’est-à-dire” in (2) is longer but much more frequent, especially if we include occurrences of *c’est-à-dire que* (24 cases in the

corpus once combined), a variant which also illustrates a stronger degree of syntactic integration (or clause association in Schourup's (1999) terms) with the presence of the complementizer *que*.

- (1) je sais pas je demanderai (0.237) **dis** et là dans ton oreille qu'est-ce que tu peux mettre d'autre
*I don't know I will ask **dis** 'tell me' there in your ear what else can you put* (FR-conv-03)
- (2) il dit (0.626) fantômes de désir **c'est-à-dire** il emploie un mot merveilleux
*he says ghosts of desire **c'est-à-dire** 'that is to say' he uses a wonderful word* (FR-intr-03)

Examples (3) and (4) show different degrees of prosodic independence of the DM "so", first co-occurring with a pause at its left periphery, then with pauses at both sides; *so* can also occur without any pause or major intonation contour, thus being prosodically integrated.

- (3) you could do it between here and there (0.420) **so** there are a lot of different ways (EN-clas-04)
- (4) this thing has to be an interval (0.300) **so** (0.253) that tells you then that x goes (EN-clas-04)

As expressed in the definition of the DM category above (see also Section 3.1.4), the main criterion for DM status is a functional one, the other features being formal characteristics applying to the most typical members of the category.

In a next step, the confrontation of this definition to 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 as much as possible individual biases and the inherent ambiguity of speech, 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 (see next section). 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 (5)) but excluding a number of constituents such as relative clauses, infinitive, nominal and prepositional phrases (as in Example (6)) except when these are acting as a-verbal predicates.

- (5) 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)
- (6) 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 (5) 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 (6), 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 thesis and may therefore not directly suit other purposes. Another consequence is that the findings of this thesis 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 bottom-up identification procedure, provided they are motivated and documented.

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 will be 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 confrontation with other existing 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* – see Section 4.3.2 below). I will only expand here on the cases of partial inclusion, namely interjections, answer particles and general extenders. The first two follow the same functional criterion: interjections and *yes-no* particles are considered as DMs when they perform one of the thirty possible functions of the DM category (see next section). These cases are quite rare and very restricted: in *DisFrEn*, only the English interjection *oh* was found to express a DM function (namely introducing reported speech) as in Example (7).

- (7) well Matthew’s saying **oh** I’ll take them on the back of my bike and I’m going **oh** uh yeah ok (EN-phon-09)

In (7), the two occurrences of “oh” do not express surprise or disappointment, as the typical interjectional use would be, but rather signal the beginning of a reported speech segment, here a dialogue between Matthew (“I’ll take them on the back of my bike”) and the speaker (“yeah ok”). In the case of answer particles, in addition to their basic meaning of agreeing or disagreeing (which are themselves possible functions of DMs), the candidate DMs must also express another DM function (e.g. *topic-resuming*) simultaneously, thus excluding fully propositional particles which only answer a question. Answer particles in English and French represent less than 100 occurrences mostly combining an interpersonal or structuring function (see next section) with their basic value of agreement or disagreement, as in Example (8).

- (8) I run an amphibious tour operation in Plymouth (1.170) it’s called “ducks ‘n’ drake” (0.560) right (1.100) and (0.613) and uh **yeah** we run around Plymouth (EN-intf-02)

In (8), “yeah” does not answer any *yes-no* question (the previous question from the interviewer was “could you introduce yourself please and tell us a little bit about what you do here”, at the very beginning of the interview). The speaker merely punctuates his description, thus signalling the continuity with the previous theme and possibly referring back to the first mention of Plymouth, in the beginning of the excerpt. Such uses of interjections and particles therefore appear highly compatible with the present definition of DMs. As for general extenders, their annotation as DMs depends on their idiosyncratic nature: recurrent forms such as *and so on* or French *et tout ça* are selected, but more innovative creations which are not produced by more than one speaker in the corpus, although functionally similar, as in Example (9), are not selected.

- (9) I work for deals with translation of patents and occasionally trademarks (0.387) but mainly patents and kind of related legal work (0.249) so oppositions and all that kind of stuff and court judgments **and all that kind of jazz** (*Backbone* bb_en026)

Here, the speaker is clearly building on the more conventional form *and all that kind of stuff* (also present in this excerpt) to signal that the list is not complete, which is otherwise annotated as *approximation*. However, in order to remain consistent in the identification phase and avoid hesitations on the boundaries of the category, such borderline cases and others were excluded. More details can be found in Crible (2014) and Appendix 1.

4.2.1.2 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. 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 only existing in spoken data (cf. Section 3.2.2). 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.

To structure the multifunctionality of DMs into a reasonable number of macro-values (manageable for quantitative analysis), four domains have been identified from a review of the literature, mostly based on the seminal distinction between ideational/textual/interpersonal by Halliday (1970), but also Sweetser's (1990) content/epistemic/speech-act distinction and González's (2004, 2005) own four (ideational, rhetorical, sequential, inferential). These various proposals all have in common the dichotomy between semantic and pragmatic coherence relations, where the "source of coherence" (Sanders et al. 1992) comes either from the external world (semantic, objective) or from the speaker's subjectivity (pragmatic, subjective). The other distinctions that these authors make do not perfectly overlap with one another: "textual" in Halliday (1970) does not correspond to "epistemic" in Sweetser (1990) and might be broader than "sequential" in González (2005), for instance (cf. Section 3.2.1). Differences in research focus and/or degree of granularity between these systems and the present research called for the need to propose the following system with thirty functions grouped in their respective domain:

- **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 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 3.2).

In the revised version of the PDTB 2.0 by Zufferey & Degand (in press), this distinction is the last decision in the annotation tree: for instance, the generic value of "cause" is assigned before its discrimination as either semantic or pragmatic, thus restricting the hesitations related to this decision at the lower level of the annotation (cf. Table 3.3). Although each system has its own benefits and pitfalls, 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, since (dis)fluency and register variation can be affected by this dichotomy (Sanders 1997).

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 specificities of the spoken mode, which involved two types of revisions:

- simplification of previous distinctions to avoid ambiguity and over-specification: for instance, the PDTB 2.0 distinguishes six types of conditional relations; González (2005) makes a distinction between personal comment and personal evaluation, although their definitions are very similar (“introduce a personal comment” vs. “introduces a personal evaluation or comment”; 2005: 60, 62);
- re-categorization of functions in different domains, a problem often found in González’s (2005) proposal: for instance, “evidence” and “justification” are categorized in two separate domains (rhetorical and inferential, respectively) although the labels suggest a strong conceptual similarity; the sequential domain is defined as “delimit[ing] discourse segment boundaries” while the inferential one is defined as “facilitat[ing] contextual shifting onto new segment” (2005: 58), which might again seem very similar.

The revisions mentioned above mainly concern categories borrowed from other works which did not fully satisfy the present purposes. As such, the revision process is mainly based on the theoretical grounds of semantic coherence, as well as practical considerations for an operational annotation. 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 polysemous DMs as they emerged from the pilot study and to the mapping of semantic and pragmatic equivalents (see Appendix 1).

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 and multimodal data (see Section 4.2.2.3 for an assessment of replicability).

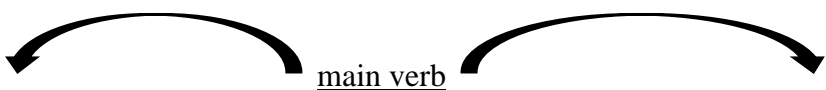
4.2.1.3 Three-fold positioning system

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 4.9.

Figure 4.9: Macro-syntactic segmentation for DM position

periphery	dependency structure			periphery
				
<i>but I mean</i>	<i>if it's empty</i>	<i>I'll just you know buy</i>	<i>fruit and like sweets</i>	<i>and so on</i>
pre-field	left-integrated	middle field	right-integrated	post-field
"PRE"	"LEFT"	"MID"	"RIGHT"	"POST"

We see that, 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.²⁹ 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. For instance, medial DMs such as "like" in Figure 4.9 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

²⁹ 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).

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.

Two values have been added to the original system of five slots in Lindström (2001) to better cope with the complexity of spoken data, following the recommendations of the MDMA project (Bolly et al. 2015), namely independent (only item in the unit, “IND”) and interrupted (position unclear due to incompleteness or interruption, “INT”). Detailed criteria and special cases are provided in the guidelines (Appendix 1), 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 (10). This variable consists of five values: initial, medial (preceded and/or followed by non-optional elements), final, independent, interrupted.

(10) 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. These positional variables resort to a rather non-consensual terminology (cf. Footnote 29) and might appear non-intuitive or even contradictory in cases such as (10) where the DM is considered both right-integrated and initial depending on the level of syntactic analysis. That being said, this very flexibility 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), an approach which is compatible with the general endeavor of this thesis to identify recurrent patterns at different levels of precision or abstraction. 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.

4.2.1.4 Other variables

Besides functions and positions, three other manually assigned variables are covered in the protocol, following the same methodology as that applied to the definition and revision of values. They were either borrowed from other works or applicable more directly to the data, and therefore required less operationalization and less innovation than the variables introduced so far. They are briefly presented in the following.

First, 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. 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 directly borrowed from the PDTB’s guidelines in Santorini (1990), with the exceptions of interjections (restricted to “primary” interjections following Norrick (2009)), prepositional phrases and subordinating conjunctions, which were originally grouped together for unknown reasons. 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 4.1 with examples.

Table 4.1: List of all part-of-speech tags for DMs (with examples)

CC	Coordinating conjunction	<i>and, but, or...</i>	<i>et, mais, ou...</i>
RB	Adverb	<i>so, actually, now, anyway...</i>	<i>donc, enfin, alors...</i>
VP	Verbal phrase	<i>you know, I mean...</i>	<i>tu vois, je veux dire...</i>
SC	Subordinating conjunction	<i>because, if, although...</i>	<i>parce que, même si...</i>
WP	Pronoun	---	<i>quoi, un, et tout</i>
NN	Noun phrase	<i>sort of</i>	<i>genre</i>
JJ	Adjective	<i>right</i>	<i>bon</i>
PP	Prepositional phrase	<i>in fact, for example...</i>	<i>au fond, par contre...</i>
UH	Interjection	<i>okay, yeah, oh...</i>	<i>hein, ben, ouais...</i>

This variable was included in the protocol to account for multi-word units and to correct potential errors in the automatic tagging provided by TreeTagger. It should be noted that some POS-tags are restricted to very few expressions, especially pronouns, noun phrases and adjectives, which should be kept in mind during analyses of DM diversity (e.g. Section 5.2.2.1).

Secondly, another functional feature was added to provide an even more generic filter into the functions of DMs, by means of a binary variable called “type” of DMs. As mentioned before, the definition of the DM category adopted here includes connecting devices that signal a discourse relation as well as items functioning on other pragmatic levels such as text structuring, metadiscourse or interpersonal management. This distinction between relational and non-relational functions is rarely tackled explicitly, with the notable exception of Degand & Simon-Vandenberg (2011) who address this issue in terms of a scale between two extremes, “non-relational” and “strictly relational”, the former showing no linking function but rather (inter-)subjective purposes such as monitoring (typically expressed by *you know*) while the latter are “grammatical items in the traditional sense of the term”, i.e. conjunctive (2011: 289). In *DisFrEn*, this scale is simplified and turned into two opposite levels, with no discrimination regarding the type of segments modified by the DM (implicit or explicit, single or complex). Connectivity is understood here in a broad sense, “not limited to relations between neighbouring utterances” (Hansen 2006: 25) and was assigned with a bias towards the relational type during the annotation. As a consequence, the non-relational value was restricted to DMs unequivocally taking scope over one unit only, thereby removing from the final protocol the option of an “in-between” value which would be rare and poorly informative. However, simultaneous functions with different scopes (i.e. one relational and one non-relational meaning expressed by a single DM) are labeled as “both” in order to account for the functional flexibility and complexity of spoken DMs (cf. Section 3.2.1).

Finally, 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 addition, the actual combination of items, for instance “and so actually”, will also appear in the annotation, making it easier to retrieve these sequences for future analysis. The values for this variable are: “Yleft” (co-occurrence at the left of the DM); “Yright: xxx” (co-occurrence at the right where “xxx” stands for the sequence of DMs in context); Ylr (co-occurrence at both left and right); “NO” (no co-occurrence).

Many examples are added in the guidelines for reference during the annotation of all the variables presented so far and can be found in Appendix 1. However, given the broad scope of this annotation protocol, it was not possible to detail the range of all possible meanings for each DM, provided such a listing is even possible or advisable given the changing nature of language in use (see Appendices 3 and 4 for the full list of annotated DMs and functions found in *DisFrEn*). Table 4.2 lists all the annotation tiers of DM-level variables as they are used in this thesis, with their definition and number of possible values.

Table 4.2: Overview of the annotation tiers specified by the DM-level protocol

Tag	Tier definition	Nbr of values
DM	full-word orthographic transcription of the DM	open
POS	source grammatical class of the DM	9
TYPE DM	position of the DM on the scale of relationality	2
DOMAIN 1	functional domain of the DM	4
FUNCTION 1	specifies the function of the DM	30
DOMAIN 2	possible second domain of the DM	4
FUNCTION 2	specifies the possible second function of the DM	30
POSITION macro	macro-syntactic position of the DM	7
POSITION micro	micro-syntactic position of the DM	5
POSITION turn	position of the DM in the turn of speech	4
CO-OCC	whether the DM co-occurs with another and where	4

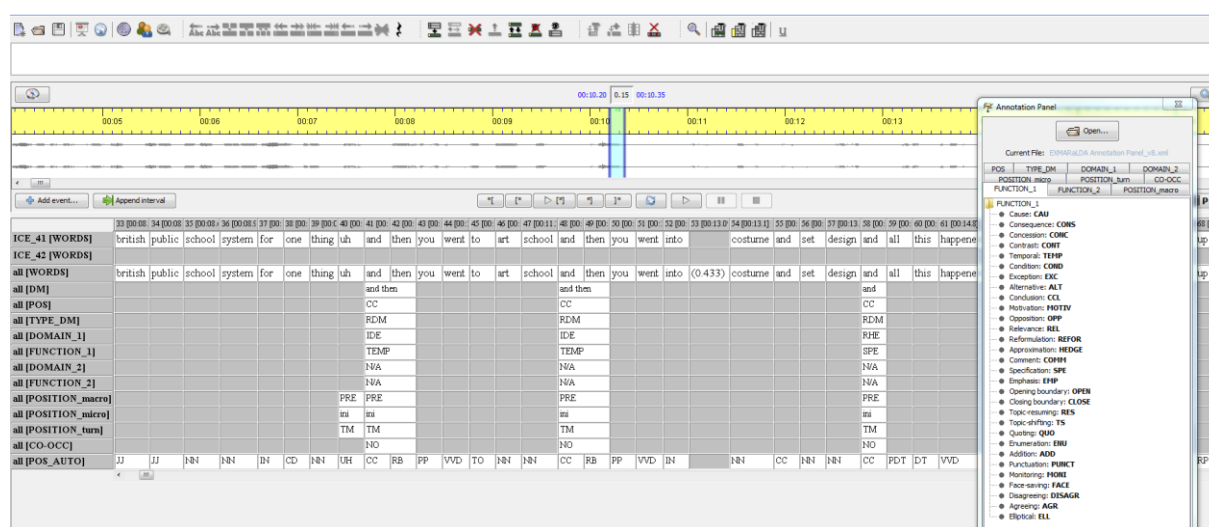
In conclusion for this protocol, the fact that many authors have tried to describe DMs illustrates the discrepancy between the complex mechanisms responsible for language production and the rigidity of corpus annotation. The present proposal hopefully contributes to this issue by striving to respect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

4.2.2 Annotation procedure

4.2.2.1 Technical aspects

As mentioned before, 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 4.10, where the DM “and then” is treated as one unit (see “complex” DMs above in section 4.2.1.1).

Figure 4.10: Partitur Editor annotation interface



Each variable presented above corresponds to one tier, with the automatic output of TreeTagger in the bottom line. All DM tokens are repeated in the [DM] tier, for practical purposes related to easy extraction of annotations, and also to standardize the transcription (original *cos* in ICE-GB is spelled *because* in this tier, elided *alors qu'* becomes the full form *alors que*, etc.).

On the right side of the screenshot, we can see the annotation panel (otherwise called controlled vocabulary), a list of all the values covered by the protocol and manually encoded as an XML file, which prevents spelling mistakes by allowing the annotator to automatically assign a value to its cell by clicking on it. A short definition of each value can be added at the bottom, thus facilitating the memorization of the taxonomy and the other variables in the protocol.

4.2.2.2 Disambiguation method

The annotation protocol for DMs specifies a number of conventions regarding the disambiguation procedure. Technically, first, annotation tiers should not be annotated in a specific order, although out of habit and convenience the vertical order presented in the EXMARaLDA files was usually followed. The annotation was only carried out once for the

whole corpus: although I acknowledge the methodological benefits of second rounds (to check for intra-coder consistency) and gold standards with other annotators (see Crible & Degand under review), the time cost of the present annotation (in addition to the required expertise) did not make it possible to conduct a second annotation round. Even though biases and decisions are documented in the protocol and have been followed as strictly as possible during the annotation, this limitation should be kept in mind in the remainder of the thesis (see next section for scores of intra-annotator agreement computed on a sample of the corpus).

For sense disambiguation, the resort to the audio file was systematic and necessary, in line with the finding in Bolly & Crible (2015, forthc.) that it significantly improves the accuracy of the annotation (see also Zufferey & Popescu-Belis 2004). There is no restriction in the number of times that the recording can be played nor the duration of the excerpt (before and after the annotated item), since some contexts can be more ambiguous than others and some DMs relate larger chunks of linguistic context than others. In addition, any disambiguation technique can be 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.

To keep track of the relative difficulty of this functional disambiguation task, a score of complexity from 1 to 3 was assigned to each DM during the annotation process. It is entirely subjective and reports on the annotator's perception of the ambiguity of the item, roughly following these instructions: 1 means no hesitation in the functional value(s); 2 corresponds to moderate ambiguity, often when two values are assigned; 3 marks cases where the annotator is not entirely satisfied with the assigned values and may involve either a completely bleached meaning or a failure of the taxonomy to account for the particular meaning in context (these cases are rare). This score of complexity was added for meta-analytic purposes (see next section) and is also involved as an independent variable in several analyses in Chapter 5.

Finally, 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 (in which case the domain and type would still be assigned twice so as to keep the number of functions equal to the number of domains and types assigned) 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 will not be 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).

Following this protocol and procedure, a total of 8,743 DMs were identified and annotated in *DisFrEn*, which amounts to about 100,000 manual annotations for this first level of analysis only. All the results in this thesis and related publications (e.g. Crible et al. 2017a; Crible under review) are the outcome of this methodology and need to be considered with the limitations discussed in this section. Some other possible applications of this dataset include any case-study approach to specific levels of the variables in the protocol (e.g. extraction of all adverbial DMs, all *topic-shift* DMs or all DMs in final position of a turn), lexicological investigations of the *n-grams* in which DMs are inserted (using the POS-tagging by

TreeTagger) and in particular strings of adjacent DMs expressing different degrees or types of co-occurrence (cf. Cuenca & Marín 2009).

4.2.2.3 Replicability of the functional taxonomy

In Crible & Degand (under review), we conducted an annotation experiment applying the functional taxonomy for DMs 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 ($\kappa = 0.406$, 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; van Enschoot et al. under review).

Turning to intra-annotator agreement, we can expect the scores to be higher given that a single annotator should resort to the same biases across different annotation rounds. 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) was annotated a second time in the same EXMARaLDA interface where all annotation tiers had been hidden beforehand, leaving only the DMs already identified in order to avoid disagreements due to item selection. The sample contained 1,194 instances of DMs (i.e. again about 15% of the whole dataset). The kappa-scores were computed on a simplified version of the dataset, that is, where partial agreement on double tags (e.g. “SEQ-RHE” at round 1 vs. “SEQ” at round 2) counts as agreement. For functional domains, the agreement is substantial ($\kappa = 0.779$, 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 ($\kappa = 0.74$, 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. DMs to which the intermediary degree of complexity (i.e. 2) was assigned during the first annotation round were found to be frequently involved in disagreements (31 out of 63 cases), while degree 1 is not problematic in a large majority (914 agreements vs. 195 disagreements) and degree 3 is too rare to be analyzed. No difference could be observed between the agreement in English and French texts.

Although the results from the inter- and especially intra-annotator agreement experiments are promising, this annotation model clearly favors precision over replicability: the original proposal as used in *DisFrEn* paved the way for a revised version (Crible & Degand under review) where a number of theoretical and practical revisions have been implemented, to be used in future research projects. 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.

4.3 Annotation protocol at fluenceme-level

The flourishing literature on (dis)fluency results in a panel of annotation protocols, albeit rarely comparable or generalizable to data of a different type. While a componential approach to (dis)fluency is generally shared amongst recent authors (e.g. Shriberg 1994; Götz 2013; Moniz 2013), the scope and format 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, see Section 2.2.3 for a detailed review).

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 the project members A. Dumont, I. Grosman and I. Notarrigo) and benefits from the input of various frameworks, thus overcoming methodological and theoretical monism. While a complete review of available protocols is beyond the scope of this chapter, the final decisions regarding the annotation of fluencemes will be discussed at length. In this section, all fluencemes covered by the protocol will be presented with their criteria and annotation procedure, starting with some governing principles that constitute the main innovations of this proposal.

4.3.1 Internal structure of fluencemes

Categories of (dis)fluent phenomena have been extensively studied in the past twenty years, including in corpus-based studies, so much so that 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 specificity of our protocol therefore lies 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 fluencemes³⁰), all annotations are assigned at word-level, with tags for every graphic unit categorized as fluenceme. This systematic segmentation avoids lexicological issues of defining words or phrases but relies on objective word boundaries for better operationality, with the exception of complex DMs as defined above (see Section 4.2.1.1). This is made possible by the tagging system with angle brackets, letters and numbers, indicating in a simple yet informative way the type of fluenceme, its position and internal structure. Concretely, each annotated word receives 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. In Example (11), we can see all these different

³⁰ Cf. Section 2.3.2 for a discussion of the terminological use of “sequence”.

configurations, starting with two simple fluencemes comprising one unit only (a DM “<DM>” and a false-start “<FS>”), followed by a complex DM, so a simple fluenceme made up of several elements (“<DM DM>”) and finally a compound fluenceme with numbers specifying the different parts (“<RI0 RI1>”).

(11)

77 [00:1	78 [00:	79 [00:	80 [00:	81 [00:	82 [00:1	83 [00:	84 [00:	85 [00:19.2	86 [00:	87 [00:20	88 [00:20
well	it	's	the	you	know	I	I	became	a	movie	maker
<DM>			<FS>	<DM	DM>	<RI0	RI1>				
DM			FS+DM+RI								

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. In the most complex cases, when a sequence of fluencemes comprises several compound fluencemes, the interpretation of the whole sequence is structured by the identification of the most overarching fluenceme that best accounts for the structure of its parts. In (12), we see how several successive repetitions can share some of their elements, here the second “I” with two tags “RI1><RI0” respectively closing the previous repetition and opening the next one.

(12)

1545 [1546 [07:01	1547 [07:01	1548 [1549 [07:01	1550 [1551 [1552 [07:02.	1553 [1554 [1555 [07:
I	(0.353)	I	miss	(0.580)	I	miss	(0.340)	her	you	know
<RI0	<UP>	RI1><RI0	RI0	<UP>	RI1	RI1>	<UP>		<DM	DM>
	<WI>			<WI>						
RI+UP									DM	

More examples are provided below in this section, and in the guidelines (Crible et al. 2016, Appendix 2). The guidelines also specify in detail the criteria for all fluencemes in different contexts, taking up problematic examples found while testing this protocol on different corpora (French and English, native and nonnative speakers, speech and sign language) by four annotators. The applicability of this protocol to multimodal data vouches for the operability and cognitive soundness of the categories identified and prevents language-specific preferences. In the following sections, I will present the categories that have been annotated in *DisFrEn* and the criteria chosen, following the collaborative protocol in Crible et al. (2016).

4.3.2 Simple fluencemes

Simple fluencemes are only composed of 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.

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, leaving out more thorough prosodic analyses.

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.

Discourse markers (“DM”) are identified following the definition and the criteria detailed in Section 4.2.1.1.

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 (see Appendix 2 for more examples). 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.

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. 2013a). If any lemma is repeated (even modified) in the next segment, it is categorized as a repetition and/or substitution (see below section 4.3.3). False-starts are tagged at the last word of the incomplete segment.

Finally, truncations (“TR”) are interruptions that only apply to words and not segments as in false-starts. 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. When a truncation is repeated (directly or after an insertion or some other fluenceme), it is numbered as in (13).

(13)

1864 [08]	1865 [08]	1866 [08:07]	1867 [08]	1868 [08]	1869 [08:08]	1870 [08]	1871 [08]	1872 [08]	1873 [08:08]	1874 [08]	1875 [08]
and	uh	(0.610)	the	uh	(0.940)	th	th	th	the	uh	the
<DM>	<FP>	<UP>	<RI0	<FP>	<UP>	<TR0	TR1	TR2	TR>RI1	<FP>	RI2>
				<WI>	<WI>					<WI>	
DM+FP+UP+RI+TR											

4.3.3 Compound fluencemes

Compound fluencemes function with a structure in at least two parts, namely the *reparandum* and the *reparans*. These categories do not exclude other simultaneous annotations, namely in

the case where a discourse marker is truncated or repeated. The numbering system mentioned before gives two types of information: identical numbers correspond to words from the same part of the fluenceme (either *reparandum* or *reparans*) while increasing numbers represent the number of times the segment has been repeated and/or substituted (see below for examples). Compound fluencemes include two types of repetitions and two types of substitutions (as well as completed truncations).

Identical repetitions (“RI”) include any words formally similar to each other and directly contiguous, whether intentionally (e.g. because of an overlap) or not, so that we avoid any judgment as to their function and relative fluency at this stage. 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*). Members of identical repetitions can only be separated by non-propositional elements, i.e. pauses, DMs, incomplete truncations and parenthetical insertions (see below). Repetitions can only apply to complete lexical elements, excluding truncated words and filled pauses.

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 (14). 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.

(14)

1495 [06:41]	1496 [06]	1497 [06]	1498 [06]	1499 [06]	1500 [06]	1501 [06]	1502 [06]	1503 [06]	1504 [06]	1505 [06:43]	1506 [06]	1507 [06:43]
(0.350)	from	the	coach	from	the	from	the	tour	c	(0.080)	tour	coach
<UP>	<RM0	RM0	RM0	RM1	RM1	RM2	RM2	<IL>	<TR		<IL>	TR>RM2>
UP+RM+IL+TR												

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.

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 (15), while the segment next to a FS has no syntactic connection with the previous one.

(15)

1306 [05:55..]	1307 [05]	1308 [05]	1309 [05:55]	1310 [05]	1311 [05:55]	1312 [05]	1313 [05:55]
anything	that	will	(0.200)	could	possibly	go	wrong
RM1>		<SP0	<UP>	SP1>			
			<WI>				
		SP+UP					

To determine the end or right boundary of a substitution, we rely on a minimal word-to-word mapping when possible, otherwise on semantic criteria, stopping as soon as a semantic equivalence can be found between the different segments, as in (16), where “is truly Liverpoolian” is replaced by “has a Cheshire accent”.

(16)

583 [02:35.6]	584 [02:35.6]	585 [02:35.6]	586 [02:35.6]	587 [02:36.5]	588 [02:37.1]	589 [02:37.1]	590 [02:37.7]	591 [02:37.7]	592 [02:37.7]	593 [02:38.1]	594 [02:38.5]	595 [02:39.1]
somebody	who	's	truly	Liverpudlian	(0.330)	and	somebody	who	has	a	Cheshire	accent
<RMO	RMO	<SP0	SP0	SP0	<UP>	<IL>	RM1	RM1>	SP1	SP1<RMO	SP1<SP0	SP1>RMO
					<WI>							

P-IL-R]

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 what is traditionally included in other frameworks, with no additional complexity for the annotators. The practical and theoretical 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.

4.3.4 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 (17). Multi-word insertions are tagged on every word. They can only appear within the *reparans* or between the two parts of a compound fluenceme, and never at the end of it.

(17)

2217 [10:00.8]	2218 [09:59.3]	2219 [10:00.8]	2220 [10:00.8]	2221 [10:00.8]	2222 [10:00.8]	2223 [10:00.8]	2224 [10:00.8]	2225 [10:00.8]	2226 [10:00.8]	2227 [10:00.8]	2228 [10:00.8]
the	monitors	go	off	wh	even	when	we	put	our	hands	in
				<TR	<IL>	TR><DM>					
				TR+IL							

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. They are not syntactically integrated, as in (18), which is the main difference with lexical insertions, another difference being their secondary informational status. Since parenthetical insertions do not directly modify the content of the repeated words, the latter are considered identical repetitions.

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.

4.3.5 Replicability of the fluenceme-level annotation protocol

The level of precision and variety of features covered by this protocol might raise concerns about the reliability of the annotation procedure. To evaluate how replicable it is, inter-annotator agreement was computed on a sample of about 7,000 words of French radio interviews which was annotated independently by myself and another (expert) annotator following the same guidelines.³¹ Pauses were excluded from the analysis on the grounds that they were identified according to different thresholds and segmentation methods depending on the research agenda of each annotator. Overall, 2,241 words were tagged as (part of) a fluenceme by at least one annotator, which amounts to 32.52% of the data. At the most inclusive level, we reached an agreement of $\kappa = 0.67$ which includes disagreements on boundaries and identification (e.g. one annotator identifies a fluenceme while the other does not) in addition to disagreements on fluenceme types (e.g. one annotator identifies a discourse marker while the other annotates the same word as an explicit editing term). 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, and that many of the assigned labels were very rare (mostly cases of double tagging such as “<RI<FS”), we decided to remove from the analysis the labels which had fewer than 10 occurrences. The dataset was thus reduced from 34 categories to 13, while the deleted labels were transformed into “no annotation” (symbol _ in the confusion matrix reported in Table 4.3). 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 operability of the guidelines.³² Comparison with other annotation frameworks assessing inter-annotator agreement is limited by the differences in granularity and category types, in addition to the sometimes opaque reports. Still, Besser & Alexandersson (2007) show a better agreement of $\kappa = 0.93$ (not restricted to true positives) with four annotators on 792 “segments” (yet relative scores show agreement higher than 70% only for four categories out of 15 in their scheme). Hough et al. (2015) found kappa-scores ranging from 0.94 to 0.99 with three annotators (two non-experts), but they used a more coarse-grained mark-up system which only identifies structural boundaries (“reparandum”, “repair”, “fillers”) and not specific categories. Their scores therefore rely on agreement for binary categories: word being part of the reparandum or

³¹ I wish to thank my colleague Iulia Grosman, who was the second annotator and who carried out the statistical analysis reported in this section.

³² The full scale presents the following interpretations: values ≤ 0 indicate no agreement; 0.01–0.20 none to slight; 0.21–0.40 “fair”; 0.41–0.60 “moderate”; 0.61–0.80 “substantial”; 0.81–1.00 “almost perfect” agreement (McHugh 2012).

not; word being part of the repair or not; word being part of the editing phase or not. All in all, it appears that our protocol is at least comparable to other annotation endeavors in terms of replicability, especially considering its fine granularity.

Table 4.3: Confusion matrix for the inter-annotator agreement on the reduced dataset

	–	D	DM	E	FP	F	I	I	RI	RIS	R	S	SP	T	Tot
–	471	18	5	7	6	10	1	3	13	0	98	17	10	14	534
DM	52	41	9	15	0	0	0	1	1	0	0	0	1	0	489
DM	0	0	13	0	0	0	0	0	8	0	0	0	0	0	21
FP	1	1	0	0	13	0	0	0	0	0	0	0	0	0	141
FS	9	0	0	0	1	10	0	0	1	0	2	7	4	1	35
IL	10	2	0	0	0	0	2	0	1	0	0	0	4	0	42
IP	0	2	0	0	0	0	1	1	0	0	0	0	0	0	22
RI	15	3	13	0	0	1	0	0	38	5	2	6	3	1	437
RM	26	0	0	0	0	0	0	0	57	1	10	8	2	4	201
SM	2	0	1	0	0	0	0	0	0	1	0	36	2	0	42
SP	11	0	0	0	0	0	0	0	0	1	0	21	17	1	51
TR	6	0	0	0	1	2	0	0	2	0	1	0	0	52	64
Total	484	60	41	22	14	23	4	5	58	8	20	95	13	73	689

Closer analysis of the confusion matrix (Table 4.3) reveals interesting sources of disagreement. First and foremost, discourse markers (DM) strike as relatively problematic to identify, with many cases tagged as “no annotation” (and some as explicit editing terms) by one of the annotators, as could be expected from the complexity of the DM category. Words which were “wrongly” categorized as discourse markers by the second annotator were, for instance, conceptual adverbs such as *clairement* ‘clearly’ and *maintenant* ‘now’ or intra-clausal uses of conjunctions, while DMs such as *c’est-à-dire* were instead labeled as explicit editing terms by the same annotator, which is yet another testimony of the need for heavy training and expertise on DMs. Another problematic category is that of modified repetitions (RM), where disagreements often involve a lexical insertion or a truncation, in which case the second annotator labeled them as identical repetitions contrary to the instruction in the guidelines (see above). On the other hand, filled pauses (FP), parenthetical insertions (IP), identical repetitions (RI), morphosyntactic substitutions (SM) and truncations (TR) all show a relative agreement above 80% (taking my annotations as reference, in the vertical column), which vouches for their operational definition. Overall, 86% of all annotations were agreed upon by the two annotators (5,927 out of the total 6,892, which corresponds to all the circled cells), including agreement on “no annotation”. The very good scores presented in this section attest to the replicability of the fluency-level protocol and allow us to interpret the results in this thesis with satisfying confidence in their objectivity.

4.3.6 Scope of the annotation in *DisFrEn*

The present thesis investigates the formal and functional behavior of DMs as one type of fluenceme and how they combine with other fluencemes (such as pauses or repetitions) across languages and registers. As such, it does not target all types of sequences equally but rather focuses on DMs and their immediate linguistic context. In addition to this theoretical motivation, the size of the corpus, as well as the number and relative complexity of the variables in the two annotation protocols presented in this chapter, compel me to restrict the scope of the annotation in *DisFrEn*. These restrictions only concern the application to some subcorpora of the fluenceme-level protocol presented in this section, while DM-level annotations were entirely produced throughout the whole corpus. Concretely, 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.

This restriction builds on our common definition of a sequence as a string of strictly adjacent words identified as fluencemes. Any item – different from the categories presented above – inserted between two different fluencemes creates a sequence boundary, except if this item occurs within a compound fluenceme, as in (22), where “sometimes” cannot be categorized as a phenomenon covered by the coding scheme but occurs within a modified repetition, thus not breaking the sequence unit.

(22)

2864 [13:47]	2865 [1]	2866 [13:47]	2867 [13:48]	2868 [13:48.5]	2869 [1]	2870 [13:48]	2871 [13:48]	2872 [1]	2873 [13:49.9]
(0.340)	the	physical	support	sometimes	the	mental	support	of	somebody
<UP>	<RM0	<SP0	RM0		RM1	SP1>	RM1>		
UP+RM+SP									

These exceptions are very rare and usually correspond to words (other than fluencemes) in the *editing term* position of a compound fluenceme, that is, between the first (*reparandum*) and second (*reparans*) parts of a sequence. Example (23), however, shows occurrences of simple fluencemes which are not affected by this exception of non-annotated words: although the DM “and then” and the following silent pause are only separated by one word (“suddenly”), they constitute two sequences of simple fluencemes; the filled pause “uhm” and the DM “as” are directly contiguous and thus form one sequence.

(23)

148 [0]	149 [0]	150 [00:45.8]	151 [00:46]	152 [00:46.1]	153 [00:46]	154 [0]	155 [00:46.9]	156 [00:46]	157 [0]	158 [00:48]	159 [0]	160 [0]	161 [0]	162 [0]	163 [0]
and	then	suddenly	(0.273)	everything	seemed	to	disintegrate	which	is	(0.180)	the	plot	uhm	as	the
<DM	DM>		<UP>							<UP>			<FP>	<DM>	
DM			UP						UP			FP+DM			

Boundaries of sequences are encoded by merging all the cells belonging to the sequence in a separate tier, the bottom one in the examples. All the fluencemes under the span of a merged cell in the bottom tier are considered to belong to the same sequence. In Example (22) above, the sequence starts with the silent pause and ends with “support”. Inside this merged cell, a

summary of the content of the sequence is manually encoded: each type of fluenceme is written once, separated by a “+” sign, in their syntagmatic order. In (24), the tag for silent pauses is written only once, even though there is a second pause at the end of the sequence. The internal structure of each fluenceme is not indicated in this summary (number of words in the fluenceme, number of repetitions, etc.). This system constitutes a first filter into the numerous patterns of combination of fluencemes, and it is completed by other “macro-labels” which will be detailed in Section 4.4.2.

(24)

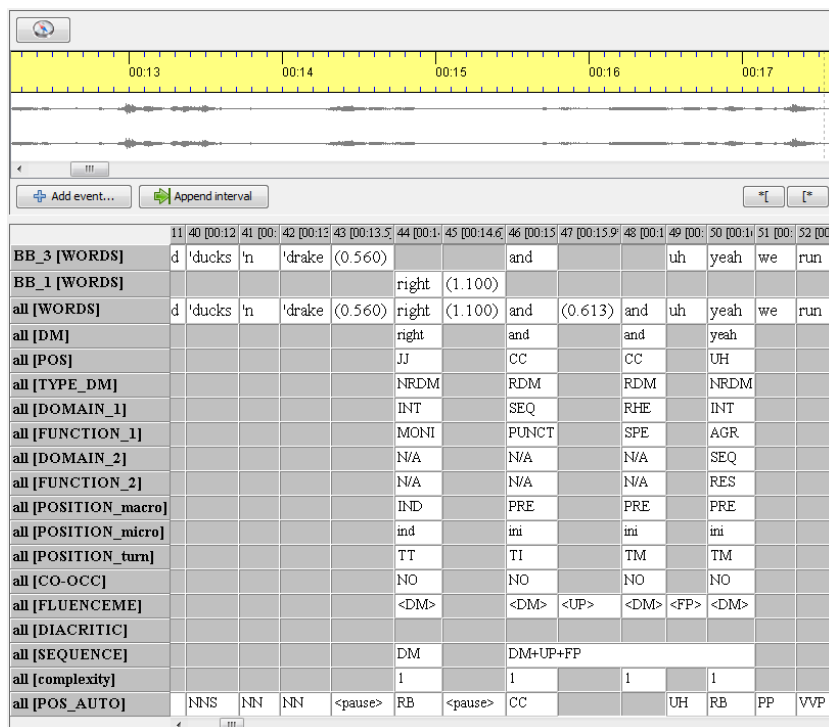
(0.540)	so	that	urn	(0.670)
<UP>	<DM	DM>	<FP>	<UP>
UP+DM+FP				

Examples (22) and (23) come from the subcorpora of radio and face-to-face interviews, which have been fully annotated at fluenceme-level, i.e. including “non-DM sequences”. In these two fully annotated subcorpora, 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 (see section 4.2.1.3), with some modifications inspired by Degand & Gilquin (2013):

- initial and final positions are no longer distinguished, since they all correspond to boundaries: pre-front field and post-field are tagged “PRE”, initial and final micro-positions are “ini”; the distinction between the left- and right-integrated position in the macro-syntax is maintained;
- silent pauses are never tagged as turn-final or turn-initial;
- filled pauses can be turn-final or turn-initial if they are strictly the first or last uttered element in the turn. DMs can still be tagged as initial or final even when filled pauses occur respectively before or after them;
- micro-syntactic boundaries are understood in a less restrictive way than for DMs, including relative and infinitive clauses;
- any interruption that leads to the beginning of a new unit (either because of a false-start, an external intervention or for any other reason) is considered as a boundary, regardless of where the interruption occurs in the first segment.

This syntactic information will be useful for the future analysis of these fluencemes, which are very frequent in speech and which frequently co-occur with DMs (see Section 6.3). An example of its application can be found in Crible et al. (2017a).

In *DisFrEn*, a total of 10,862 sequences of fluencemes have been annotated, out of which 7,244 contain at least one DM. The remaining 3,618 cases were therefore identified in the two subcorpora of interviews only. The coding interface in Partitur Editor contains fifteen annotation tiers (plus the automatic output of TreeTagger) and can be seen in Figure 4.11.

Figure 4.11: Annotation interface in Partitur Editor with both protocols in *DisFrEn*

4.4 Data extraction and post-treatment

The annotation protocols presented in the previous sections were designed in the perspective of efficient quantitative treatment with as few technical manipulations as possible. Yet a number of post-treatment steps were necessary to turn the set of annotated files into a structured dataset readily available for analysis. In this section I will present the extraction method and the transformations that were applied to the dataset in its final enriched form.

4.4.1 EXAKT concordancer by analytical level

The EXMARaLDA suite contains its own extraction and concordancer tool, EXAKT, to query the corpus. First, all .exb files from the annotation interface have to be converted into a .coma file readable by EXAKT, and this conversion requires the COMA tool, which binds all annotated files into one corpus. This .coma file can then be queried by annotation tier.

In *DisFrEn*, two types of units are investigated: DMs – small units (the longest in the corpus are four-word units, e.g. *on the other hand*) with many variables – and sequences of fluencemes, of very different sizes (from a unique fluenceme up to 43 tagged words in the corpus) and containing other annotations, i.e. the fluencemes and diacritics tags. The only way to return every variable with its associated unit is to extract DMs and sequences separately as two different analytical levels. These two units correspond to the two levels of analysis presented above, namely “DM level” and “sequence level”.³³

³³ “Sequence level” refers to the unit under consideration for the analysis, while “fluenceme level”, as used earlier in this chapter, refers to the annotation protocol and the word-level tagging method.

Firstly, all DMs are queried with their syntactic and functional variables appearing in separate columns, with an eighty-character context to the left and to the right of the item as well as the name of the file where it was found (Figure 4.12).

Figure 4.12: EXAKT annotation extraction interface at DM level

Left Context	Match	Right Context	DM	POS	TYPE_DM	DOMAIN_1	FUNCTION_1	DOMAIN_2	FUNCTION_2	Filename*[T]
ue maintenant c'est incon même pas envie de polém...	parce que	suis au courant maintenant de tout ça fait quand même...	parce que	SC	RDM	IDF	CAU	N/A	N/A	FR-intr-03 s...
0.085) avant tout (13.604) tout simplement on peut (...)	alors	(0.424) après il y a le scandale le problème (0.670) et...	alors	RB	RDM	RHE	OPP	N/A	N/A	FR-intr-03 s...
out (13.604) tout simplement on peut (0.172) alors (...)	après	il y a le scandale le problème (0.670) et qui est de plus...	après	RB	RDM	SEQ	ENU	N/A	N/A	FR-intr-03 s...
dale le problème (0.670) et qui est de plus en plus lo...	puisque	maintenant (0.427) on sait que sa passion (0.397) et l...	puisque	SC	RDM	RHE	MOTIV	N/A	N/A	FR-intr-03 s...
puisque maintenant (0.427) on sait que sa passion (0...	et	là dessus Jean - Daniel me disait une chose marrante l...	et	CC	RDM	RHE	COMM	N/A	N/A	FR-intr-03 s...
ellement désespéré réellement souffrant (12.042) m (...)	et	il a quand même accroché (1.033) sa souffrance (0.284...	et	CC	RDM	SEQ	ADD	N/A	N/A	FR-intr-03 s...
truc à une race (0.830) au lieu de l'assumer (10.108) e...	c'est-à-dire	d'assumer un truc tragique parce que quand on est à c...	c'est-à-dire	VP	RDM	RHE	SPE	N/A	N/A	FR-intr-03 s...
er (10.108) euh c'est-à-dire d'assumer un truc tragique	parce que	quand on est à cette hauteur de sensibilité (0.703) que...	parce que	SC	RDM	RHE	MOTIV	N/A	N/A	FR-intr-03 s...
e (0.428) c'est comme Rimbaud c'est pas des vies fac...	quand	on est à cette hauteur de sensibilité (0.703) que tradui...	quand	SC	RDM	IDF	TEMP	N/A	N/A	FR-intr-03 s...
1) c'est qu'il a quand même (0.271) décomposé le mo...	hein	(12.597) ah parce que le problème (0.228) qui est terri...	hein	UH	NRDM	INT	MONI	N/A	N/A	FR-intr-03 s...
est qu'il a quand même (0.271) décomposé le morcea...	quoi	c'est à dire qu'il a (0.408) il a déplacé l'agencement e...	quoi	WP	NRDM	RHE	HEDGE	N/A	N/A	FR-intr-03 s...
u'il a (0.408) il a déplacé l'agencement euh (0.607) euh	c'est à dire qu'	il a (0.408) il a déplacé l'agencement euh (0.607) euh et	c'est à dire qu'	VP	RDM	RHE	SPE	N/A	N/A	FR-intr-03 s...
(0.408) il a déplacé l'agencement euh (0.607) euh q...	et	que donc il compte de toute évidence (0.254) et il a di...	et	CC	RDM	SEQ	ADD	N/A	N/A	FR-intr-03 s...
0.607) euh et que donc il compte de toute évidence (...)	donc	il compte de toute évidence (0.254) et il a dit une chos...	donc	CC	RDM	RHE	CCL	N/A	N/A	FR-intr-03 s...
là là x x se trouve là le drame de Céline c'est son génie	et	il a dit une chose sur Proust qui est tellement sublime ...	et	CC	RDM	SEQ	ADD	N/A	N/A	FR-intr-03 s...
serait- ce que (0.103) une phrase la comme ça il l'écrit	quoi	le génie absolu (0.383) s ne serait- ce que (0.103) une ...	quoi	WP	NRDM	SEQ	PUNCT	N/A	N/A	FR-intr-03 s...
meine ça il écrit hein en mille- neuf- cent- trente- deux	hein	en mille- neuf- cent- trente- deux donc (0.595) dans le ...	hein	UH	NRDM	INT	MONI	N/A	N/A	FR-intr-03 s...
igne en disant Proust (0.991) mi revenant lui même (0...	donc	(0.595) dans le Voyage au Bout de la Nuit il commence...	donc	CC	NRDM	SEQ	CLOSE	N/A	N/A	FR-intr-03 s...
vide (0.379) fantômes de désir partouards indécis (...)	déjà	c'est sublime mi revenant lui même (0.368) c'est perd...	déjà	RB	RDM	RHE	CONDM	N/A	N/A	FR-intr-03 s...
52) (0.267) après (0.206) il dit (0.626) fantômes de d...	bon	(1.282) parler de la cten parler de l'univers de Proust d bon	bon	JJ	NRDM	SEQ	CLOSE	N/A	N/A	FR-intr-03 s...
son opposé fantôme désir (0.269) c'est rien de plus r...	c'est à dire	il emploie un mot merveilleux avec son opposé fantôm...	c'est à dire	VP	RDM	RHE	SPE	N/A	N/A	FR-intr-03 s...
de plus réel et fantôme (0.173) c'est l'angoisse (0.216)	et	fantôme (0.173) c'est l'angoisse (0.216) et après il réu...	et	CC	RDM	IDF	CONT	N/A	N/A	FR-intr-03 s...
onne définition de Proust (5.123) tout à fait (0.2...	mais	après il réunit tout ça (0.320) partouards indécis m b...	mais	CC	RDM	SEQ	ADD	N/A	N/A	FR-intr-03 s...
uis plus beaucoup ça me parle beaucoup plus Céline (...)	alors	alors comment est- ce qu'on peut avoir aimé euh Céli...	alors	CC	NRDM	SEQ	OPEN	N/A	N/A	FR-intr-03 s...
la mondanité et des codes de la société je euh person...	alors	comment est- ce qu'on peut avoir aimé euh Céline et...	alors	RB	RDM	RHE	CCL	N/A	N/A	FR-intr-03 s...
ondanité et des codes de la société je euh personne oh ...	mais	je trouve Cé euh Proust essentiel (0.336) sur (0.996) l...	mais	CC	RDM	IDF	CONC	N/A	N/A	FR-intr-03 s...
ains et les histoires de tatas (0.293) et de pédés (0.289)	ben	d'ailleurs c'est les mêmes (0.189) on pourrait presque...	ben	UH	NRDM	SEQ	TS	N/A	N/A	FR-intr-03 s...
243) ou de clodos (6.834) ou de gens très pauvres (0...	d'ailleurs	c'est les mêmes (0.189) on pourrait presque dire que ...	d'ailleurs	RB	NRDM	RHE	EMP	N/A	N/A	FR-intr-03 s...
	et puis	il y en a un autre qui s'est branché sur des histoires de	et puis	CC	RDM	IDF	CONT	SEQ	ENU	FR-intr-03 s...
	mais	ils ont été deux ethnologues (0.163) absolus (0.170) s...	mais	CC	RDM	RHE	OPP	N/A	N/A	FR-intr-03 s...

Then all sequences are extracted, showing as the queried item all the elements in the sequence. For each sequence, the fluenceme tags, as well as all the other annotations (diacritics and DM-level variables) are concatenated in one cell only, as can be seen in Figure 4.13. We see that, in the case of a sequence containing several fluencemes, it is not possible to know from the concordancer which tag applies to which element, a limitation which is especially problematic for DMs since it prevents any quantitative treatment of the different variables at this level. This problem was solved by a manual identification code assigned in post-treatment (see section 4.4.4 below). All annotations are then exported as two Excel tables, one per analytical level.

Figure 4.13: EXAKT annotation extraction interface at sequence level

Left Context	Match	Right Context	SEQUENCE	FLUENCEME	DIACRITIC	DM	POS	Filename*[T]
l'emploie un mot merveilleux avec son opposé fantôm...	(0.269)	c'est rien de plus réel et fantôme (0.173) c'est l'angoi...	UP	<UP>				FR-intr-03 s...
son opposé fantôme désir (0.269) c'est rien de plus r...	et	fantôme (0.173) c'est l'angoisse (0.216) et après il réu...	DM	<DM>		et	CC	FR-intr-03 s...
st rien de plus réel et fantôme (0.173) c'est l'angoisse	(0.216) et	après il réunit tout ça (0.320) partouards indécis m b...	UP+DM	<UP> <DM>		et	CC	FR-intr-03 s...
173) c'est l'angoisse (0.216) et après il réunit tout ça	(0.320)	partouards indécis m bon (0.192) on pourrait quand ...	UP	<UP>				FR-intr-03 s...
zards indécis m bon (0.192) on pourrait quand même	(0.224)	que chez Proust (0.194) on tr il y a une bonne définiti...	UP	<UP>				FR-intr-03 s...
92) on pourrait quand même se dire (0.224) que chez	(0.194)	on tr il y a une bonne définition de Proust (5.123) tou...	UP	<UP>				FR-intr-03 s...
rait quand même se dire (0.224) que chez Proust (0.19	tr	il y a une bonne définition de Proust (5.123) tout à fai...	TR	<TR>				FR-intr-03 s...
il y a une bonne définition de Proust (5.123) tout à fait	(0.289) mais alors	comment est- ce qu'on peut avoir aimé euh Céline et...	UP+DM	<UP> <DM> <DM>		mais alors	CC RB	FR-intr-03 s...
(0.289) mais alors comment est- ce qu'on peut avoir	euh	Céline et Proust alors (0.673) il moi je suis pas du tou...	FP	<FP>				FR-intr-03 s...
ine et Proust alors (0.673) il moi je suis pas du tout m...	Prou moi moi je je	suis plus beaucoup ça me parle beaucoup plus Céline ...	TR+RI	<TR> <RI0 RI1> <RI0 RI1>				FR-intr-03 s...
l moi je suis pas du tout moi Prou moi moi je je suis p...	beaucoup	ça me parle beaucoup plus Céline (0.314) mais je trou...	FS	<FS>				FR-intr-03 s...
je je suis plus beaucoup ça me parle beaucoup plus C...	(0.314) mais	je trouve Cé euh Proust essentiel (0.336) sur (0.996) l...	UP+DM	<UP> <DM>		mais	CC	FR-intr-03 s...
up ça me parle beaucoup plus Céline (0.314) mais je t...	Cé euh	Proust essentiel (0.336) sur (0.996) la manière de l'an...	TR+FP	<TR> <FP>				FR-intr-03 s...
plus Céline (0.314) mais je trouve Cé euh Proust esse...	(0.336)	sur (0.996) la manière de l'analyse de la grille de la mo...	UP	<UP>				FR-intr-03 s...
(0.314) mais je trouve Cé euh Proust essentiel (0.336)	(0.996)	la manière de l'analyse de la grille de la mondanité et d	UP	<UP>				FR-intr-03 s...
e de la grille de la mondanité et des codes de la société	je euh personne	oh ben d'ailleurs c'est les mêmes (0.189) on pourrait ...	FS+FP	<FS> <FP> <FS>				FR-intr-03 s...
la mondanité et des codes de la société je euh person...	ben d'ailleurs	c'est les mêmes (0.189) on pourrait presque dire que ...	DM	<DM> <DM>		ben d'ailleurs	UH RB	FR-intr-03 s...
société je euh personne oh ben d'ailleurs c'est les mê...	(0.189)	on pourrait presque dire que c'est pas loin (0.319) ce...	UP	<UP>				FR-intr-03 s...
mêmes (0.189) on pourrait presque dire que c'est pas	(0.319)	contrairement à ce qu'on pense (0.250) il y en a un q...	UP	<UP>				FR-intr-03 s...
ue c'est pas loin (0.319) contrairement à ce qu'on pe...	(0.250)	il y en a un qui s'est branché sur les mondains et les hi...	UP	<UP>				FR-intr-03 s...
s'est branché sur les mondains et les histoires de tatas	(0.293)	il y en a un autre qui s'est branché sur les mondains et les	UP	<UP>				FR-intr-03 s...
les mondains et les histoires de tatas (0.293) et de péd...	(0.289) et puis	il y en a un autre qui s'est branché sur des histoires de	UP+DM	<UP> <DM DM>		et puis	CC	FR-intr-03 s...
a un autre qui s'est branché sur des histoires de misère	de (0.274) de de	putes en Angleterre ou de (0.243) ou de clodos (6.834...	RI+UP	<RI0 <UP> RI1 RI2>	<WI>			FR-intr-03 s...
s histoires de misère de (0.274) de de putes en Anglet...	ou de (0.243) ou de de	clodos (6.834) ou de gens très pauvres (0.237) mais il...	RI+UP	<RI0 RI0 <UP> RI1 RI1>	<WI>			FR-intr-03 s...
.274) de de putes en Angleterre ou de (0.243) ou de cl...	(6.834)	ou de gens très pauvres (0.237) mais ils ont été deux e...	UP	<UP>				FR-intr-03 s...
ou de (0.243) ou de clodos (6.834) ou de gens très pa...	(0.237) mais	ils ont été deux ethnologues (0.163) absolus (0.170) s...	UP+DM	<UP> <DM>		mais	CC	FR-intr-03 s...
ens très pauvres (0.237) mais ils ont été deux ethnol...	(0.163)	absolus (0.170) sur deux humanités (0.342) et ce qui e...	UP	<UP>				FR-intr-03 s...
eus ethnologues (0.163) absolus (0.170) sur deux hum...	(0.342) et	ce qui est étonnant chez Proust (0.348) c'est (0.784) l...	UP+DM	<UP> <DM>		et	CC	FR-intr-03 s...
deux humanités (0.342) et ce qui est étonnant chez Pr...	(0.348)	c'est (0.784) l'analy alors par moments moi il me gon...	UP	<UP>				FR-intr-03 s...
(0.342) et ce qui est étonnant chez Proust (0.348) c'e...	(0.784)	f'analy alors par moments moi il me gonfle comme to...	UP	<UP>				FR-intr-03 s...

A third unit of analysis is that of the fluencemes themselves, which can be extracted in two ways: querying all “<” signs will return the total number of annotated fluenceme tokens but without any information about the internal structure and number of words in the same fluenceme (e.g. a three-word DM counts as one fluenceme token); querying the regular expression “\$” in EXAKT will return all elements (words and pauses) that have received a tag including the internal structure of fluencemes (e.g. a three-word DM counts as three tags). This second way gives an idea of the rate of fluencemes in each text (number of words categorized as fluencemes out of the total number of words in the text) but does not specify which fluencemes belong to the same sequence, and loses sight of the great variation in terms of sequence size. This level of analysis was mainly used for the quantitative investigation of fluenceme rates (Section 6.1.1).

4.4.2 Macro-labels

Once this extraction is done, a number of modifications were made to the existing annotations, all in the perspective of filtering the many values of certain variables and summarizing the annotations in different ways. Modified variables do not replace the originals but form their own new column in the dataset. At DM level, they merely consist in the grouping of certain values:

- double values for domains and functions are merged: an item annotated as both *conclusion* “CCL” and *topic-resuming* “RES”, for example, will receive the merged tag “RES-CCL”. This merging always follows the same systematic order based on the frequency of domains in the pilot corpus, from the least frequent to the most: interpersonal, ideational, rhetorical, sequential. Functions from different domains are merged following the order of their respective domain, while functions from the same domain are ordered alphabetically. Items that were assigned one domain and/or function only are not affected by this modification.
- the variable for co-occurring DMs is simplified as a binary yes/no category, by merging all types of co-occurring cases (“Yleft”, “Ylr” and all actualizations of “Yright: xxx”).

All other modifications are related to the 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.

The first addition to the sequence-level dataset is a reduced version of the full sequence of fluenceme tags, returning automatically a compact tag using a VBA form under Excel. It keeps only the opening tags and the following two letters, thus removing all information of internal structure. For instance, the sequence annotated “<DM> <FP> <RI0 RI0 RI1 RI1>” will return <DM<FP<RI. It offers a first simplification of the data which keeps the original syntagmatic order of fluencemes and the number of fluenceme tokens (i.e. it does not compress several repetitions of the same fluenceme type into one tag only).

The next category is a reformulation of the synthesis encoded during the annotation with a change of order so that all fluenceme types appear in the same order across all sequences. Its

purpose is to reduce the variation of the content of sequences by standardizing the order of fluenceme types. Concretely, for a given sequence annotated as “RI+UP+DM” in Partitur Editor, it will be re-ordered as [DM+UP+RI] in this first macro-label. The systematic order is the following: DM, UP, FP, TR, FS, ET, RI, RM, SP, SM, IL, IP, DE. It roughly follows a cline of increasing complexity (from simple to compound fluencemes), with related phenomena (insertions and deletions) deliberately left at the end of this hierarchy and DMs at the beginning.

The third category 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”.

In a slightly different perspective, the next set of macro-labels describes the internal structure of the elements in a sequence and looks at three types of information: (i) whether the sequence contains simple or compound fluencemes; (ii) whether the sequence contains one or several fluencemes; (iii) whether the sequence containing compound fluencemes also contains simple fluencemes, and the position of the latter with respect to the former. Truncations are either considered simple or compound depending on their completion, following the annotation protocol. In the case of compound fluencemes applying to a simple fluenceme (as in the repetition of a DM: “and and”) the tag of simple fluenceme will not be taken into account. This category has 10 different values that cover any type of sequence. The values, their definition and examples are provided in Table 4.4.

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 completes the information on the co-occurrence of DMs and offers a broad filter, a first answer to the hypothesis that DMs occur more frequently in sequences than in isolation.

These macro-labels rely on conceptual categorization and therefore require manual encoding, especially given the many different types of combinations that can be covered by a single value. They are complementary to each other and very helpful to analyze distributions and tendencies (especially in relation to the usage-based notion of schemas), making the

quantitative analysis manageable in spite of the high variation of patterns of fluencemes in the data.

Table 4.4: 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>

4.4.3 Additional variables in Excel

In addition to the original annotations and the macro-labels, other variables were automatically encoded in the dataset using formulas in Excel. The first group of these variables is numeric and corresponds to some extent to certain macro-labels presented above:

- number of fluenceme types: counts how many different fluencemes are contained in a sequence, excluding repetitions of the same fluenceme and internal structures. It is based on the synthetic summary of a sequence created during the annotation, e.g. [UP+FP+DM] (3 types);
- number of fluenceme tokens: counts how many fluencemes are contained in a sequence, excluding their internal structure but accounting for repetitions of the same fluenceme. It is based on the reduced sequence (see previous section) and counts the number of opening angle brackets (“<”), e.g. <UP<FP<DM<DM (4 tokens);
- number of tags: counts how many elements are in the span of a sequence, including pauses, internal structures of complex and compound fluencemes and repeated types, in other words any element which received one or several tags (in the latter case, it is counted only once). It is based on the full sequence of fluenceme tags, e.g. the sequence “(0.580) uhm but I mean” reads <UP> <FP> <DM> <DM DM> (5 tags);
- number of DMs: counts how many DMs are contained in a sequence, not accounting for their internal structure. It is based on the full sequence of fluenceme tags, retrieving all “<DM” strings of characters.

Finally, a manual coding system is dedicated to categorizing the different configurations of DMs and pauses, both silent and filled. This focus on DMs and pauses is motivated by the high frequency of these fluencemes separately or clustered in a sequence. It indicates the type of pause(s) and their position with respect to the DM, and was assigned to each DM in the corpus. This variable has fourteen different values that summarize the most frequent configurations, leaving in a “mixed” category more complex patterns:

- “N/A”: the DM is not preceded nor followed by a pause;
- “UPL”, “UPR”: there is one Unfilled Pause at the Left/Right of the DM, and no other pause at no other position;
- “FPL”, “FPR”: there is one Filled Pause at the Left/Right of the DM, and no other pause at no other position;
- “UPB”, “FPB”: there is one Unfilled/Filled Pause at Both sides of the DM, and no other pause;
- “UFL”, “UFR”: there is an Unfilled pause followed by a Filled pause at the Left/Right of the DM, and no other pause at no other position;
- “FUL”, “FUR”: there is a Filled pause followed by an Unfilled pause at the Left/Right of the DM, and no other pause at no other position;
- “UDF”: there is an Unfilled pause, followed by a Discourse marker, followed by a Filled pause, and no other pause at no other position;
- “FDU”: there is a Filled pause, followed by a Discourse marker, followed by an Unfilled pause, and no other pause at no other position;
- “MIX”: any other type of configuration. It must include both types of pauses, with at least one type that occurs twice, either at one or both sides of the DM. Examples of “MIX” types are: <DM<FP<UP<FP, <UP<FP<UP<DM<UP.

These categories of DMs and pauses have been elaborated for the purpose of this research (Section 6.3; see also Crible et al. 2017a) and thus reflect its focus and interest in DMs over other fluencemes. In my opinion, there is no simpler system (i.e. with fewer possible values) that would summarize these configurations and account for both the types of pauses and their respective positions.

4.4.4 Identification codes

After adding these variables to the manual annotations and macro-labels, the *DisFrEn* dataset in its final form comprises twenty-seven variables in total: fifteen at DM level and twelve at sequence level. The challenge and final step of post-treatment which results from this high number of variables was to make all these values communicate with each other across the two analytical levels. As mentioned before, the EXMARaLDA concordancer was not able to return such information. A manual solution was found in the form of unique identification codes attributed to all DMs and all sequences in the corpus.

These codes have a similar systematic structure that indicates the following metadata: subcorpus (task label), language and text ID code. Each sequence or DM is then given a number which corresponds to its chronological order of appearance in the text. For instance, the sequence ID “ClaEN2t27” means that this is the twenty-seventh sequence of the second text in the English subcorpus of classroom lessons (EN-clas-02). The DMs contained in this sequence (if any) will have the sequence ID plus “DM” and their specific number, as in “ClaEN2t27DM32” (thirty-second DM in the text). DMs from the same sequence will have the same code except for the last two figures.

Although these codes need to be assigned manually, they are very useful since they make it possible to retrieve the annotations of one level (e.g. the macro-labels of a sequence) and attribute them to the items on the other level (the DM(s) in this sequence), thus connecting all variables together. The automatic inter-connection between the two levels is possible in Excel through formulas based on the identification codes. As a final example illustrating the richness of the dataset and the use of these codes, I will conclude this section with the complete description (including metadata) of a random DM in *DisFrEn* as it appears after the entire annotation and post-treatment process:

Column name	Content	Value
DM level		
DM_ID	Unique code of the DM	InfEN4t135DM47
CONTEXT_LEFT	Left context	... the website (0.620) ok
ITEM	Original item	and
CONTEXT_RIGHT	Right context	(0.060) so what is ...
DM	DM expression	and
POS	POS	CC
TYPE_DM	Type of DM	NRDM
DOMAIN_1	Domain	SEQ
FUNCTION_1	Function	TS
DOMAIN_2	2 nd domain (if any)	N/A
FUNCTION_2	2 nd function (if any)	N/A
DOMAINS	Merged domains	SEQ
FUNCTIONS	Merged functions	TS
POSITION_macro	Macro position	PRE
POSITION_micro	Micro position	ini
POSITION_turn	Position in the turn	TM
CO-OCC	Co-occurring DMs	Yright: and so
CO-OCC2	Yes/No co-occurring	YES
COMPLEXITY	Score of complexity	1
FLUENCEME	Fluenceme tag	<DM>
DIACRITIC	Diacritic tag (if any)	N/A
Sequence level		
SEQ_ID	Unique code of the sequence	“ClaEN2t1”
SEQ_FULL	Sequence span	ok (0.380) and so
FLUENCEMES	All fluenceme tags	<DM><UP><DM><DM>
REDUCED	Reduced sequence	<DM<UP<DM<DM
SEQ_SYNT	Synthetic sequence	DM+UP
SEQ_ORD	Ordered sequence	[DM+UP]
STR_ELE	Internal structure	sim.seq
SEQ_CAT	Sequence category	P

CLUSTER	Type of DM cluster	in sequence
NB_FLU_TYP	Number of fluenceme types	2
NB_FLU_TOK	Nr of opening brackets	4
NB_TAG_SEQ	Nr of tags in the sequence	4
DIACRITIC_SEQ	Diacritics in the sequence	N/A
NB_DM	Number of DMs	3
DM_PAUSE	Configurations of DM & pauses	UPL
Metadata		
LANGUAGE	Language of the text	EN
TEXT_CODE	Text code	EN-intf-04_s.exs
REGISTER	Speaking task	intf
D.ELICIT	Degree of elicitation	semi.supervised
NB.SPK	Number of speakers	dialogue
D.PREP	Degree of planning	semi.prepared
D.INTERACT	Degree of interactivity	semi.interactive
D.MEDIA	Degree of broadcasting	not.media
C.PRO	Professional or not	pro

4.5 Conclusion of the chapter

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 satisfying inter- and intra-annotator agreement;
- the extraction of annotations in two steps or levels, viz. DM level and sequence level, and the mapping of these levels with additional categories or macro-labels.

Shortcomings are mostly due to practical reasons which, as we know, often interfere with theoretical ambitions in empirical studies. Nevertheless, the numerous revisions, based on a pilot study as well as confrontations with the literature and experts in the field of discourse annotation, prevent major pitfalls and make *DisFrEn* a reliable dataset for the study of discourse markers as fluencemes.

Chapter 5: Portraying the category of discourse markers

Introduction to the chapter

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 bottom-up exploratory research and answering some of the hypotheses laid out in Section 3.4. The present analyses are mainly quantitative and frequency-based; more complex (multivariate) statistical models are introduced when they shed a complementary or synthesizing light onto the data.

Results are provided in raw and relative frequency per thousand words (henceforth ptw).³⁴ They are either computed across register-based or feature-based subcorpora (e.g. frequency in conversations vs. interviews; frequency in professional vs. non-professional settings). The use of one metadata system over another in the different sections of this chapter depends on which system is most relevant to the research question or hypothesis at stake.

Each section builds on the results from the previous one(s) according to the following structure: the overall frequency of DMs will first be compared across languages and registers without considering particular variables besides part-of-speech tags and the DM expressions themselves (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 three functional variables (type, domain, function) are introduced in Section 5.3, where they will be individually detailed and mapped onto register and positional preferences; the different configurations of co-occurring DMs are examined in Section 5.4 in combination with both syntax and functions; and finally Section 5.5 summarizes and discusses the main results of this chapter in the usage-based perspective of uncovering form-function patterns at various degrees of abstraction.

5.1 General frequency 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 could be formulated (cf. Section 3.1.3). 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 bottom-up approach to English and French DMs. Frequency of DMs by register, on the other hand, is highly documented and should show major effects of the degrees of preparation and interactivity, following hypotheses on fluencemes in general and DMs in particular: high

³⁴ 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.

frequency and variety of DMs should be associated with spontaneous discourse and interactive registers, given the role of DMs in planning processes and interpersonal strategies of exchange management. Table 5.1 reports on the distribution of DMs in the *DisFrEn* corpus.

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

In *DisFrEn*, 8,743 DMs were identified and annotated, amounting to an overall relative frequency of 54 DMs ptw. 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 $p < 0.05$ ($LL = 101.76$, $p < 0.01$).³⁵ 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 large 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 3.2.2).

At this basic stage of the analysis, the observed difference between English and French cannot be interpreted any further than a quantitative difference in frequency, which I will strive to relate to more qualitative variables later on in the chapter. However, one possible explanation regards a theoretical and methodological decision on definition and identification of DM tokens: I mentioned in Section 4.2.1.1 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 de los Angeles Gómez González (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 SDRT framework (Asher & Lascarides 2003, cf. Section 3.2.1), which was originally developed for discourse

³⁵ All log-likelihood tests in this chapter and the following ones were computed with the on-line calculator provided by Lancaster University, accessible at <http://ucrel.lancs.ac.uk/llwizard.html>.

relations and 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 are an English specificity and 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. The fact remains that the inclusion of tag questions in the DM category could have considerably affected (i.e. smoothed out or even reversed) the frequency results and quantitative difference noted above.

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, this view of DM use is rather negative and overlooks more "fluent" or strategic roles of DMs, which are attested by their not-so-rare presence in very formal registers such as news broadcasts. Functional annotations should prove highly informative in this regard by filtering the many domains of use of DMs across different registers (see Section 5.3). Furthermore, 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 intermediary 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; Sections 2.3.3 and 2.4). This effect is not necessarily a negative one, as further analyses will show in the following sections. 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 qualifies 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) in English.

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 English and French 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.

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 3 for the same table with register information). 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).

Table 5.3: Top five most frequent DMs in English and French

	English	French
(1)	<i>and</i>	<i>et</i> ‘and’
(2)	<i>but</i>	<i>mais</i> ‘but’
(3)	<i>so</i>	<i>donc</i> ‘so’
(4)	<i>well</i>	<i>alors</i> ‘so/well’
(5)	<i>you know</i>	<i>hein</i> ‘right’

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. Only a selection of register-based specificities is listed here:

- 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.
- The subordinating conjunction *if* and its French equivalent *si* 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.
- *Indeed*, *however*, *for*, *meanwhile*, French *car* ‘since’, *pour que* ‘so that’ or *en effet* ‘indeed’ are some of the DMs which 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.

These results are purely descriptive: the DMs listed above can actually cover many different uses, and such a generic level of analysis does not further our understanding of the research questions under scrutiny. All DMs found in *DisFrEn*, their frequency and annotated functions can be found in Appendix 4. So far, the hypothesis of higher frequency in spontaneous and interactive registers has been confirmed, with the qualification 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 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) annotated as detailed in Section 4.2.1.3. Figures 5.1 and 5.2 represent the proportions of POS-tags in news broadcasts and conversations, in the two languages combined. 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.

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

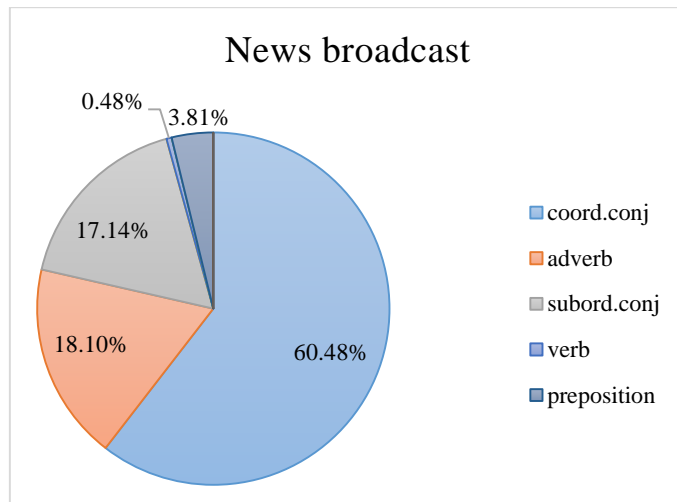
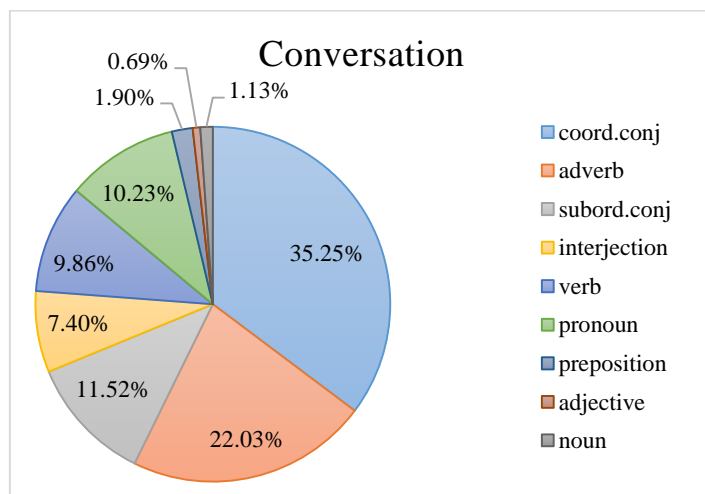


Figure 5.2: Proportions of part-of-speech tags in conversations



Most POS-tags are shared across languages and registers, although to a lesser extent in news broadcasts, political speeches and sports commentaries, pointing to an effect of broadcasting on the choice and variety of DMs. Crosslinguistically, the only notable specificity is the pronoun category, which is only possible in French: it is almost only represented by *quoi* ‘you know’ (also by *et tout ça* and other rare variations thereof) and is therefore restricted to conversational discourse.

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 Positional variables

The previous frequency results allowed us to test general hypotheses on language and register variation, beyond what case studies can provide, and confirmed the similarity between English and French as well as the prevalence of conjunctions in the DM category. Another aspect where a bottom-up approach to the category can prove enlightening is the positional behavior of DMs, in particular their supposed tendency towards utterance-initial position, as often claimed in general DM definitions. Apart from initiality, other syntactic slots (medial, final) will be carefully investigated with respect to different units (turn, dependency structure, clause) in order to uncover some formal restrictions to their use, as well as their tentative connection to fluency. No functional variable will be integrated at this stage of the analysis, so as to explore the descriptive power of formal, objective factors independently of a more qualitative and interpretative approach to the data (see Section 5.3).

5.2.1 DMs across units

As explained in the description of the methodology (Section 4.2.1.3), the position of DMs is annotated according to a tripartite system which distinguishes three reference units relevant to the behavior of DMs, namely clause (a minimal propositional unit, including subclause), dependency structure (a main clause and its constituents, roughly corresponding to an utterance) and turn (the span between two changes of speaker). The main hypothesis in this regard is that initial position should 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

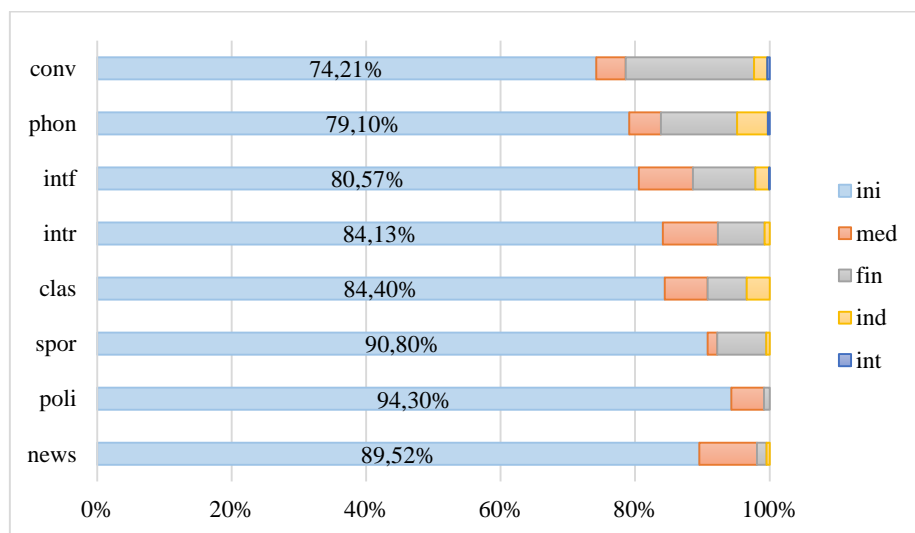
	English		French		Total	
	DMs	ptw	DMs	ptw	DMs	ptw
initial	3639	42.11	3417	45.39	7056	43.64
medial	292	3.38	223	2.96	515	3.18
final	245	2.84	730	9.70	975	6.03
independent	65	0.75	118	1.57	183	1.13
interrupted	8	0.09	6	0.08	14	0.09

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 mentioned in Section 5.1. 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. 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). Further evidence of the weight of *quoi* on this distribution is provided by Figure 5.3 representing the proportions of DM position in the clause by register. We see that the final slot takes up its largest proportion in conversation, which is precisely the register where *quoi* preferably occurs, as identified above.

Figure 5.3: Proportions of DMs in micro-position (clause-level) by register

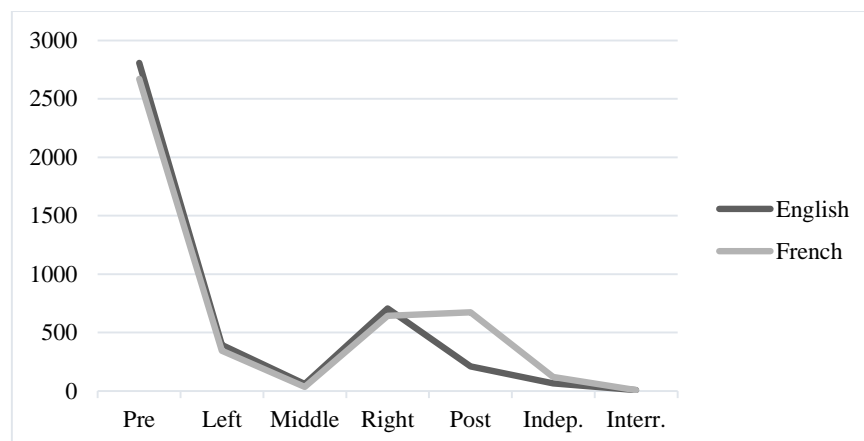


This figure also shows that the prevalence of initial position should be somewhat qualified by register variation: while initial DMs are always the majority, the proportion varies from 74% (in conversation) to 94% (in political speech), with a general cline towards larger proportions in formal (professional, broadcast) registers, which is further proof of the restrictions on DM use in these speaking tasks.

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. As a reminder, in this system, left and right positions are divided into integrated (“LEFT”, “RIGHT”) and not integrated (“PRE” for pre-field, “POST” for post-field), in addition to a middle-field position (“MID”, within the verb phrase) as well as the independent (“IND”) and interrupted values (“INT”) taken up from the micro-syntactic system (cf. Section 4.2.1.3). Figure 5.4 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. 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).

Figure 5.4: Macro-position (dependency-level) of DMs



- (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)

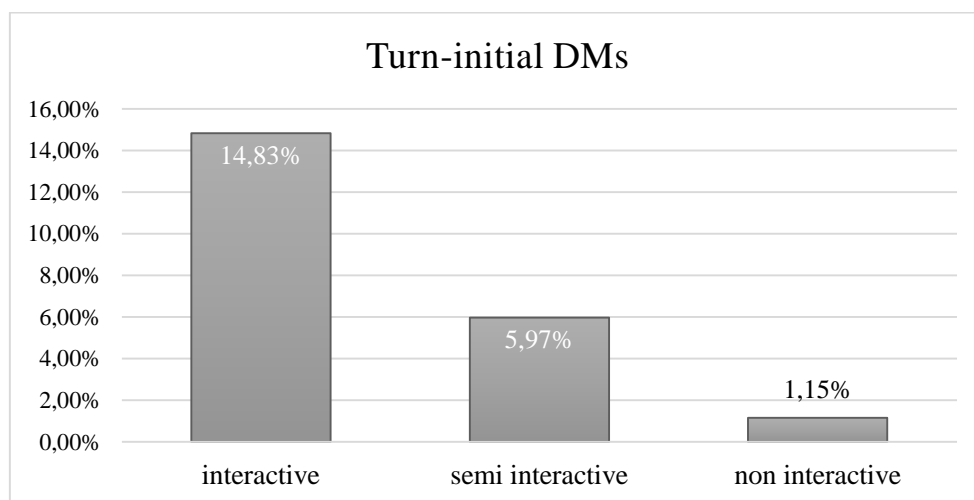
In a micro-syntactic sense, both occurrences of “so” in these examples are initial: they introduce a (sub)clause, that is, a grammatical unit expressing a proposition and containing at least a predicate and its subject (“we’ve got a whole lot of uh clergy scientists poets” and “they stretch the imagination a bit”). However, macro-syntactically, the unit introduced by “so” in (1) is a full independent utterance and the DM is not governed by the previous clause, whereas in (2), “so” subordinates the second segment to the first in a relation of syntactic dependency, here with a sense of purpose which could be substituted by *so that* (I like fantastic things so that they can stretch the imagination). *So* is one of the rare DMs which can occur both in integrated and non-integrated contexts, 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.

Register information does not strongly qualify these findings (see Appendix 5 for the distribution of macro-syntactic slots by register). The only notable effect 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). A final observation at the macro-syntactic level is crosslinguistic and suggests a higher variation of positional behavior in French than in English, where the pre-field slot takes up larger proportions across most registers (except radio and news).

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 5.5 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 5.5: 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 monological 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.³⁶

No major crosslinguistic (quantitative) difference is worth noting at turn-level: the proportions of the different slots only vary according to the general DM distribution and the higher frequency of French DMs overall, while the ranking of positions in the turn is the same in English and French (medial, initial, final, independent). The only difference 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 tend to start their turns with the more multifunctional *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 ($N = 40$). These variations might correspond to different functions (e.g. *opening boundary* vs. *topic-shift*) which will not be discussed any further here.

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 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.

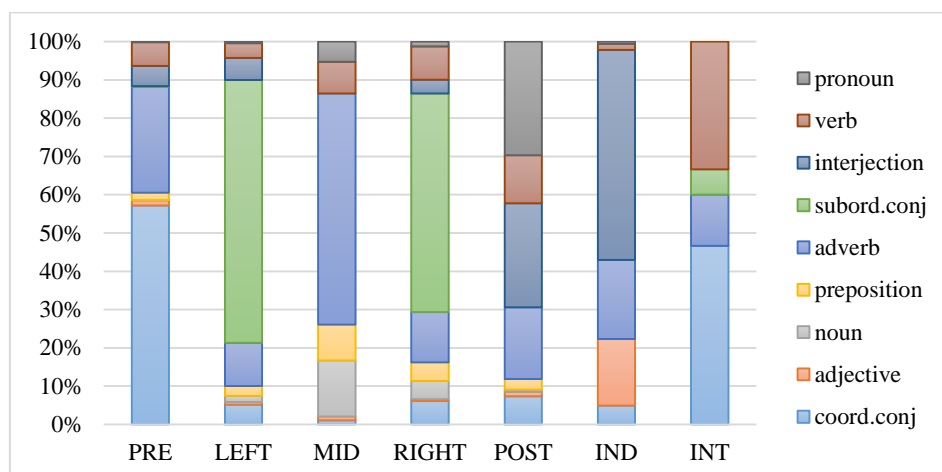
5.2.2 Formal and cognitive restrictions of less typical positions

5.2.2.1 Variation within macro-syntax: focus on post-field DMs

So far, the prevalence of initial position has been established and nuanced by the variation in language, register and unit type. While this tendency towards initiality can be easily explained by the relational or structuring nature of DMs, the remaining less typical positions (namely medial and final) could benefit from further investigation in order to uncover the specific formal restrictions to their use (see Section 5.3.4 for an integration of positional and functional variables). Medial and final 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. The following analysis is only interested in formal diversity, that is, the specific syntactic classes of DMs which are attested in each position.

Cross-tabulating macro-syntactic position with POS-tags offers a first filter into this variation, as represented in Figure 5.6. We see 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. What is more, 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 forbids to discuss them any further).

³⁶ The z-ratio is used to test the significance of the difference between two independent proportions. Z-scores and their associated p-values reported in this thesis were computed on http://vassarstats.net/propdiff_ind.html.

Figure 5.6: Proportions of POS-tags across macro-syntactic positions

Concretely, 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: this configuration is most frequent in political speeches, where it takes up 88% of all middle-field DMs, against around 50% in the other registers. Lastly, the pattern illustrated in (6) is the rarest and formally most restricted 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.

The only remaining position where no favorite POS-tag can be distinguished is the post-field slot, where the situation is more complex 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 a specificity of 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 *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, less typical positions, especially the post-field macro-syntactic slot, are 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 qualifies the difference between pre- and post-field, especially in French where the former covers more different DM types than the latter.

5.2.2.2 Variation within micro-syntax: focus on medial DMs

The micro-syntactic final position does not bring any new insights of formal diversity of DMs since it corresponds to the macro-syntactic post-field already discussed in the previous section. Medial position, however, is more interesting than the macro-syntactic middle field since its definition is less restrictive, thus potentially covering more different DMs than the majority of adverbs observed in Figure 5.6. Adverbs, which are known for their syntactic mobility, still take up a large proportion of medial DMs (41% in English and 35% in French, against 24% and 22%, respectively, in initial position). Yet, they no longer represent the majority of cases, mainly because of 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 French (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 forbids such generalization, or at least qualifies it depending on the POS-tag of the DM: adverbs, verbs and prepositions 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) there was something much more complex if you looked **for instance** at twentieth century painting it got (0.213) very very far away (EN-intr-04)

Formally, we see similarities between Examples (7) and (9), which both occur before a prepositional phrase (“for care”, “at twentieth century”), 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 (see Section 5.3.4).

Apart from these three POS-tags which are shared crosslinguistically in substantial frequencies, other types of DMs can also be found in medial position, although in much smaller proportions. The most striking specificity of this syntactic slot 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 (Example 11).

(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**!³⁷

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 qualified 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.

(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,

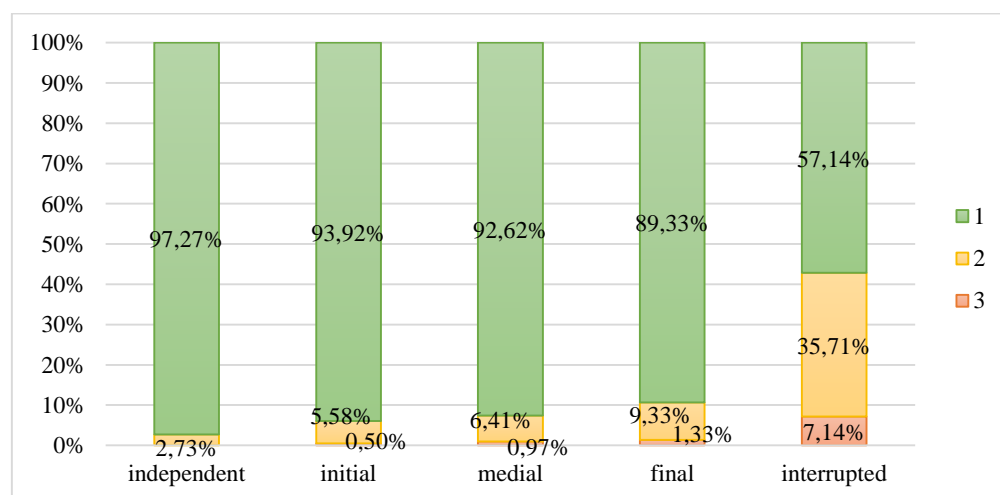
³⁷ <http://m.downatthetmac.proboards.com/thread/12655/happy-sort>

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 5.3.4). At this formal stage of the analysis, what can be asserted with high certainty is that there are some recurrent patterns and restrictions to the types of DMs which can occur in various positions of the speech string.

5.2.2.3 Position and score of complexity

A tentative interpretation of the impact of position in cognitive processing and fluency is provided by analyzing the relation between position and the score of complexity assigned during the annotation. As explained in Section 4.2.2.2, this score or degree of certainty expresses the annotator's subjective impression of ease vs. difficulty during the sense disambiguation task. Although it does not target specifically positional variables, coding complexity could be interpreted as an indirect sign of processing cost, at least for the annotator if not for the participants of the interaction as well. I would like to propose that this meta-annotation is somewhat related to the typicality of the DMs: frequent uses of DMs should become more accessible to the annotator, much like they are to a hearer, in concordance with the role of frequency in cognitive entrenchment (and fluency) as developed in Section 2.3.4. The distribution of complexity scores across the different micro-syntactic slots is reproduced in Figure 5.7.

Figure 5.7: Proportions of scores of coding complexity by micro-syntactic positions



This graph shows that the more the DM occurs to the right of the utterance, the higher the proportion of complex cases (scores 2 or 3). In other words, the proportion of complex, hence presumably atypical, DMs increases as the DM moves along the speech string, from initial to final and interrupted DMs. In particular, we see that 11% of final DMs are rated 2 or 3; final DMs also take up a quarter (24%) of all score-3 DMs and 17% of all scores 2. Interestingly, the right periphery has been identified by Levelt (1983) as a potential slot for disfluency since the speaker's attention towards their previously uttered speech is heightened towards the end of

constituents, thus leading to more interruptions or reformulations. The coding complexity of DMs in final position could therefore be considered as a first tentative cue of the link between online processing and offline annotation. Although additional proofs are required (see Section 6.6), this graph suggests that not all slots are equal in terms of processing complexity, at least for the task of annotation, which should be expected from the non-linear rhythm of spoken utterances.

This result is a first step in the methodological endeavor to establish a link between frequency and fluency. However, to date, there is no empirical evidence that annotation complexity should be related to on-line processing complexity, given the different nature of the tasks (off-line vs. on-line, situated vs. decontextualized, multimodal vs. transcript-based, etc.). An interesting research avenue would consist in experimentally testing the extent to which the types of cognitive effects observed for the processing of DMs transfer to the annotation process.

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, and the related formal restrictions in each of these contexts. 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. The first sections of this chapter thus illustrate the types of conclusions which can be drawn from a bottom-up frequentist approach to pragmatic categories. Such findings are mainly limited to confirming previous claims from the literature or coming to terms with the heterogeneity of DMs and their differences between English and French, assuming that the *DisFrEn* dataset is representative of these languages. The resulting picture of such a complex pragmatic category is therefore only partial. However, 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 was never 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 Functional variables

Similarly to the positioning system, functional variables are divided into three levels which vary in their degree of granularity, from two types (relational or non-relational, and the option of combining the two), to four domains (and their combination) and 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 (Non-)relational type

As mentioned in Chapter 3, relationality is often considered criterial to DM status, especially in writing-based taxonomies of discourse relations targeting “connectives” rather than the broader functional spectrum suggested by speech-based studies of DMs (cf. Section 3.2.1). Before I provide the corpus-based results to test the representativeness of the relational feature, a methodological precision should be noted: since the occurrences of simultaneous relational and non-relational types for a single DM are very rare in the corpus (136 DMs out of 8,743), they were merged with non-relational values, given that these cases are less frequent than relational DMs and that non-relationality is more marked than the default relational interpretation of DMs. This variable is therefore binary.³⁸ Bearing this modification in mind, the distribution of relational (RDM) and non-relational (NRDM) types by language and register is reported in Table 5.6.

Table 5.6: Relative frequency (ptw) of (non-)relational DMs by language and register

	Relational			Non-relational		
	English	French	Total	English	French	Total
conversation	34.67	48.42	41.53	19.91	38.78	29.33
phone	37.65	43.64	40.11	24.83	34.5	28.8
interview	49.66	51.38	50.54	13.02	20.62	16.92
classroom	42.44	31.96	39.47	12.41	18.53	14.15
radio	44.91	37.67	41.36	9.69	14.73	12.16
sports	37.88	32.33	35.48	2.19	6.85	4.2
political	21.73	18.53	20.21	0.58	1.66	1.09
news	12.92	15.17	14.02	0.99	1.33	1.16
Total	37.09	39.24	38.09	12.08	20.45	15.98

³⁸ Double domains and functions, however, will be treated as separate categories so as not to lose the information of DMs expressing two simultaneous types and meanings (see Section 5.3.3).

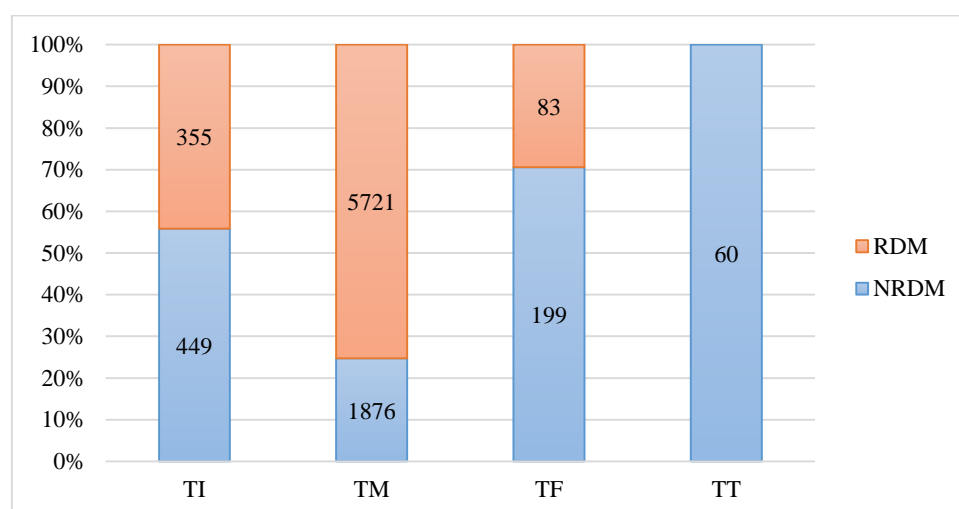
Unsurprisingly, we see that relational DMs are much more frequent overall than non-relational DMs ($LL = 1505.54$, $p < 0.001$) in English and French and across all registers. This finding reflects the centrality of the relational feature in the DM category. Nonetheless, substantial frequencies of non-relational DMs can be observed in a number of registers which are characterized as (semi-)interactive dialogues: proportions of NRDMs decrease along with the general frequency of DMs in all registers, from informal (around 40%) to intermediary (around 25%) and formal registers (around 7%).

Crosslinguistically, we see that NRDMs are almost twice as frequent in French as in English overall, although the difference is smaller in some registers. This result again evokes the high frequency of *hein* ‘right’ or *quoi* ‘you know’, which are typically non-relational. The most frequent NRDMs are in fact *well* and *you know* in English and *hein* ‘right’, *quoi* ‘you know’ and *ben* ‘well’ in French: in both languages, non-relational functions seem to be expressed by speech-specific expressions. Some NRDMs are located in final position, as in (13). Other typically relational DMs can also express non-relational meanings in initial position, as in (14).

- (13) <ICE_19> most people learn to drive by the time they’re seventeen **you know**
 <ICE_20> well not me (0.547) I mean it’s completely (EN-phon-02)
- (14) <VAL_11> moi j’ai mangé les satés et vous avez mangé euh
 <VAL_12> **mais** non mais je me demande si on n’a pas tout mangé
 <VAL_11> *I ate the sates and you ate uh*
 <VAL_12> **mais** ‘but’ no but I wonder if we didn’t eat it all (FR-conv-01)

In Example (13), “you know” takes scope over the previous utterance to establish common ground on a well-known fact (“most people learn to drive”); there is no other segment in the scope of the DM, so that the meaning is non-relational. In Example (14), “mais” does not seem to express its usual contrastive meaning but instead signals disagreement, as indicated by its co-occurrence with “non”. In this context, the DM does not so much relate the previous utterance to the next as introduce a new one, which is also a new turn after the interruption of <VAL_11>. It is particularly interesting that these two examples illustrate cases of turn-final and turn-initial DMs, respectively: non-relational meanings are to be expected when there is a change of speakers, even if it is entirely possible for a DM to signal a relation between two units from different turns and speakers.

It remains that NRDMs seem, by definition, particularly associated with the beginning and end of turns, as shown in Figure 5.8. We see that the proportions of RDMs and NRDMs are fairly similar in turn-initial position (“TI”) and in favor of NRDMs in turn-final position (“TF”), as opposed to the majority of turn-medial DMs (“TM”) which are overwhelmingly relational. DMs constituting independent turns (“TT”) stand apart with their low frequency (only 60 occurrences) and exclusively non-relational uses, which usually correspond to punctuating and/or backchanneling functions (e.g. *okay*, *right*). The high frequency of NRDMs in interactive registers can therefore be explained by this connection with turn changes, which are scarcer or even nonexistent in semi- or non-interactive contexts (cf. Section 5.2.1).

Figure 5.8: (Non-)relational type by position in the turn

Besides these positional preferences, RDMs and NRDMs share a number of expressions which can be used in either type. In fact, the eight most frequent DMs in the whole corpus can all express a relational or a non-relational meaning, but always with a preference for one use. This preference usually favors RDMs (namely for *and*, *but*, *so*, French *et* ‘and’, *mais* ‘but’, *donc* ‘so’, *alors* ‘then’) except for *well* which is predominantly non-relational. Lower in the frequency ranking, most DMs show strong preferences for one type or the other, for instance:

- typical NRDMs include *you know*, *sort of*, French *hein* ‘right’, *quoi* ‘you know’, *ben* ‘well’, *bon* ‘well’, *tu vois* ‘you see’, *voilà* ‘there’, etc.;
- typical RDMs include: *if*, *because*, *I mean*, *when*, French *parce que* ‘because’, *enfin* ‘I mean’, *quand* ‘when’, *si* ‘if’, etc.

Very few expressions show a similar frequency in each use except *now* (19 vs. 21 occurrences of RDM and NRDM, respectively) and, to a lesser extent, *actually* and *en fait* ‘in fact’. Examples (15) and (16) illustrate the two possibilities for the DM *now*.

- (15) the idea of a monopoly in Jacksonian times is something which is actually literally created and condoned by the government (0.280) **now** Jacksonianism is about breaking that up (EN-clas-02)
- (16) finding out just how much support there is in the Arab world for military action to end the crisis (0.240) **now** how much is this all part of the general psychological pressure (EN-news-08)

The DM “now” in Example (15) signals a contrastive relation between “the idea of a monopoly” and the political programme of Jacksonianism; the connection is further marked by the anaphorical “that” referring to part of or the whole previous utterance. By contrast, “now” in Example (16) does not express its original temporal meaning nor its contrastive variant but rather serves as a segmentation cue or introducer of the upcoming interrogative act. However, it should be noted that examples such as (16) or (14) are rather challenging to categorize into relational or non-relational types, especially since the DMs can be used both ways. As developed in Chapter 3, relationality should be seen as a continuum, as suggested by Degand

& Simon-Vandenberg (2011), leaving many occurrences in an in-between place on the scale. Therefore, despite interesting association patterns with formal variables such as position in the turn or particular DM expressions, I would like to conclude on a precautionary note regarding this variable, which might need further operationalization and testing, especially given the natural tendency to interpret every DM as relational by default. Bearing in mind these potential limitations of reliability, the fact remains that, once applied systematically to corpus data, the annotation of relationality allows us to confirm the centrality and representativeness of this feature in the full DM category.

5.3.2 Single domains and functions

As a reminder, the taxonomy of DM domains and their respective function values is reported below as Table 5.7. As mentioned in Chapter 4, 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.

Table 5.7: 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		

5.3.2.1 Domains

The intermediary level of granularity in functional variables is that of the domain of use, a term taken from Redeker (1990) which refers to the level of discourse targeted by the DM. In this work, I distinguish four domains, namely ideational (content, objective relations), rhetorical (speaker's attitude, subjective relations), sequential (turn exchange and topic structure) and interpersonal (speaker-hearer relationship). Domains also relate to common notions in discourse analysis and cognitive linguistics such as speaker- vs. hearer-orientation, objective vs. subjective relations or coherence vs. topic relations, distinctions which were defined earlier (Section 3.2.1) and whose relevance in discourse processing allows us to make qualitative

interpretations on the basis of more powerful statistical tools than what was used so far. Each model or method will be briefly explained when it is introduced.

Based on Denke's (2009) corpus findings (cf. Section 3.3.1), DMs are expected to attend primarily to discourse structure, in other words, the sequential domain should 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 should 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) should be strongly associated to interpersonal DMs, whereas formal registers (e.g. news broadcasts) should 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. The data is reported in Table 5.8.

Table 5.8: Distribution of single domains by language

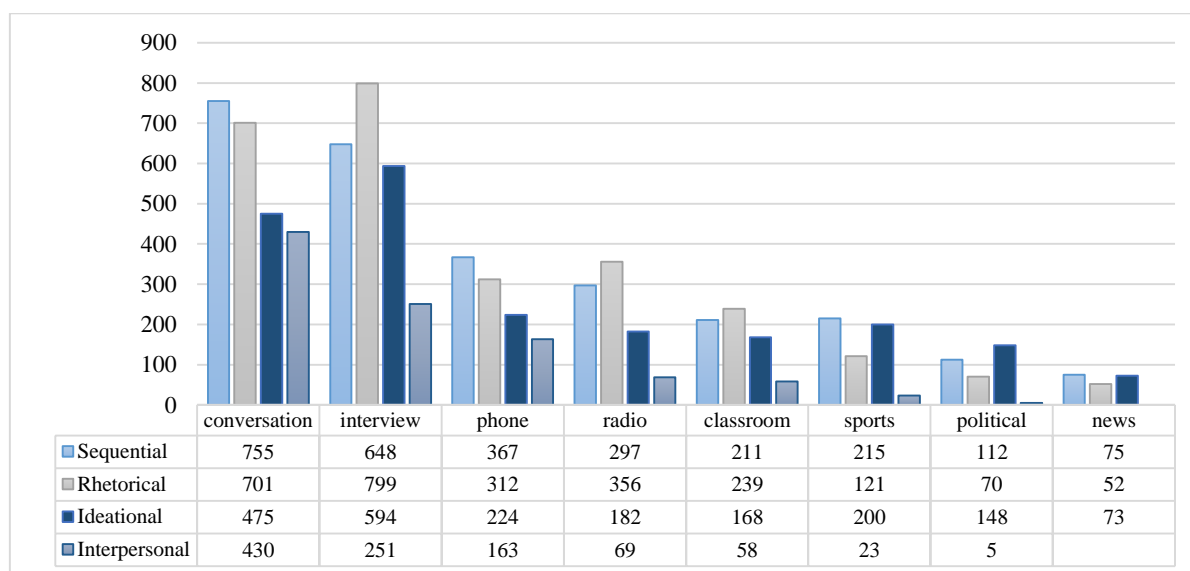
	English		French		Total	
	DMs	ptw	DMs	ptw	DMs	ptw
Sequential	1269	14.69	1411	18.74	2680	16.57
Rhetorical	1319	15.26	1331	17.68	2650	16.39
Ideational	1144	13.24	920	12.22	2064	12.76
Interpersonal	322	3.73	677	8.99	999	6.18
Total	4054	46.91	4339	57.63	8393	51.9

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 strike as 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).

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 5.7) 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 specificities of English and French,³⁹ 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. The distribution of DM domains across registers is provided in Figure 5.9.

Figure 5.9: Distribution of DM domains across registers



This graph clarifies the rivalry 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. The preference of both rhetorical and sequential DMs over ideational DMs is particularly surprising in classroom lessons which could 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 (17)-(19).

³⁹ But see above for the exclusion of English tag questions.

- (17) 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)
- (18) 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)
- (19) 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 (17) 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 (18), 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 (19) 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 5.9), 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 5.9.

Table 5.9: Relative frequency of domains (ptw) by number of speakers

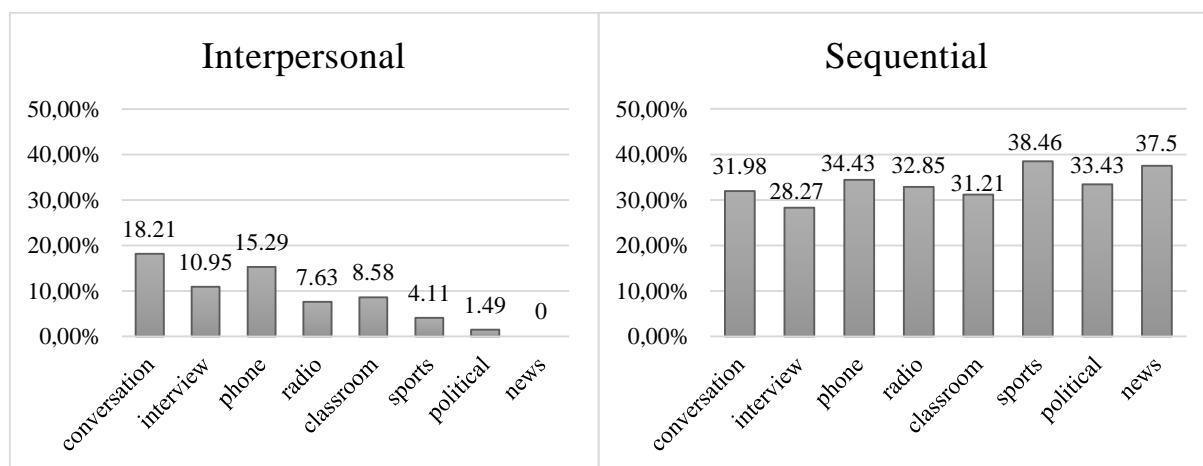
	Sequential	Rhetorical	Ideational	Interpersonal	Total
monologue	10.18	8.07	10.27	1.22	29.74
dialogue	19.88	20.64	14.07	8.71	63.30
multilogue	8.69	9.31	8.69	2.48	29.17

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 registers. 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).

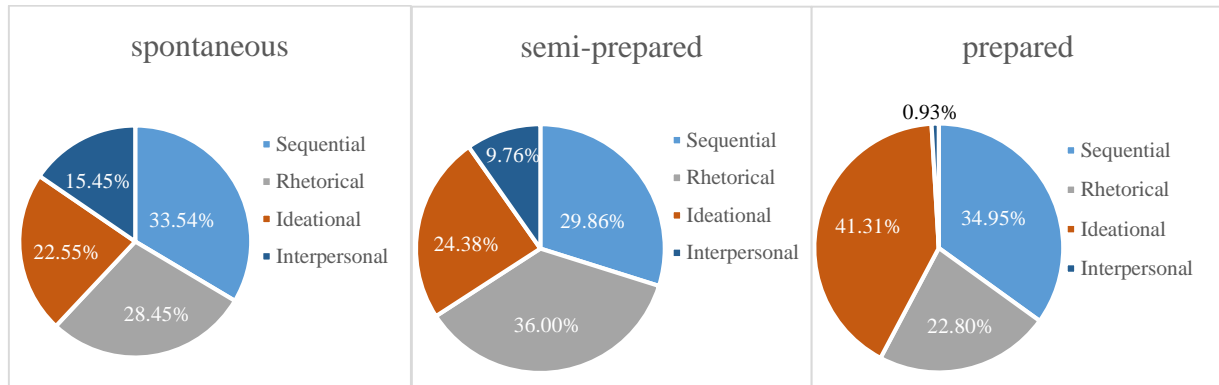
The interpersonal domain, which has scarcely been discussed so far, appears with a low frequency across all registers, especially those related to formal features (monologues, prepared, non-interactive). By definition, interpersonal DMs are connected to dialogue, which is reflected in Table 5.9. In fact, 84% of all interpersonal DMs occur either in conversations,

face-to-face interviews or phone calls. Their appearance in other registers is anecdotal, 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 Figure 5.10. In these pie charts, we see that sequential DMs constantly take up about 30% of all DMs in each register, whereas the situation is much more irregular for the 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.

Figure 5.10: Proportions of interpersonal and sequential DMs in each register



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 5.11 (the variation by degree of interactivity is highly similar).

Figure 5.11: 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 5.3.2.2 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.

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, could serve as robust cues for the automatic disambiguation of DM domains (see also Bolly et al. 2015, in press 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 5.10.

Table 5.10: 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 multifunctional 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 (20).

- (20) 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.⁴⁰

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 qualifies 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 above-mentioned 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 5.3.4). Examples (21)-(23) illustrate these interpersonal patterns.

⁴⁰ This is one of the reasons why Crible & Degand (under review) propose to re-structure the functional taxonomy and annotate these cases as “interpersonal contrast”.

- (21) 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)
- (22) yeah I'm just phoning up and doing that thing I was talking to you about **you know**
(0.300) recording (EN-phon-05)
- (23) 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*, cf. Section 5.2.2.2), 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 5.3.4).

Finally, I will combine the two functional variables investigated so far, namely (non-) relational type and domain, in order to uncover potential associations and restrictions. In Section 5.3.1, I identified particular DM expressions which are specific to one type (RDM or NRDM) and a few ones which can be used in both ways. Such associations should reasonably be reflected on the cross-tabulation of domains and types, as reported in Table 5.11.

Table 5.11: (Non-)relational type by domain

	Non-relational			Relational		
	EN	FR	Total	EN	FR	Total
Sequential	454	648	1102	815	762	1577
Rhetorical	143	116	259	1176	1214	2390
Ideational	0	0	0	1144	920	2064
Interpersonal	322	677	999	0	0	0

Clear conclusions can be drawn from this table. First, we see that, in both languages, two domains are restricted to one of the two types, namely interpersonal (NRDM) and ideational (RDM). This reflects the very definition of the functions in these domains, as well as the speech-specific nature of interpersonal DMs as opposed to ideational relations which also exist in writing. More interestingly, sequential and rhetorical DMs can be used both in relational and non-relational contexts, although not to the same extent: there is no restriction in language nor register for the duality in the use of sequential DMs, whereas rhetorical DMs can only occur as both types in intermediary and informal registers (all settings except news broadcasts, political speeches and only three NRDMs in sports commentaries, see Appendix 6). A more fine-grained analysis at function-level should reveal what specific function values are responsible for these dual uses (see next section).

In sum, analyses at domain-level reveal clear tendencies of variation across languages, registers and additional variables such as POS or (non-)relational type. The multifunctionality 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). However, this multifunctionality is not random but follows the systematic effects of methodological and theoretical considerations. More complex statistical models, with additional variables including syntax, will be provided in Section 5.3.4 to further our understanding of this variation.

5.3.2.2 Functions

The third and most 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, cf. Section 4.2.1.2). 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 one (non-)relational type and to particular 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 7. Only the ten most frequent functions are reported in Table 5.12.

Table 5.12: 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

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 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

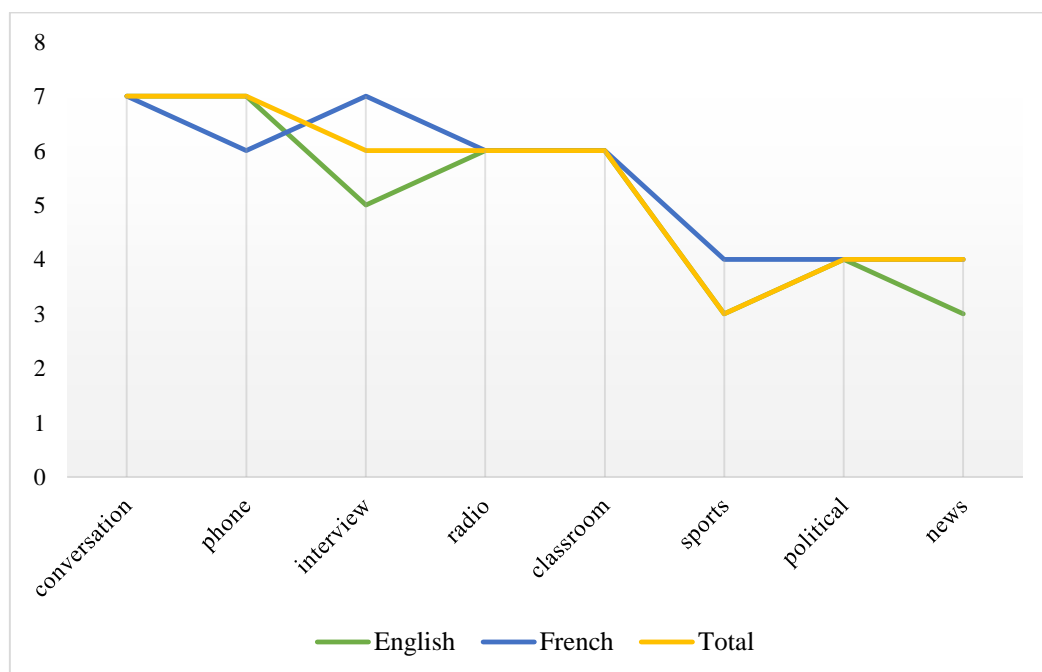
different ranks. The main crosslinguistic difference concerns the *monitoring* function, which ranks 2nd in French against only 7th in English with a highly significant gap in frequency (LL = 123.32, $p < 0.001$). *Monitoring* is mostly expressed by *you know* in English (180/236) and *hein* ‘right’, *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%; $z = -1.195$, $p > 0.05$).

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 below (see Appendix 8 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 could be interpreted as a result of the artificial nature of this type of language where topics are usually explicitly changed.

Apart from these specificities, another source of contrast between registers might be their varying functional diversity. As announced in the previous section, an analysis of DM function ratio could 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 large 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 5.12.

Figure 5.12: 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 four overall, with a large drop occurring in sports commentaries (only three types in English and overall). Although the differences are small, they could 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” which was already used in Section 5.2.2: 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 5.13. 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. 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 5.13, which place conversations at an intermediary level of functional diversity against the more varied range of radio interviews, for instance.

Table 5.13: 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

Finally, restrictions of (non-)relationality, already identified at domain-level, will now be analyzed at the more fine-grained function-level in order to identify type-specific functions and functions that can be used in both contexts. Bearing in mind the methodological precautions developed in Section 5.3.1, we can identify the specific function values responsible for the restrictions of type per domain, namely the restrictions of interpersonal and ideational domains to NRDM and RDM, respectively, while sequential and, to a lesser extent, rhetorical DMs can be both relational and non-relational. In the data, exclusively relational functions include, among others, *addition*, *conclusion*, *temporal* or *consequence* (in other words, typical discourse relations); exclusively non-relational functions are *monitoring*, *punctuating*, *closing*, *approximation*, *ellipsis* and a few less frequent ones (in other words, speech-specific functions). The *approximation* function in particular takes up most occurrences (59%) of non-relational rhetorical DMs previously discussed, as in the following example:

- (24) I come back talking a little bit more like a Liverpoolian (0.390) but I (0.270) **kind of** lose that in a in a short time (EN-intf-03)

This *approximation* function is the only one which is exclusively non-relational in the rhetorical domain, while all other functions are either exclusively relational (e.g. *conclusion*) or mostly relational (e.g. *specification*). This observation of DMs which can be either relational or non-relational raises the question whether it is coherent to allow for this duality in a single function: should occurrences of relational and non-relational *specification* actually be grouped in a single category or rather form their own (sub)type of function? Examples of the DM *actually* are particularly telling in this matter:

- (25) I had t- t- this grounding in theatre design and so I you know even on (0.373) a film like the Tempest it was **actually** to explode that whole idea of design (EN-intr-04)
- (26) you may have heard of the uh BBC wildlife programmes (0.350) all the BBC wildlife programmes are **actually** based at Bristol (EN-intf-05)

In both examples, there is a common referent between the two utterances in the context of *actually* (“design” and “programme”). However, in (25), the relation or continuity marked by the DM does not so much apply between the two utterances as between the left-dislocated element “a film like the Tempest”: the first utterance functions more as background information, while the meaning of *actually* is more intra-sentential and flirts with nuances of

emphasis and counter-expectation. By contrast, the relation introduced by *actually* in (26) is fully inter-sentential and can be reformulated by a relative clause (“you may have heard of the BBC wildlife programmes, which are based at Bristol”) typical of elaboration or specification relations. There is probably more than one way to interpret these examples, which are open to debate: *actually* is one of the most ambiguous DMs to annotate compared to other DMs with a stronger core meaning. This evasive meaning of *actually* is corroborated by authors such as Aijmer (2002: 265) who identified a large functional spectrum including contrast, justification, elaboration, evaluation, self-correction, topic change and softener. In other words, non-relational cases of *specification* may in fact be the result of the underspecification of the DM: the relation is less strongly marked in examples like (25) compared to (26), leading to a categorization as non-relational, although this decision is more concerned with a methodological bias than an observable linguistic difference.

The duality also applies at the sequential domain, especially with the *topic-shift*, *topic-resuming* and *enumeration* functions. Again, the difference between RDM and NRDM uses of these functions may be more a question of degree than a true functional divide. Since sequential functions apply at a higher level of discourse structure, topic relations or lists can take scope over elements which are quite distant from each other. Qualitative comparison of RDM and NRDM *topic-resuming*, for instance, reveals that NRDM cases tend to correspond to long-distance relations, whereas RDMs are more adjacent (although, by definition, *topic-resuming* involves an interruption of the ongoing topic). Examples (27) and (28) illustrate these typical patterns.

- (27) now it's very much 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 <laughing/> (0.160) **but** it's things like 'yeah' and 'you what' and (0.230) whereas you know mine is 'yes' 'pardon' or whatever (EN-intf-03)
- (28) <ICE_44> what I really remember is once (1.420) uh just after the war when we'd moved back to our house in Sheffield which we'd left because of fear of the bombs and my father (0.360) had come out of the air force and he was carving a huge great joint (0.340) and he suddenly recited the whole (0.313) of Keats's ode to the nightingale [...]
- <ICE_43> so you and your your uh brother and two sisters enjoyed all that did you
- <ICE_44> oh yes I think uhm (0.167) I don't know what my mother would have done if we had not come out naturally bookish but we did come out naturally bookish and
- <ICE_43> and i- is that something that you've tried to continue with your own family do you do you (0.400) m- (0.427) make literary allusions as you go about the o- ordinary domestic business
- <ICE_44> uhm I do indeed and it annoys almost all of them one of my daughters (0.420) at a party opened the front door to the guests and said uhm (0.547) uhm I am the youngest daughter of this house I do not read books there are too many books in this house excuse me I will take your coat (0.400)

and then <laughing/> uhm (0.307) I I think I'm a bit too intense about books

<ICE_43> **so** on going back to your to your childhood (0.220) it was your mother wasn't it who was the the driving force behind all of this behind this sort of intellectual rigour (EN-intr-05)

In (27), the speaker is talking about a regional accent which she describes as a “lazy speech”, before introducing a humorous anecdote about her nephew. She then resumes the description of the accent with “but” and an anaphorical pronoun “it”. In this example, the two utterances connected by “but” are only separated by the two segments about the nephew, and the proximity of the interrupted description favors a categorization as relational. By contrast, in Example (28), the interviewer <ICE_43> asked the interviewee to tell childhood memories about her literary parents. <ICE_43> asks a follow-up question (“is that something that you've tried to continue with your own family”) which leads to an anecdote about the interviewee's daughter. Finally, the interviewer comes back to her original topic of childhood with a turn-initial “so” and an explicit, lexicalized signal of topic-resuming (“going back to your childhood”). In this second case, it is not certain what the segment introduced by “so” is connected to (right before the digression on the interviewee's daughters or, before that, the original question about her parents which does not appear in this extract). In any case, the relation is quite distant and “so” seems to function as a non-relational discourse-structuring device.

A perhaps more convincing illustration of the functional divide between the relational and non-relational types is provided by the cases of *enumeration* where the former (RDM) are cases of completed lists, as in (29), while the latter (NRDM) are pointers to discourse referents which do not necessarily enter a list construction, as in (30).

- (29) <VAL_3> vous arrivez à vous rendre compte que vous avez un accent
 <VAL_2> mm
 <VAL_3> qui n'est pas celui que **un** vous aviez l'impression d'avoir et **deuxièmement** que (0.560) qui n'est pas le même que celui que vous pensiez avoir
 <VAL_3> *you realize that you have an accent*
 <VAL_2> *mm*
 <VAL_3> *which **un** 'one' is not the one you felt you had and **deuxièmement** 'secondly' which is not the same as the one you thought you had* (FR-intf-01)
- (30) <VAL_2> existe-t-il des personnes qui s'expriment très bien (0.493) et si oui qui sont-elles à quoi est-ce lié
 <VAL_3> ben je crois qu'en définitive [...] euh moi je crois que **un** c'est dû à l'éducation
 <VAL_2> mm
 <VAL_3> euh quelqu'un qui a fait des études romanes ou qui a fait des études euh (0.720) de grec ancien latin ancien etcaetera parlera mieux que celui qui a fait des études techniques

- <VAL_2> *are there people who speak very well and if so who are they what is it linked to*
- <VAL_3> *well I think in the end uh me I think that **un** ‘one’ it’s linked to education*
- <VAL_2> *mm*
- <VAL_3> *uh someone who studied Romance languages or who studied ancient Greek ancient Latin etcaetera will speak better than someone who did technical studies (FR-intf-01)*

The enumerating DM pair “un ... deuxièmement...” in (29) meets both the criteria of short distance and completed list, which favors a relational reading. In (30), however, the context is very similar (same speaker, same DM, similar topic) but the list construction is very different: only one element is introduced and is then elaborated, with no return to the enumeration (no other element is added to the list afterwards). In this case, “un” emphasizes and points to an argument, which might also be interpreted as “primarily it’s linked to education”, thus motivating the non-relational label. All in all, the annotation of (non-)relational type needs to be better operationalized to reinforce these interpretations of examples in a more systematic way. Nonetheless, this variable, when mapped with the more fine-grained function-level, shows some potential to co-vary in regular patterns with linguistic features such as underspecification, long-distance scope or discourse (in)completion.

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, observations of functional diversity and mappings with the promising (yet still unstable) variable of relationality. Functional considerations at such a fine level of granularity will be taken up in Chapter 6 in relation to hypotheses of fluency.

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, so much so that only a few observations will 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 (cf. the three levels of multifunctionality defined in Section 3.1.1). Their distribution is reported in Table 5.14.

Table 5.14: 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 (31), where *so* expresses both a conclusion and a topic-resuming function.

- (31) 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 (31) 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, could be interpreted in multiple ways, either as a result of the very high and similar frequency of these two domains in general (cf. Section 5.3.2.1), 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. This multifunctionality is compatible with the definition of DMs and their role in “local and global content and structure” (Fischer 2000: 20), respectively represented by rhetorical (local content) and sequential (global structure) functions.

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. This is reflected in the subjective score of complexity documenting the annotator's confidence (1 is high confidence and 3 is high hesitation), with 37% of all double tags being coded as levels 2 or 3, as opposed to only 5% in single tags. In addition, the intra-annotator reliability study reported in Section 4.2.2.3 showed that the categories which were most disagreed upon are double tags, namely RHE-SEQ, INT-SEQ and IDE-RHE: double tags amount to 23% of all disagreements (against their overall frequency of 4%), which points to the need to better operationalize this option in the annotation procedure.

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 5.15. 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).

Table 5.15: 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%

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 thesis will focus on the 8,393 single-tagged DMs whenever functional variables are concerned.

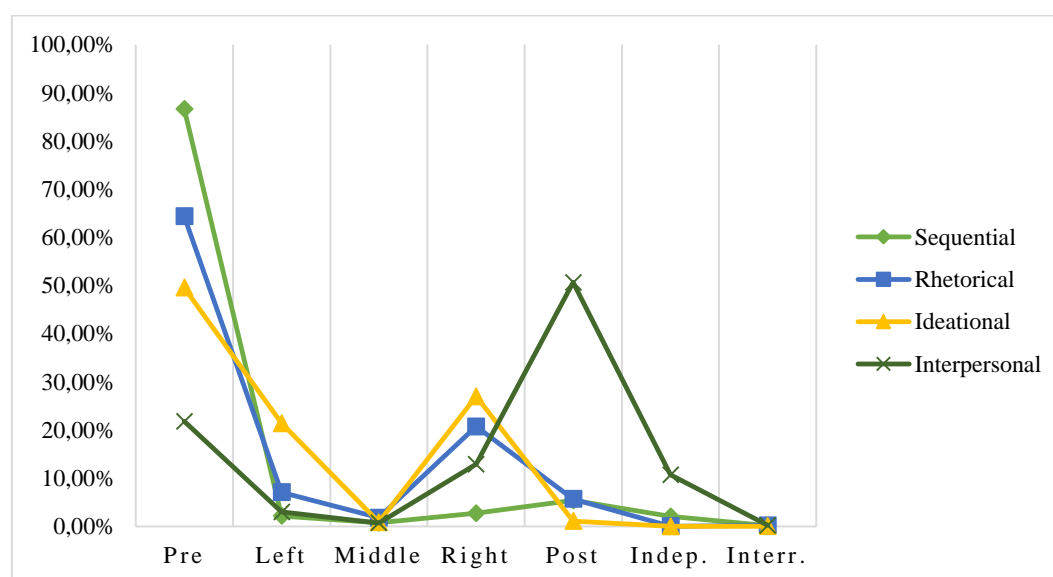
5.3.4 Integrating syntax and pragmatics

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, as developed in Section 3.4:

- the higher discourse scope of sequential DMs should 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 could 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 5.13.

Figure 5.13: Proportions of macro-syntactic slots in each domain



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” 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).

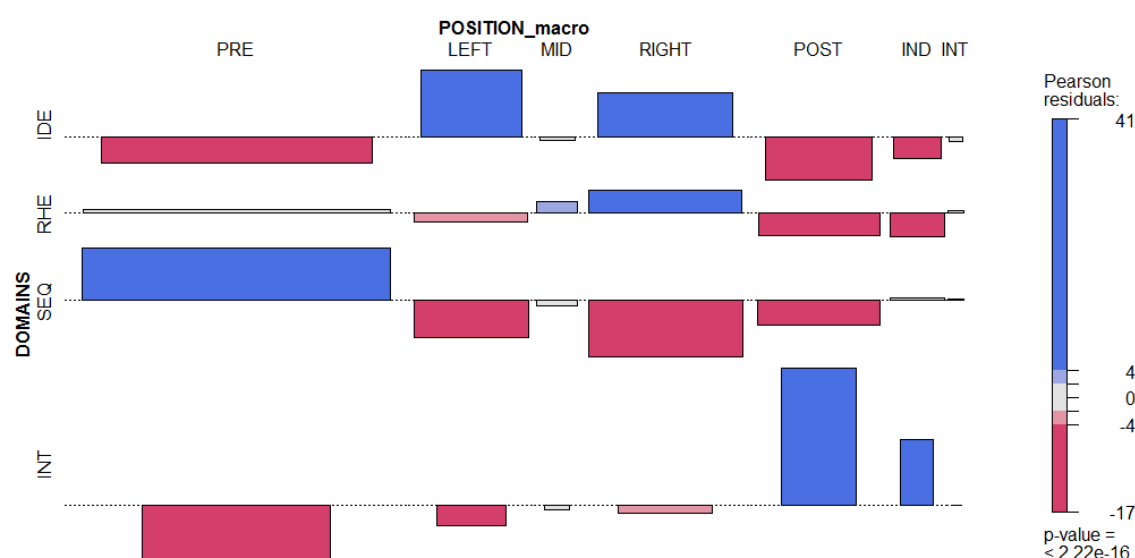
The interpersonal domain is, as hypothesized, strongly associated with final, non-integrated position (“POST”) in half of all its occurrences, as opposed to all the other domains where this slot is very rare. A substantial proportion (22%) of interpersonal DMs also occur in pre-field position, although 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 notable specificity 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 (32)-(35) illustrate the most frequent pattern for each domain:

- (32) 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)
- (33) 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)
- (34) 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)
- (35) 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 5.14) 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 residual. The color of the rectangles indicates a positive (blue) or negative (red) association (grey means no significant difference).⁴¹ Extended association plots go beyond mere frequency and show which patterns are significantly more or less frequent, relatively to their competitors.

Figure 5.14: Extended association plot of domains and macro-position



Starting with significantly positive associations, this plot confirms (i) the attraction of sequential DMs in pre-field position, (ii) the use of interpersonal DMs in post-field and independent positions and (iii) 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 5.2.2.2 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

⁴¹ 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).

(i) ideational DMs occur primarily utterance-initially (as in Example (34) above) and (ii) 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 specificity. 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 or schemas:

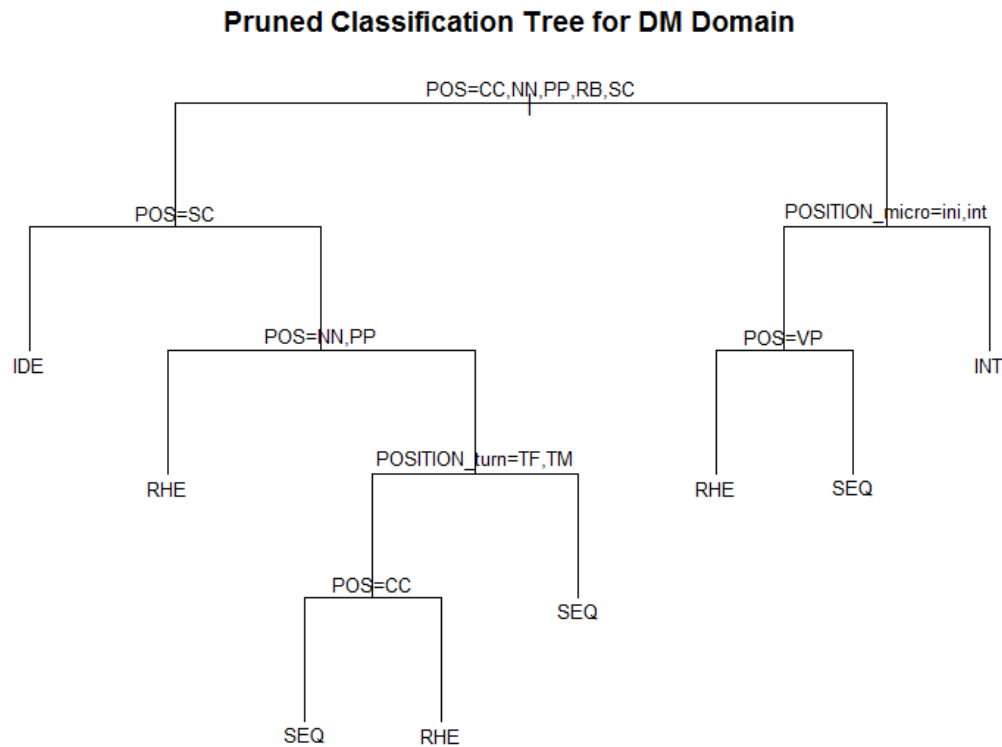
- discourse-relational functions, either objective or subjective (ideational and rhetorical), show a relative preference for (right-)integrated contexts and a dispreference – or at least absence of significance – for peripheral (pre- and post-field) positions;
- discourse-structuring functions are strongly (and relatively) associated with the initial position;
- hearer-oriented 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 5.15 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.

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.⁴²

⁴² As a reminder, these abbreviations are directly borrowed from the PDTB tagset (Santorini 1990), cf. Chapter 4.

Figure 5.15: Pruned classification tree of domains

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 discussed so far in this section. This result does however not qualify 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 multifunctionality 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 CC 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 6). 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.

5.4 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. Waltireit 2007; Pons Bordería 2008; Cuenca & Marín 2009; Dostie 2013), and relates to the positional flexibility and multifunctionality 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).⁴³ Concretely, 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. The results discussed in the following sections are based on the existing annotations in *DisFrEn* and their mapping with Cuenca & Marín's (2009) gradual model of co-occurrence presented in Section 3.2.2. Multiple factors influencing DM co-occurrence, both from metadata and DM annotations, will be progressively modeled in an all-encompassing statistical model (Section 5.4.1). In a second, more qualitative analysis, some of these annotated features will be tentatively mapped with co-occurrence degrees in order to refine our understanding of this multi-faceted phenomenon (Section 5.4.2).

⁴³ 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."

5.4.1 Modeling the factors of co-occurrence

As pointed out in Sections 4.2.1.1 and 4.2.1.4, 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. However, they will be included in the qualitative analysis as a potential mapping equivalent to Cuenca & Marín’s (2009) “composition” category (see Section 5.4.2.3).

5.4.1.1 Co-occurrence across registers and languages

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 5.16 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.⁴⁴ 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 are subject-specific, respectively.⁴⁵ 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

⁴⁴ Mixed-effect models (both linear and logistic) were computed with the {lme4} package (Bates et al. 2014) on R-studio.

⁴⁵ Regression models usually take speakers as random effects in order to control for individual variation and sociolinguistic variables. In the present state of *DisFrEn*, speakers metadata is not available (cf. Section 4.1.1). Only the text or transcript in which each DM occurred can be identified, hence the use of individual transcripts as random effects. The models are therefore slightly more coarse-grained than if speakers (instead of transcripts) had been used.

generalizability of the model to other populations (e.g. Barth & Kapatsinski in press). 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.

Table 5.16: Combinations of DMs by decreasing frequency (excluding *hapax legomena*)

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;	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

Each factor in a regression model always takes one level of the variable as reference (e.g. English is the reference level for the factor “language”) against which the other levels are compared. In R-studio, this reference level is selected alphabetically. However, in all the regressions in this thesis, the reference level was manually changed when relevant, that is, when the factor is ordered and one level is conceptually more “basic” than the others (e.g. the initial position was systematically used as reference level in models where position was included). 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.

5.4.1.2 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 was hypothesized to favor the initial position. We can see in Table 5.17 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 5.17: 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	20.3%	82.24%
medial	477	38	515	7.38%	2.18%
final	719	256	975	26.26%	14.71%
independent	168	15	183	8.2%	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 (36) and clause-medial in (37).

- (36) 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)
- (37) 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 5.16). 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). This conclusion should, however, be confirmed by a more fine-grained analysis of functions and co-occurrence degrees as carried out in Section 5.4.2.

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 (38) and (39).

- (38) 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)
- (39) 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 (38): 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 qualify the previous statement. In (39), 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 qualification 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. The analyses in Section 5.4.2 will refine these interpretations.

5.4.1.3 *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). Before turning to a more fine-grained analysis of different types of co-occurrences in the next section, I will 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, score of coding complexity, 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 5.4.1.1), 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.
- *Complexity* (numeric): the higher the coding complexity, the higher the chance of 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 map non-relational vs. relational DMs and may 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 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 under review). 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. Lastly, DMs which were subjectively coded as difficult to disambiguate seem to frequently co-occur, which might also be an effect of their attraction to final position. Another interpretation would suggest that ambiguous DMs (i.e. those that are difficult to annotate) occur in somewhat problematic or at least underspecified contexts where DMs tend to cluster, as a compensating strategy to cope with weakly encoded pragmatic meanings. The analyses in the next section will further this line of reasoning.

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, it is only by zooming in on particular patterns combining these different features that a more fine-grained view of (different types of) co-occurring DMs can be achieved, which suggests the resort to more qualitative (or rather qualitative-quantitative) methods of investigation in order to rank different DM clusters on a scale of integration or fixation and interpret the attraction of some features in light of cognitive, usage-based hypotheses.

5.4.2 The company they keep: from co-occurring to complex DMs

One underlying assumption in this research states that discourse-level elements frequently co-occurring in the speech string should bear some sort of cognitive-functional similarity and be used as meaningful complex units. In order to further our understanding of the phenomenon of discourse-level co-occurrence, the combinations identified above can be refined by integrating functional variables as well as more qualitative considerations of co-occurrence degrees, based on Cuenca & Marín's (2009) proposal. In other words, I mean to uncover whether co-occurrence (of DMs, in this case) involves redundancy or complementarity. Results from the previous section suggest that the situation is not binary, especially in light of the prominent place of underspecified DMs such as *and* or *quoi* which can combine either for pragmatic reinforcement or simply because their weak semantic content is compatible with the meaning of more specific DMs.⁴⁶ This analysis will also pursue the objective of clarifying the link between interactivity and cohesion: the former is typically expressed by final interpersonal DMs whereas the latter mainly corresponds to initial discourse-relational DMs. While these two types of DMs are very different, they frequently co-occur, which questions the apparent opposition between interactivity and cohesion. For this reason, I will focus on DM clusters extracted from the conversational subcorpus, where interpersonal DMs are most frequent.

The functional annotations in *DisFrEn* will be tentatively mapped with the three degrees of integration for co-occurrence patterns proposed by Cuenca & Marín (2009) which are

⁴⁶ The underspecification of French *quoi* 'right' might be less consensual than for *and*. However, in the data, this DM was assigned no fewer than five functions with more than 10 occurrences (viz. *monitoring*, *closing boundary*, *punctuation*, *conclusion*, and *face-saving*) and several other less frequent ones, which points to its multifunctionality and resulting underspecification.

developed below. In their paper, the authors distinguish DM clusters based on a combination of syntactic and semantic criteria, which they found to be frequently associated to patterns of positional and functional use:

- *Juxtaposition*, where two or more DMs co-occur but do not combine syntactically nor semantically (i.e. different functions, scope over different segments). This degree is often instantiated by conjunctions (coordinating or subordinating) in “act internal” position (i.e. initial position between a pair of connected utterances) and expressing propositional meanings (e.g. *and when*).
- *Addition*, where two or more DMs take scope over usually local segments but remain functionally distinct (i.e. different functions, same scope). This pattern often corresponds to combinations of conjunctions with parenthetical or pragmatic connectives (e.g. *but actually*) in act-internal position or at minor transition places, typically expressing propositional or structural meanings.
- *Composition*, which constitutes the most integrated level of co-occurrence, whereby two DMs (rarely three) are used as a single complex unit (i.e. same function, same scope) and tend to indicate a global discourse function, without being completely lexicalized (e.g. *pues vale* in Spanish). These constitute combinations of parenthetical and pragmatic connectives occurring in major transition places and expressing structural-modal meanings (or sequential-interpersonal in terms of the present taxonomy).

These categories will be put to the test of systematic corpus annotations extracted from conversational data, verifying the extent to which the distinctions and criteria used by Cuenca & Marín (2009) actually match the instantiations of co-occurrence in the corpus. Concretely, any corpus-driven combination of DM features which conceptually corresponds to one of the three degrees of integration will be qualitatively analyzed in order to verify whether the examples match the definition. In other words, this analysis investigates whether top-down categories can be confirmed by formal-functional bottom-up patterns.

Lastly, the role of frequency will also be examined, following the hypothesis that the most frequent clusters should contain DMs expressing similar functions. High frequency clusters can be interpreted as candidate “complex” units undergoing a process of lexicalization or fixation, given the capital role of frequency of use in language change (e.g. Ellis 2016). In the end, I hope to disentangle the continuum from merely co-occurring to fixed, complex DMs.

5.4.2.1 *Potential equivalents of juxtaposition*

Juxtaposition is the least integrated type of co-occurrence: the DMs just appear to be co-located in the speech string without any further connection. Based on Cuenca & Marín’s (2009) definition, the following patterns, manually identified from the data, can be expected to meet this low degree of fixation: (i) clusters of more than two DMs (low frequency); (ii) clusters of both relational and non-relational DMs (different number of segments); (iii) clusters of DMs across final and initial positions (backward- vs. forward-looking scope); (iv) clusters including a subordinating conjunction (intra- vs. inter-clausal scope). They will each be illustrated by authentic examples and confronted to the original definition of juxtaposition. Starting with the

number of DMs in a cluster, the great majority of occurrences include two DMs only.⁴⁷ Of the remaining 37 cases, only three are clusters of four DMs: these can be safely discarded as being anything more than mere juxtaposition, given their rarity and rather awkwardness or unnaturalness, as in Example (40).

- (40) <ICE_6> at the end of the day she got a very indifferent degree which I can comfort myself with (1.090) I mean
 <ICE_5> done all right? she's happy?
 <ICE_6> yes **well I mean you know so** that's what I'm trying to say that you know all these things that Linda sets such great store by at the end of the day (0.600) don't add up to a row of beans (EN-conv-08)

In this example, <ICE_6> is talking about an acquaintance (Linda) who used to boast about her education achievements and ended up having low grades (“a very indifferent degree”), which rejoiced <ICE_6>. <ICE_5> then asks whether this person ended up being happy, to which <ICE_6> answers positively and then produces a number of DMs which might express either embarrassment (for her rather mean comments), mitigation (“she is happy but not too much”) or common ground (“you know what it's like”), before she introduces a summary or conclusion of her anecdote with “so”. In terms of scopes and functions, in spite of the many interruptions and resulting difficulty of interpretation, the following readings can be proposed: *well* launches the new utterance (global scope, sequential function); *I mean* seems to introduce a mitigation or reformulation of “yes” or “happy” which is not completed (local scope, rhetorical function); *you know* is calling for cooperation either on the full story or the upcoming conclusion (unclear scope, interpersonal function); *so* summarizes the whole anecdote by “that's what I'm trying to say that...” (global left scope, conclusive function). Even though these interpretations can be challenged, it remains that the DMs in this cluster are clearly not well integrated, do not combine into a coherent fixed unit but rather express distinct meanings pointing in different directions.

The situation is not as straightforward for the 34 clusters of three DMs, which can include stronger ties between two out of the three components, as in (41).

- (41) moi je j'aime pas euh me battre et cette ambiance de de bagarre et tout ça je n'aime pas quoi (0.440) **et voilà quoi** j- je m'en voulais vraiment d'avoir fait ça
I don't like to fight and that fighting atmosphere and all that I don't like it you know
et voilà quoi ‘and that's it you know’ I was angry at myself for it (FR-conv-05)

The string *voilà quoi* is quite frequent, either on its own or clustered with one (or two) other DMs, which might point to a stronger degree of fixation (see Section 5.4.2.3). However, taken as a whole, the three-DM cluster in (41) does seem to meet the criteria of juxtaposition with different scopes and functions at least between *et* and *voilà quoi*: the former connects the previous context to the following DMs in a relation of conclusion; the latter, by contrast, seem

⁴⁷ Occurrences of complex DMs (e.g. *and then*, *et puis*) were counted as one DM: a cluster such as *and then when* is therefore categorized as a co-occurrence of two DMs. 46 complex DMs were extracted from conversations based on the aforementioned closed list.

to function independently as closing signals by which the speaker expresses his unwillingness (or inability) to expand on the ongoing conflicting topic.

This example also illustrates the next pattern, namely the co-occurrence of relational and non-relational DMs, which concerns 100 cases among clusters of two DMs. RDMs necessarily take scope over two abstract objects whereas NRDMs only apply to one segment (of varying size). This intrinsic difference in scope bounds such cases to be categorized as juxtaposition. Both orders (NRDM-RDM, RDM-NRDM) are represented in the following examples, respectively:

(42) with her present job she's sort of uhm (0.650) having (0.810) high jobs being wafted under her nose as as a sort of uh incentive **you know and** she said you know even vice president 'so what?' she said (EN-conv-08)

(43) je dis 'dégage d'ici' et (0.330) il s'est assis (0.630) ben je dis ben je vais prendre mes affaires et m'en aller **alors hein** (0.480) parce qu'ou sinon j- j' allais vraiment (0.610) vraiment frapper dedans

*I say 'get out of here' and he sat down well I say well I will take my things and go **alors hein** 'then right' because otherwise I was really going to hit him* (FR-conv-05)

In (42), “you know” shows a backward-looking scope, typical of its monitoring function, whereas *and* connects the previous utterance to the next to allow the narration to go on. In (43), “alors” is also backward-looking but in a relational sense, connecting the fact that he sat down to his consecutive leaving; “hein”, on the other hand, only takes scope over the latter. These examples are clear cases of juxtaposition. Nonetheless, a substantial number (32/100) of cases in this pattern involve a *quoi* followed by a conjunction, mostly *mais*: the high frequency of this cluster, already discussed above, could be a sign of its higher integration or fixation relatively to other juxtaposed patterns (see Section 5.4.2.3).

The next pattern of juxtaposition often coincides with the previous one and includes clusters of final and initial DMs. This means that the DMs in the cluster not only have different scopes but are attached to different segments altogether, which results in different syntactic positions, as in Example (42) above. This pattern concerns 51 cases (out of 251 two-DM clusters), including 31 *quoi* + conjunction (e.g. *quoi mais*, *quoi parce que*, *quoi quand*). Functionally, the large majority (41/51) of these cases initiate with an interpersonal DM (*quoi* but also *tu vois*, *you know*) and continue with either a rhetorical DM (27 cases, e.g. *I mean*) or an ideational one (12 cases, e.g. *parce que* ‘because’). This can be taken as evidence of the link between interpersonal and more cohesive, discourse-relational functions, which frequently co-occur together although the DMs belong to different segments.

As suggested in the previous section, this tendency qualifies the apparent gap between interactivity and cohesion. These two domains or general functions of language, although expressed by different forms, often co-locate in the speech string, which may be interpreted as a sign of their conceptual proximity or at least complementarity. It might even be said that there is more than one way to appear coherent and hearer-oriented during an interaction, and that both interpersonal and discourse-relational DMs are necessary for communicative success. The analysis at fluency-level in the next chapter should enlighten whether these functions of DMs

are associated to similar or different sequences of fluencemes and whether interpersonal DMs are necessarily more disfluent than relational DMs, as a writing-based definition of fluency would suggest.

Coming back to NRDM-RDM clusters, we see that 21 of them (27%) include a subordinating conjunction as second DM in the cluster, which also maps a number of final-initial cases discussed above. The syntactic disparity between the components of such clusters is striking: they originate from different grammatical classes, occur in different positions and take scope over segments of different syntactic status (governed or not, utterance-internal or not). Zooming in on these clusters, we see that, apart from the frequent *quoi*-clusters, most of them are turn-initial, as in Example (44).

- (44) <ICE_7> it really does have to be uhm (2.893) two twenty (1.300) well (0.613) two thirty maybe you know
 <ICE_8> **well if** he's gone against the agent's advice already and slapped another fifty (0.407) on top he's hardly likely to suddenly come right down again (EN-conv-01)

Here, “well” opens the new turn while “if” starts a conditional relation with the main clause “he’s hardly likely to come down”. Again, the difference in function and scope seems to call for a categorization of this type of co-occurrence as juxtaposition. In sum, patterns of syntax and functions can be straightforwardly linked to a low degree of integration between co-occurring DMs, with the potential exception of very frequent *quoi*-clusters (*voilà quoi, quoi mais*) whose recurrence vouches for a stronger degree of fixation.

5.4.2.2 Potential equivalents of addition

The second degree of co-occurrence, *viz.* addition, is defined as a similarity in scope and a difference in function. In the data, these cases are more challenging to identify on the basis of syntactic and functional annotations of DMs, especially in the absence of a systematic annotation of DM scope. An indirect access to these clusters would be to select co-occurring DMs functioning in the same domain (e.g. two ideational DMs with different functions), following the hypothesis that the conceptual and semantic similarity of values within one domain should be reflected by a similarity of scope: ideational DMs will tend to relate local segments; sequential DMs will tend to function at a higher level of organization; interpersonal DMs will tend to be non-relational and backward-looking; rhetorical DMs are more fluctuating between local relations (e.g. *motivation*) and non-relational meanings (e.g. *approximation*).

Only 47 clusters in the sample share a domain and not a function, from which we have to exclude 11 cases of three- and four-DM clusters and eight cases involving a subordinating conjunction (previously categorized as juxtaposition). Among the remaining 36, two main patterns can be distinguished: interpersonal DMs combining *ellipsis* with *monitoring*, typically involving the French general extender *et tout ça* ‘and all that’ (45); sequential DMs combining *closing* with *topic-resuming* or *punctuation* (46).

- (45) <VAL_20> je triche un peu quoi
 <VAL_19> pourquoi

- <VAL_20> ben je prends des gens **et tout ça quoi** c'est pas bien hein
 <VAL_20> *I cheat a little you know*
 <VAL_19> *why*
 <VAL_20> *well I take people with me et tout ça quoi 'and all that you know' it's not good is it (FR-conv-05)*
- (46) <ICE_77> the reason I think is that uhm modern medicine (1.120) now enables people to cope
 <ICE_76> you're getting cheese on your (0.900) jumper
 <ICE_77> it'll improve the flavour it'll improve the flavour uhm it enables <laughing/> people to come through uhm
 <ICE_76> improve appearance more
 <ICE_77> yeah (1.227) **now then** uhm
 <ICE_76> by the way Liz is okay for going to the uhm Verdi (EN-conv-04)

On the whole, these examples and others in the sample seem to confirm that DMs from the same domain tend to add – but not merge – their respective functions into a single discourse operation. Another type of additive co-occurrence concerns precisely the *addition* function typically expressed by the conjunctions *and* / *et* and accounting for 27 cases. The weak semantic content of this function, simply signaling continuity, makes it compatible with more specific meanings such as *temporal* or *consequence*, as in (47).

- (47) Dick's written on the on the the minutes or whatever student action (0.480) on (0.253) uh what do the students think of the course in general and the BA and and what could be done to improve it (0.380) **and so** (0.240) Bob drafted this questionnaire and gave it to Dick (EN-conv-06)

In this example, the two DMs connect the same segments and “so” specifies the nature of the relation by a causal (consecutive) inference (Bob drafted the questionnaire because Dick wanted to know the students' opinion). The low semantic content of *and*, in addition to the high frequency of this cluster in the two languages (cf. *et donc*), could even suggest a higher degree of integration than mere addition. It is hard to properly distinguish the contributions of “and” and “so” in examples such as (48), given that the addition or continuity signaled by the former is implied in the consecutive meaning of the latter, which should be seen as a specification or reinforcement of “and” rather than an entirely different function.

Similarly, another borderline case of additive co-occurrence is *et alors* ‘and then’ (19 occurrences) where the second DM mostly expresses *specification*: *addition* and *specification* do not seem so much distinct as complementary functions entertaining a relation of hyperonymy (the latter being a subtype of the former). I would therefore suggest that these *and*-clusters be considered as an intermediary degree of co-occurrence between addition and composition. This line of reasoning could be extended to other underspecified DMs besides additive conjunctions, such as the case of *quoi* discussed above. Such corpus-based evidence seems to suggest a link between underspecification and co-occurrence which could, in turn, motivate a specific degree of integration for these patterns, not as fixed as complex DMs – although the high frequency of some of them suggests an ongoing process of lexicalization – but more complementary than mere addition of meanings.

In sum, this second degree of co-occurrence does not fully map feature-based patterns from the annotated DMs, with exceptions and borderline cases. It appears that addition is not as much a formal-functional pattern as the other two degrees, against which it is rather negatively and gradually defined: addition is more than juxtaposition but not quite composition.

5.4.2.3 Potential equivalents of composition

The last and most integrated type of co-occurrence is composition, which corresponds to unity of syntactic and functional behavior. Obvious candidate features for this degree are “complex” DMs such as *and then* or *mais bon*, which are presently considered to function as one DM unit instead of two independent DMs (46 cases) and co-occurring DMs expressing the same function (16 cases), as in Example (48).

- (48) je me dis le Laveu uh (0.520) ça peut marcher et ça lui fait un nom **quoi tu vois**
*I think the Laveu uh it could work and it makes his fame **quoi tu vois** ‘you know’ (FR-conv-05)*

Both “quoi” and “tu vois” were annotated as *monitoring* in this example. Once more, *quoi* is well represented in this category (7/16). However, some same-function clusters are *hapax legomena* such as *as it were like, you know I mean* or *enfin voilà* ‘well that’s it’, which do not appear highly integrated: the top-down criterion of functional similarity does not seem systematically satisfactory. However, frequency alone is not a sufficient criterion either to decide on the low or high degree of integration for a particular pattern, since corpus data cannot be used as a direct mirror of entrenchment. For instance, in *DisFrEn*, only two occurrences of *ou sinon* ‘or else’ were found, even though the cluster meets several criteria for lexicalization (Crible 2015), including mutual pragmatic reinforcement (or functional similarity). If high frequency were a criterion for high integration, the following conversational DMs would be under consideration for compositional status (by decreasing order of frequency): *quoi mais* ‘you know but’, *et puis* ‘and then’, *et alors* ‘and then’, *and then* and *voilà quoi*, which all have more than ten occurrences (*et donc* ‘and so’ follows with nine cases). Two of these are complex DMs (*and then*, *et puis*) which are obvious matches for composition since it is no longer possible to distinguish a separate function for each component. By contrast, the different positions and scopes of the components in *quoi mais* seem somewhat contradictory with composition. In sum, functional similarity alone and high frequency alone do not always map a high degree of integration and fixation.

Another, more flexible way of extracting functionally similar co-occurring DMs is to select those expressing the *emphasis* function, which is precisely defined as “depend[ing] on another co-textual expression which it reinforces” in the annotation guidelines (Appendix 1: 352). In the conversational data, 11 such cases were found, mostly unique clusters such as *mais d’un autre côté* ‘but on the other hand’ (expressing *opposition*) or *you know sort of* (expressing *monitoring*). Most of these cases include a conjunction, either *and* or *but* (and their French equivalents *et* and *mais*), and thus contradict Cuenca & Marín’s (2009) typical patterns for composition which never include conjunctions but parenthetical or pragmatic connectives

instead. In their view, examples such as (49) should be categorized as addition instead of composition.

- (49) I always thought that she's she's taking the piss **but in fact** she's absolutely dead serious (EN-conv-01)

Here, the two DMs both express a relation of concession and it is not certain which one of them reinforces the other. This DM cluster is particularly promising because of its existence in French as well, either as *mais en fait* (not attested in the corpus) or *et en fait*, in spite of its rare frequency in *DisFrEn*. By comparison with same-function clusters (cf. Example (48) above), this more flexible understanding of functional similarity seems closer to an intuitive interpretation of high co-occurrence degree. Therefore, it seems that defining co-occurrence as redundancy is not particularly convincing when confronted to systematic corpus annotations, which rather reveal a higher role of complementarity, either in the form of emphasis (49) or underspecification (47). When taking frequency as a cue to mutual attraction, it appears that these cases of near-identity are closer to fixed complex DMs than fully redundant clusters. In other words, the hypothesis that frequent co-occurring DMs are functionally similar is only confirmed if we accept a more flexible sense of similarity as complementarity, which answers the research question at the onset of this analysis of co-occurrence.

To conclude, this study comparing corpus annotations with Cuenca & Marín's (2009) three-fold model has confirmed the existence of different degrees of co-occurrence, although their top-down definition does not necessarily map feature-based clusters extracted from the subcorpus of conversations. In particular, I showed that (i) there might be a missing intermediary level between addition and composition for cases such as *and so* or *et alors* and (ii) formal restrictions as provided in the original model are contradicted by a wealth of variation and counter-examples in the data. An interesting perspective to this study would be to test the cognitive reality of these different degrees by experimentally measuring their impact on cognitive processing. Another avenue, already suggested by Cuenca & Marín (2009), is to complement the syntactic-functional patterns with prosodic parameters, which might draw further distinctions or, on the contrary, merge several degrees or patterns together. Presently, the study of co-occurrence will be refined by analyzing the fluencemes in the co-text of DMs, which include a basic annotation of pauses (Chapter 6).

5.5 Summary and interim discussion: the potential of bottom-up research

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 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 favor 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, relationality, 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 multifunctionality 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.

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 4.1.2). 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 (cf. also the lack of speakers metadata, Section 4.1.1) 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 (Sections 5.1, 5.2) remain rather limited to a partial frequency-based portrait of the DM category, the integration of syntax and pragmatics in multivariate models (Section 5.3.4) proved more innovative and relevant to the investigation of form-function patterns undertaken in this usage-based research. The last section on co-occurrence was more interpretative, working on a smaller sample and trying to bridge the gap between systematic corpus annotations and qualitative categories in order to uncover the link between corpus-driven

patterns of co-occurrence and theoretical notions such as underspecification, interactivity and cohesion (a similar endeavor is undertaken in Chapter 7 for qualitative repair types). 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 presently considered on equal grounds with other factors (categorical variables and metadata) as potentially telling of underlying cognitive processes.

Beyond its descriptive and literature-confronting 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 (i) maintain a level of objectivity regarding the results, avoiding the circularity of “finding what one is looking for” and (ii) 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, the potential schemas (i.e. form-function patterns) identified so far will be refined and potentially qualified by the inclusion of another set of surface variables, namely the syntagmatic behavior of DMs with regard to their competitors in the fluencemes typology.

Chapter 6: The (dis)fluency of discourse markers: insights from the clustering of fluencemes

Introduction to the chapter

The results discussed in this chapter take the corpus-based profiles of DMs from Chapter 5 one step further and integrate them into a broader view of DMs as one type of fluencemes within the typology, along with pauses, repetitions or truncations for instance. The hypothesis of fluency-as-frequency lies at the core of the following analyses, looking for evidence in support of its cognitive validity as well as its limitations. Overall, this chapter can be described as a continuation of the investigation of discourse-level co-occurrence, moving from a within-category (cf. Section 5.4 on co-occurring DMs) to a between-category perspective: what can we conclude about the (dis)fluency of DMs on the basis of corpus frequency and clustering patterns? Do fluencemes clustered in a sequence show a similar degree of (dis)fluency? To what extent can such conclusions be generalized across languages, registers and degrees of granularity?

This chapter is structured as follows: starting with general observations on the clustering of fluencemes, Sections 6.1 – 6.3 focus exclusively on sequence-level variables (i.e. fluenceme types and various macro-labels) in order to describe the inter-relationship between members of the typology; Sections 6.4 – 6.6 are more fine-grained and integrate DM-level variables in order to draw interpretations of relative (dis)fluency in light of functional patterns; the main findings are summarized and discussed in Section 6.7.

6.1 Paradigmatic annotation of fluencemes 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 3.3.1). 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 (cf. Section 2.3.2). This tendency is expected to be largely due to the pervasiveness of unfilled pauses, as well as the high frequency of DMs in dialogues developed in the previous chapter. 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 2.3.4. Only then can we model the inter-relationships between each fluenceme type, identify recurrent patterns of combination and interpret these clusters in light of their features and distribution.

To meet such a paradigmatic programme, 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 of which was found to have a significant effect on the co-occurrence of DMs (cf. Section 5.4.1.1). 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. This section will therefore focus on identifying both language-specific and shared patterns across different degrees of abstraction or granularity, striving to test the hypothesis on fluenceme clustering and providing tentative interpretations of (dis)fluency.

6.1.1 Fluenceme rates

As explained in Section 4.4.1, the corpus-based extraction of fluencemes in *DisFrEn* is quite flexible insofar as the content of the sequences can be queried with varying degrees of granularity, both at fluenceme and sequence level. Starting with the former, 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 (see Table 6.1), 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).⁴⁸ 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 and disfluent devices alike, similar structures which are ambivalent (e.g. stuttering repetition vs. enumerating repetition), discourse markers signaling local cohesive relations and others

⁴⁸ 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.

more related to the interactive and spontaneous nature of speech, in line with the approach and annotation procedure designed in Crible et al. (2016). 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 which were previously found to be potentially more disfluent because of the heightened degree of speaker's attention towards their own production, coupled with the low degree of preparation (Broen & Siegel 1972; cf. Section 2.3.3). 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 (cf. above) 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.

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.

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

	English		French		Total	
	ftf	radio	ftf	radio	ftf	radio
UP	110.88	78.31	87.62	64.52	98.92	71.56
DM	62.86	54.60	71.88	52.16	67.50	53.41
FP	30.96	13.56	22.83	19.84	26.78	16.64
RI	11.84	16.98	14.96	17.94	13.45	17.45
TR	5.34	5.02	4.77	6.06	4.87	5.53
RM	4.98	3.53	5.82	6.18	5.58	4.83
FS	3.87	3.19	7.43	6.18	5.30	4.65
SP	3.05	1.94	2.94	2.26	3.39	2.09
SM	0.76	0.34	2.94	2.85	1.88	1.57
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

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 have been 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”.

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 face-to-face interviews than English (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, so much so that 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.

Regarding the effect of broadcasting on fluenceme frequency, we can observe an overall difference in favor of non-broadcast (face-to-face) interviews in both languages. However, this difference does not affect all fluencemes equally. Unfilled pauses are more frequent in face-to-face interviews in the two languages (41% more frequent in English, 35% in French; LL = 99.99, $p < 0.001$). A similarly large difference is found for DMs (LL = 37.48, $p < 0.001$), especially in French. Filled pauses are twice as frequent in English non-broadcast as 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 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.⁴⁹

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 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 to the present study.

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. 2017b) where the same unique typology was applied to data in native French, native and learner English and Belgian French Sign Language by four different annotators. We found that relative frequencies of individual fluencemes are highly similar, 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 between 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 (cf. Section 2.2.1).⁵⁰ We can reformulate this finding in terms of (non-)linearity: non-linear processes are omnipresent in speech production, covering about one fifth of the sound signal, and such momentary interruptions of the linear unfolding of speech mostly attend to the upcoming rather than previous material.

⁴⁹ 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.

⁵⁰ Cf. also Levelt’s (1983) “covert repairs”.

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 their ambivalent definition. 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 should 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 above, thus resorting 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 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 rather contrasted 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 (see Chapter 7).

6.1.2 Sequence length and most frequent clusters

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 2.3.2) 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), which 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.

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

	face-to-face				radio			
	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

We see that the bulk of sequences in the data include up to three fluenceme tokens (together more than 90% of all sequences), 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. 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, although the observed differences between each subcorpus remain valid and in line with previously obtained results.

It now remains to uncover what specific clusters lie behind this numeric information and 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 (cf. “reduced sequence”, Section 4.4) and leave more abstract categories or macro-labels for tentative interpretations of relative fluency in the next subsection. 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. This 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) will be the focus of Section 6.3.

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).

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.

A last observation at this level 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" of 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 fluenceme, 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 qualify 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 (i) the general ranking of sequences, (ii) the proportion of isolated and clustered fluencemes and (iii) the difference of relative "independence" between DMs and other fluencemes.

6.1.3 Fluency-as-frequency across different degrees of granularity

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 their fluency 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 degree(s) of abstraction better fit(s) the fluency-as-frequency hypothesis.

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 (i) the structure of the fluenceme(s) (simple or compound), (ii) the number of fluencemes (one or multiple) and (iii) 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 could 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” reformulations. 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, with only three significant contrastive differences marked by stars in

the table, and will therefore not be commented any further.⁵¹ 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 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)

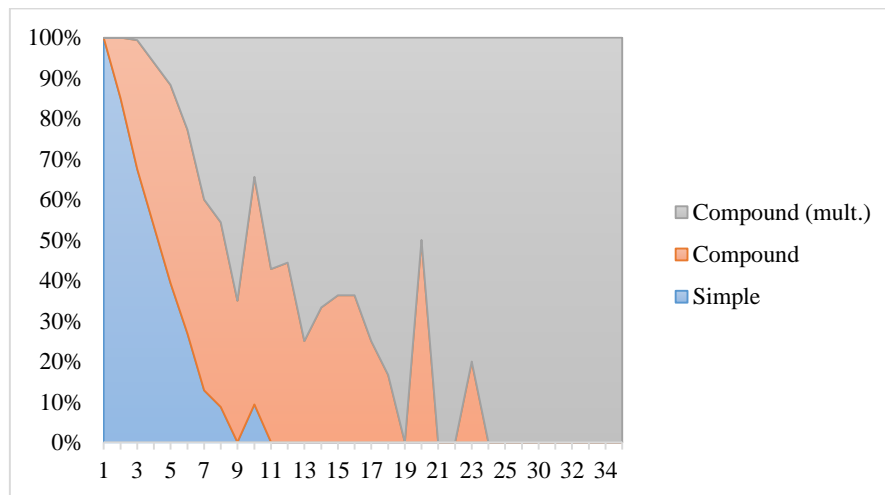
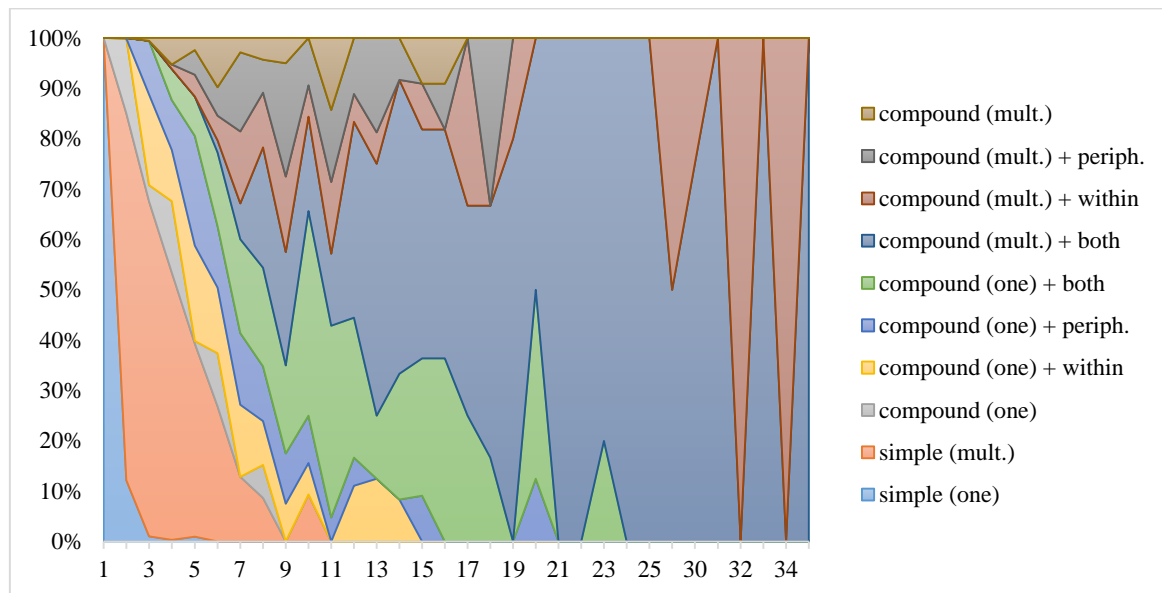
The truncation of “savaient” in (3) and the substitution of “in” by “with” in (4) generate non-linear retracing which involves partial repetition of anteposed (“ils”) or postposed material (“all their”). Repeating available linguistic material has been experimentally shown to be generally associated 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 remains 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:

⁵¹ The notation system is as follows: * for $p < 0.05$; ** for $p < 0.01$; *** for $p < 0.001$; no star when $p > 0.05$. These LL statistics were computed for the difference between languages for each type of sequence structure, both settings combined.

- (5) il raconte une histoire **ehh (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 qualification 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 strike as particularly disruptive despite their very low frequency. In terms of abstraction degrees, we can say that the apparent association between frequency and structural complexity suggested by Table 6.4 with rather broad categories is somewhat qualified by zooming into specific annotation labels (e.g. RM+TR) and actual instantiations of sequences.

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 degree of granularity of the observations 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 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.

Figure 6.1: Proportions of sequence type (coarse-grained) by sequence length**Figure 6.2:** Proportions of sequence type (fine-grained) by sequence length

From the coarse-grained approach (Figure 6.1), we see that very short sequences represent the quasi-monopoly of simple fluencemes, which become inexistent after 10-word sequences. The small rise of the blue area at this spot (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”) 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”).⁵² 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 linearity, we no longer see an effect of stagnation and interruption, which is instead replaced by an elaborate interrogative structure built on a repetition for a contrastive construction opposing “mieux” to “moins bien”. 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.

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, brown; multiple compound with peripheral fluencemes, dark grey) 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 blue and red), 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

⁵² The conjunction “et” (‘and’) is not annotated as a DM in this example because it functions intra-sententially by connecting two utterance-internal object complements.

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
 <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 (FR-intf-03)*

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 linearity, 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 the pattern in light green (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 (see the approach in Chapter 7). 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 section, 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.

6.2 DM-based sequences across registers

This section will test the hypothesis according to which unplanned discourse should lead to more frequent and more varied fluencemes than planned speech, while intermediary registers should be more similar to spontaneous dialogues. 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. It is hypothesized that sequences which are specific to informal situations should be typically disfluent, while sequences shared across all registers should be more ambivalent. The explanatory power of register variation will sometimes be complemented by the secondary metadata system describing each register in terms of situational features such as degree of preparation or number of speakers, when such an alternative approach is relevant. 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 6.5 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 (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. 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.

Table 6.5: Relative frequency of DM-based sequences ptw in *DisFrEn*

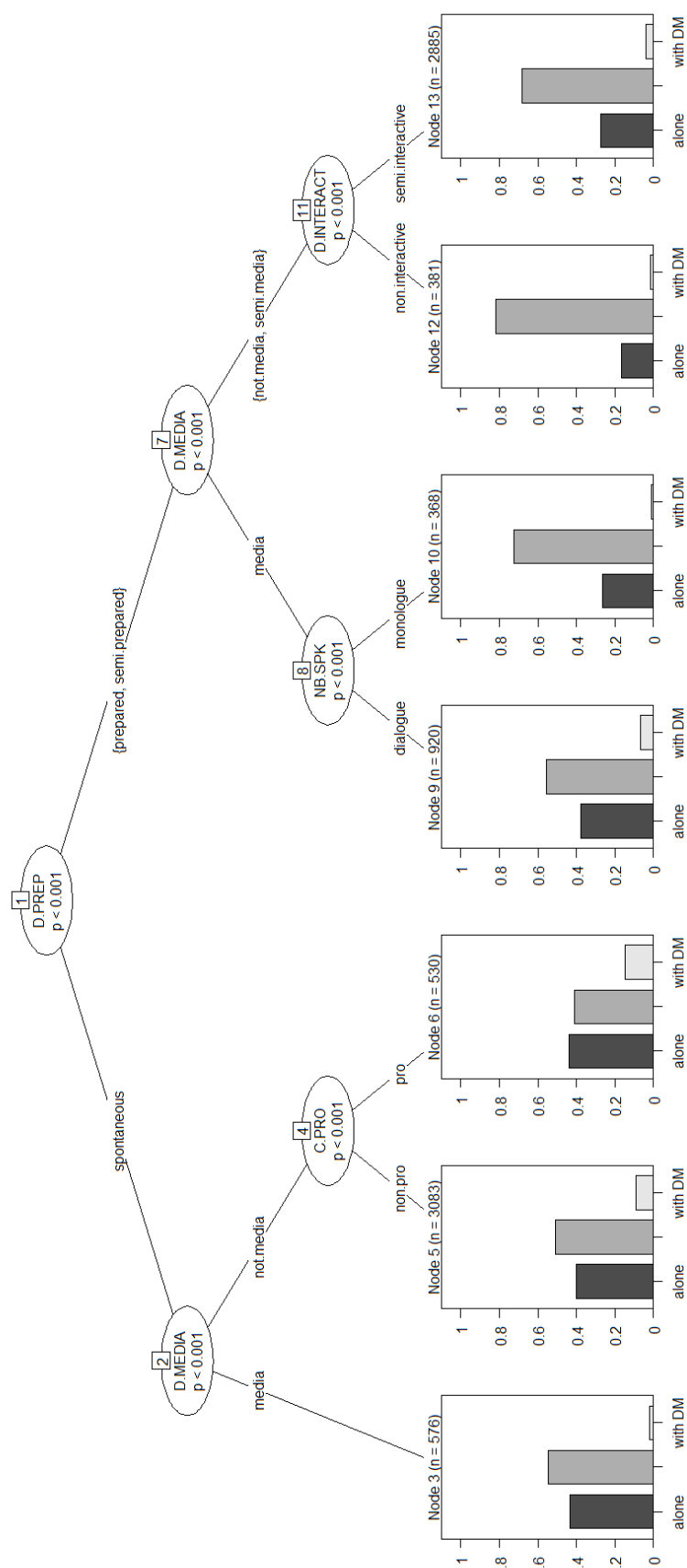
	English	French	Total
conversation	46.11	66.83	56.46
phone	52.73	62.66	56.81
interview	53.30	55.87	54.62
radio	47.87	42.78	45.38
classroom	43.93	39.48	42.67
sports	37.76	36.47	37.20
political	21.04	18.79	19.97
news	13.62	16.35	14.96
Total	42.26	47.71	44.80

To identify the specific clusters of DMs and fluencemes behind this frequency table, the following analysis investigates fluenceme sequences according to three degrees of granularity, ranked by decreasing order of abstraction:

- 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 6.1.3.

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. Starting with the most coarse-grained degree of abstraction (“cluster”), 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 6.3 reports on a conditional inference tree (a type of decision tree 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 abbreviations 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.

Figure 6.3: Conditional inference tree for isolated, clustered and co-occurring DMs

Turning to the second degree of granularity (“sequence category”), 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 6.4 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”).⁵³ 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.

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 answer to 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 6.5, where differences in proportions are shaded according to the statistical significance of Pearson’s 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, less 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 5.5). 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. 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).

⁵³ This first grouping only partially matches the overall frequency of DMs in the corpus as shown in Chapter 5, where DMs were not distinguished according to their clustering pattern (isolated or clustered with other fluencemes). In fact, when both isolated and clustered DMs are combined, their overall frequency is higher in face-to-face interviews, as opposed to the results in Figure 6.3 where clustering impacts the ranking, with more isolated DMs in radio than face-to-face interviews.

Figure 6.4: Conditional inference tree for sequence category by register

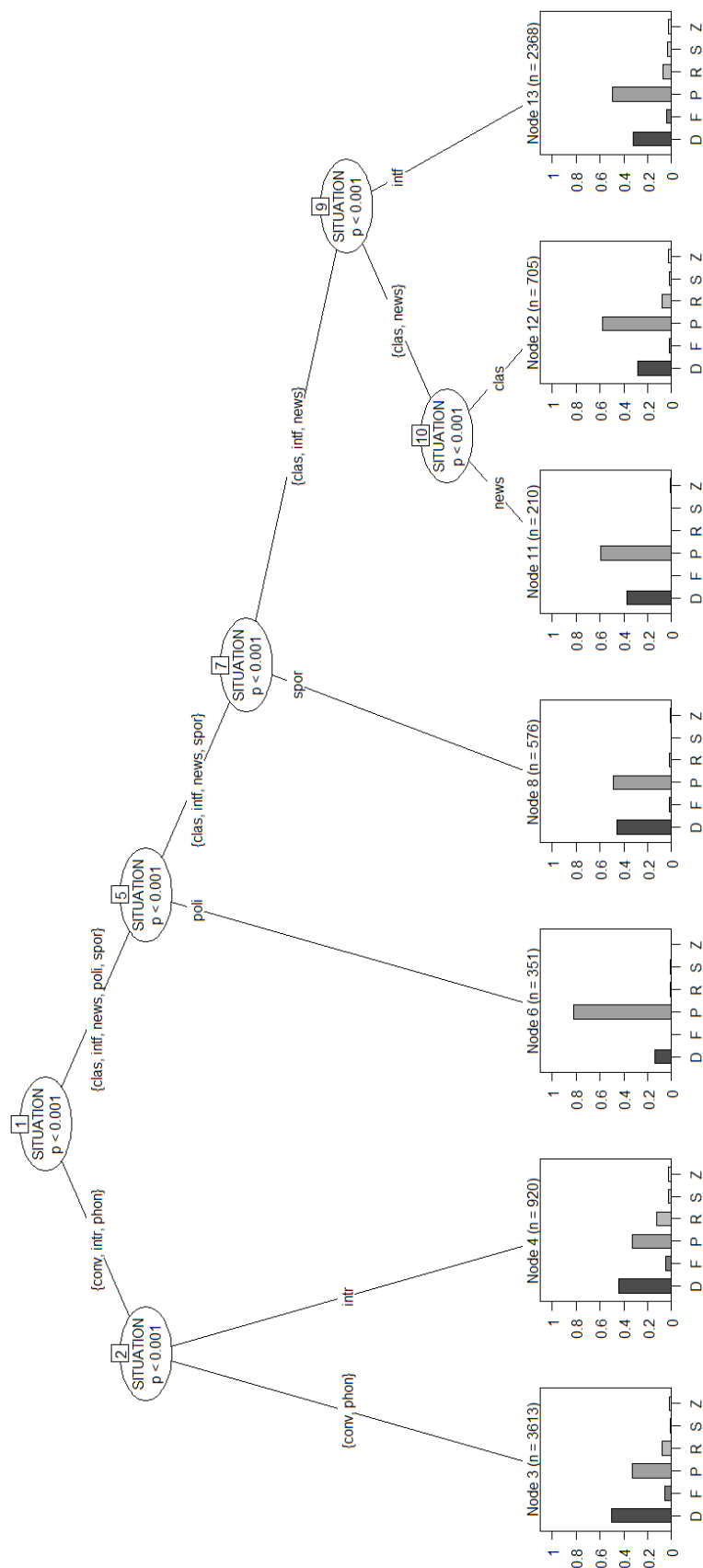
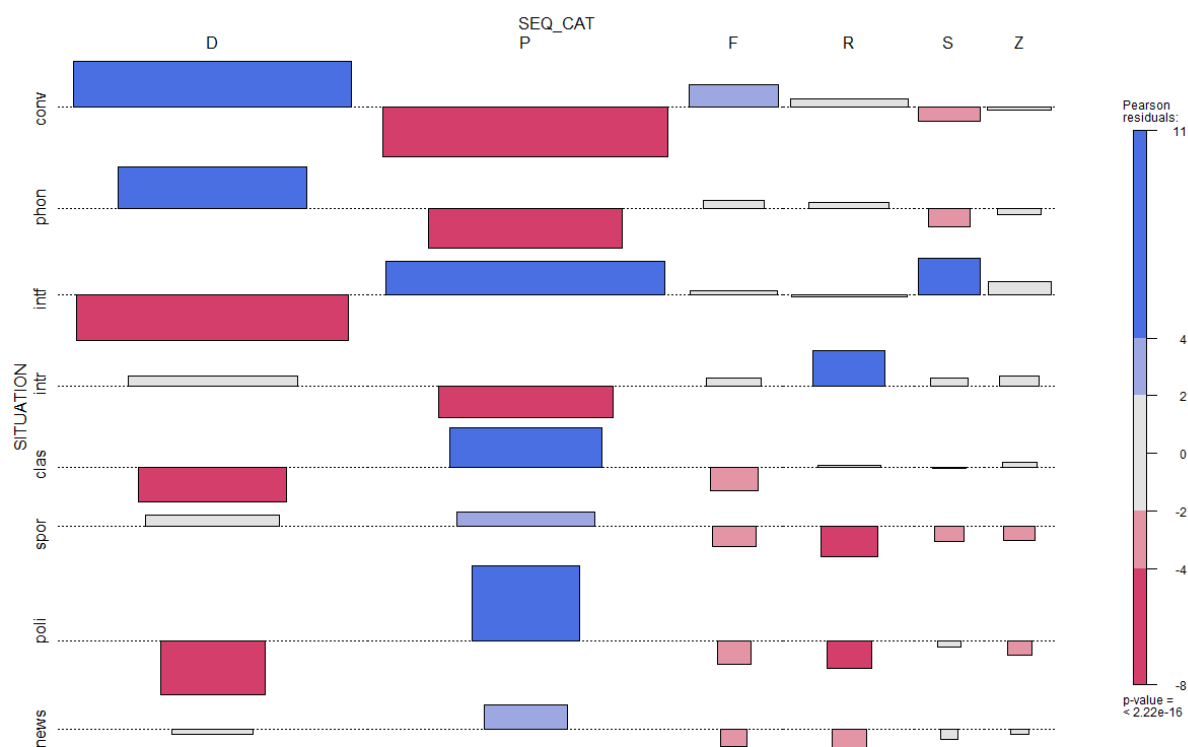


Figure 6.5: Extended association plot between sequence categories and registers

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 6.1.3) 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, so much so that 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 functional interpretation of authentic examples only partially confirms this conclusion. The specificity of F-sequences to conversations and its 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 (11).

- (11) 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 should 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 6.5) 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.

- (12) and you know **is it going to look like dad is it going to look like mum** (EN-intf-03)
- (13) **well I wasn't driving well I was driving** partly on the road but also on through open country (EN-phon-02)
- (14) 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 (12), 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 (13) 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 (14), 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 (14) 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 relatively to other registers, and its specificity to (or absence from) some settings which might be a better indicator of its relative fluency (here, relatively more disfluent in phone calls).

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:

- (15) 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)

- (16) **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 (15) 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 strike as strongly marked or disruptive, although perceptive ratings would be necessary to assert such a conclusion. Overall, such mixed sequences should be particularly marked in registers which are restricted in their diversity of sequence types and where they are significantly less frequent than in others, in other words, in broadcast formal registers. In fact, this specificity 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.

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 6.5. 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 6.1.3. 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 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 results from the paradigmatic annotations (Section 6.1.3) of this research.

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;
- *disons* (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 specificities 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 specificity or restriction in sequence types. This line of investigation will be further pursued with the integration of functional annotations in Sections 6.4 – 6.6. 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. In terms of clustering tendency, the strong association between DMs and pauses, as well as their high frequency, leads me to focus on their patterns of combination in greater detail in the next section.

6.3 From isolation to combination: focus on DMs and pauses in interviews

The very high frequency (and quasi-monopoly) of DMs and pauses within the fluenceme typology calls for a deeper investigation of the mechanisms behind their clustering tendencies. Their relative prominence compared to other fluencemes is here taken as a sign of their greater functional ambivalence and non-optionality, which is precisely the reason why they are so often excluded from disfluency typologies and annotation schemes (cf. Section 2.2.3.2.1). Such a restriction, however, overlooks an important aspect of the phenomenon as well as a very large portion of the data. The present research addresses this gap in the literature and the analyses in this section focus on the three most frequent fluencemes in *DisFrEn*, namely unfilled pauses (UPs), discourse markers (DMs) and filled pauses (FPs).

The results reported in this section are largely based on a previous study (Crible et al. 2017a) which is here replicated on the subcorpus of face-to-face interviews from the *DisFrEn*

dataset. This choice is motivated by (i) the absence of register effects found in the original paper and (ii) the paradigmatic annotations available in the interview data which make it possible to include more occurrences, especially regarding the systematic annotation of the position of pauses (instead of a sample analysis in the original paper). This analysis pursues a double objective: firstly, to study the effect of clustering on the behavior of DMs and pauses and, secondly, to further our understanding of the phenomenon of discourse-level co-occurrence already tackled in Section 5.4. In line with the general approach of this research to explore corpus-driven patterns through various degrees of abstraction, the present analyses will mostly make use of the macro-labels designed specifically for [DM+pause] clusters defined in Section 4.4.3. This section follows the original structure from Crible et al. (2017a), starting with observations of DMs and FPs used independently from each other, compared to their clustered contexts of use and, lastly, striving towards a statistical model encompassing linguistic variables impacting the clustering of DMs and pauses.

6.3.1 Independent uses of DMs and FPs

The contrastive perspective of the present thesis, as well as its relative independence from prosodic considerations, motivates a focus on filled pauses, which are vocalized elements (even lexical, according to some authors, e.g. Clark & Fox Tree 2002; Corley & Stewart 2008; Tottie 2015a) showing different forms across languages, as opposed to the purely temporal unfilled pauses. DMs and FPs will each be investigated in isolation (only element in the sequence) and in combination with unfilled pauses (and nothing else in the sequence).

6.3.1.1 DMs and DM+UP patterns

In face-to-face interviews, 665 isolated DMs were annotated, out of the 2,369 total DMs in this subcorpus (regardless of clustering): 305 in English (17.88 ptw), 360 in French (19.95 ptw), in non-significantly different frequencies (LL = 1.98, $p > 0.05$) for a total of 18.95 isolated DMs ptw, or 28% of all DMs in interviews. The five most frequent expressions within this isolated set are *and*, *so*, *actually*, *then*, *because* in English and *et* ‘and’, *mais* ‘but’, *parce que* ‘because’, *donc* ‘so’ and *ben* ‘well’ in French. We see that the two languages share a number of common expressions (*and* / *et*, *so* / *donc*, *because* / *parce que*) which are also among the most frequent DMs overall in *DisFrEn*. The other, less typical expressions (*actually*, *then*, French *ben*) rank high in the interview data, although their frequency is lower in *DisFrEn* on the whole, which points to preferences in the selection of DMs which do or do not occur alone.

Table 6.6 reports the proportion of isolated DMs across positions in the three different annotation systems, namely micro-syntax, macro-syntax and turn of speech (cf. Section 4.2.1.3). It appears that isolated DMs follow their general distribution (i.e. when isolated and clustered uses are combined, cf. Section 5.2.1): most frequent in initial position of both micro- and macro-syntactic units, in similar proportions across English and French. Some differences arise for less typical micro-syntactic positions, such as the reversed proportions for medial and final DMs between English and French. The French data is also characterized by a higher

proportion of turn-initial and turn-final DMs than in English. These results will be compared to the occurrences of DM+FP clusters in Section 6.3.2.

Table 6.6: Proportions of isolated DMs across positions

		English	French	Total
Micro	Initial	74.10% (226)	75.77% (272)	75.00%
	Medial	19.34% (59)	9.47% (34)	14.01%
	Final	5.57% (17)	14.21% (52)	10.24%
	Independent	0.98% (3)	0.56% (2)	0.75%
Macro	Pre-field	50.82% (155)	52.09% (188)	51.51%
	Left	11.48% (35)	9.47% (34)	10.39%
	Middle	4.26% (13)	0.56% (2)	2.26%
	Right	29.51% (90)	23.68 (85)	26.36%
	Post-field	2.95% (9)	13.65% (49)	8.73%
	Independent	0.98% (3)	0.56% (2)	0.75%
Turn	Initial	7.21% (22)	23.12% (83)	15.81%
	Medial	90.16% (275)	69.36% (250)	78.92%
	Final	1.64% (5)	6.96% (25)	4.52%
	Independent	0.98% (3)	0.56% (2)	0.75%

Regarding their co-occurrence with unfilled pauses (and nothing else), 534 clusters were extracted including either DM+UP or UP+DM and no other fluenceme (i.e. these are all clusters of two tokens and types). The former order (DM+UP) is much less frequent than the latter (83 vs. 451) in quite similar proportions in English and French (14% and 20%). There is, however, a significant difference in favor of English clusters ($LL = 4.5$, $p < 0.05$) which reminds us of the higher frequency of UPs in English overall (cf. Section 6.1.1). DMs alone and DM+UP clusters are almost equally frequent in English (52% of isolated) while French shows a slight preference for the isolated contexts (59%).

Compared to isolated contexts, DMs clustered with unfilled pauses occur in higher proportions in initial position (88% vs. 75% in micro-syntax; 70% vs. 52% in macro-syntax), as can be seen in Table 6.7. This difference points to the segmentation, boundary-marking role of unfilled pauses. Regarding the position in the turn, occurrences of DMs and unfilled pauses anywhere other than turn-medially become extremely rare, with an overall proportion of 98% of turn-medial clusters. This is due to the higher frequency of the UP+DM pattern which, by definition, cannot occur turn-initially since a pause cannot be considered as the first (or last) element of a turn. Absence of sound signal between two turns of speech is not attributed to any speaker since it would necessarily resort to an arbitrary decision from the analyst. Less frequent positions, as well as co-occurring DMs, show too few occurrences to be reliably compared with the isolated uses.

Table 6.7: Proportions of DMs with unfilled pauses across positions

		English	French	Total
Micro	Initial	89.44% (254)	86.80% (217)	88.20%
	Medial	6.34% (18)	3.60% (9)	5.06%
	Final	3.17% (9)	8.40% (21)	5.62%
	Independent	1.06% (3)	1.20% (3)	1.12%
Macro	Pre-field	77.11% (219)	62.80% (157)	70.41%
	Left	8.10% (23)	10.80% (27)	9.36%
	Middle	1.06% (3)	0.40% (1)	0.75%
	Right	10.21% (29)	17.20% (43)	13.48%
	Post-field	2.46% (7)	7.60% (19)	4.87%
	Independent	1.06% (3)	1.20% (3)	1.12%
Turn	Initial	1.14% (4)	0.80% (2)	1.12%
	Medial	97.89% (278)	98.40% (246)	98.13%
	Final	0.70% (2)	0.80% (2)	0.75%

6.3.1.2 FPs and FP+UP patterns

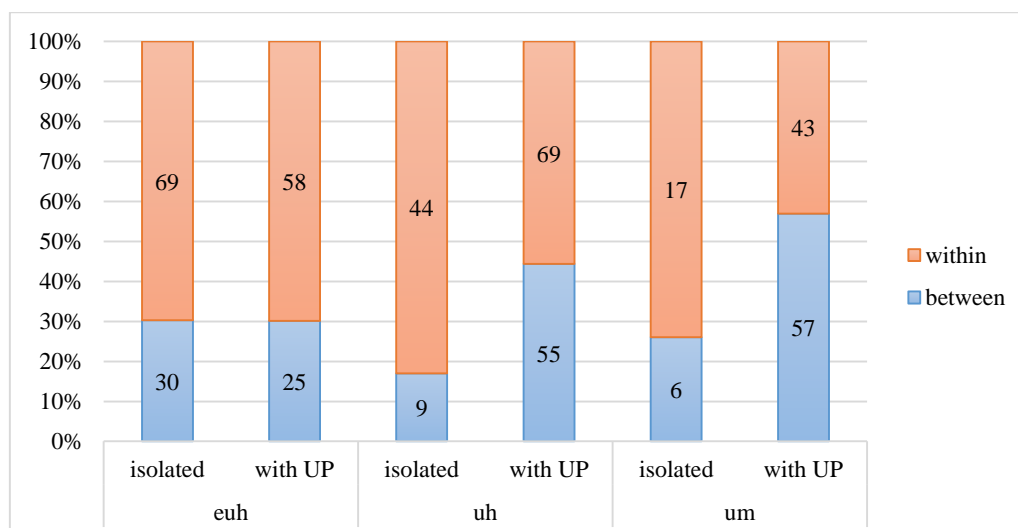
Crosslinguistic investigations of filled pauses are much less common than in the field of DM studies. Yet previous studies have shown some variation either in the form or use of FPs across languages such as English, Swedish, Mandarin Chinese, French or Spanish (see Eklund & Shriberg 1998; Zhao & Jurasfky 2005; Vasilescu et al. 2007). Shriberg (1994) and Clark & Fox Tree (2002), in particular, claim that the two major English FPs *uh* vs. *um* show positional preferences for utterance-internal and utterance-initial uses, respectively (i.e. within or between utterances). In the present interview data, 940 FPs have been annotated, including only 175 isolated ones: 76 in English and 99 in French (LL = 1.88, $p > 0.05$), which amounts to 4.46 and 5.49 FPs ptw respectively. English FPs can be further distinguished according to their vocalic form: *uh* is the most frequent form in isolation, with 53 occurrences against 23 for *um*.

When clustered with unfilled pauses (and nothing else), we find 307 occurrences of either order (UP+FP or FP+UP). These pause clusters are much more frequent in English than in French (224 vs. 83, LL = 75.44, $p < 0.01$), which is additional evidence of the higher weight of UPs in English, as already pointed out by Grosjean & Deschamps (1975). These results are summarized and cross-tabulated with the binary positioning system (within vs. between utterances) in Table 6.8. Compared to the effect of unfilled pauses on DMs, it appears that, in English, FPs are much more frequently clustered with UPs than isolated (52% of isolated DMs vs. 25% of isolated FPs). In French, however, the proportions are similar (59% and 54% of isolated DMs and FPs respectively). This higher attraction between filled and unfilled pauses in English tends to argue against the lexical status of FPs suggested by some authors (e.g. Tottie 2015a). DMs appear more independent of UPs than FPs, which confirms the categorization of FPs as pauses along with UPs, and not as words like DMs.

Table 6.8 Filled pauses (with and without UP) by position and language

		Between	Within	Total	%
English	FP without UP	20% (15)	80% (61)	76	25%
	FP with UP	50% (112)	50% (112)	224	75%
	Total	42% (127)	58% (173)	300	100%
French	FP without UP	30% (30)	70% (69)	99	54%
	FP with UP	30% (25)	70% (58)	83	46%
	Total	30% (55)	70% (127)	182	100%
Total	FP without UP	26% (45)	74% (130)	175	36%
	FP with UP	45% (137)	55% (170)	307	64%
	Total	38% (182)	62% (300)	482	100%

Table 6.8 also shows the positional preferences of FPs with and without unfilled pauses in English and French. We see that isolated FPs largely prefer the utterance-internal (or “within”) position in the two languages, which is also the case for the FP+UP clusters in French only (70%), while English FP+UP clusters are perfectly evenly distributed across the two positions (between and within utterances). The different proportions of utterance-initial (or “between”) positions between isolated (20%) and clustered FPs (50%) in English might indicate an attraction to boundaries when UPs are present, an effect which was also observed for DMs in the previous subsection. Overall, these findings on position are consistent with the study in Crible et al. (2017a) carried out on additional registers of English and French. However, they do not fully map the behavior of Dutch FPs analyzed by Swerts (1998), who found a systematic preference for boundaries when FPs cluster with UPs. Here, the “between” and “within” positions are equally filled by the English clusters with no particular preference (although there is indeed a rise in the proportion of utterance-initial uses from isolated to clustered FPs), while this effect is not observed at all in French. This situation is refined by looking at the specific FP form in Figure 6.6.

Figure 6.6: Patterns of position and clustering for *euh*, *uh* and *um*

We see that the unique French form *euh* always prefers utterance-internal position, as in Example (17). The English forms *uh* and *um* show a very similar trend, with much more “within” than “between” position without a UP and a more balanced distribution in clusters with a UP. However, we can see that occurrences of isolated *um* are rather scarce (23/123). All in all, *uh* always prefers internal positions as in Example (18), which corroborates previous works (Shriberg 1994; Clark & Fox Tree 2002), while *uhm* is more balanced between initial (with a pause, Example 19) and medial (without a pause, Example 20), although the latter pattern is less frequent overall.

- (17) il vous enverra le bouquin avec **euh** un bulletin de versement vous n’aurez qu’à payer
he will send you the book with euh a deposit slip you will just have to pay (FR-intf-03)
- (18) my role in this area then is to ensure that **uh** (0.660) we generate income (0.540) to the
 university (EN-intf-05)
- (19) and I’ve actually worked for different law firms (0.180) **um** (0.433) the accountancy
 firm I think it’s quite different (EN-intf-08)
- (20) so what is your **um** role here as manager what sort of things do you take care of (EN-
 intf-04)

In Crible et al. (2017a), we suggested that FP+UP clusters could be interpreted as “signals” rather than “symptoms” for lexical salience or other attention-getting purposes, even in utterance-internal positions. The examples (17)-(20) indeed support this view, although no interpretation of (dis)fluency could be reliably based simply on the position and type of cluster an FP occurs in. Utterance-internal FPs as in Examples (17), (18) or (20) might go unnoticed by the hearer, serve information-structuring functions or be involved in pragmatic face-saving strategies. In particular, the binary positioning system presently adopted seems insufficient to account for finer distinctions of boundaries between smaller units such as complement or noun phrases which can be chunked and marked by FPs. Such conclusions would require discourse parsing and more data than what is available here.

6.3.2 DM, FP, UP: the impact of clustering

We can now extend the previous analyses (co-occurrence with unfilled pauses and positional preferences) to DMs and FPs in combination in order to identify the effect of their clustering. In doing so, we will be able to test Beliao & Lacheret’s (2013) result on the “independence” of prosodic disfluencies such as pauses with respect to DMs and see which of the two fluencemes (DMs or FPs) attracts the other. From the 2,369 DMs extracted from face-to-face interviews, 249 are clustered with a FP, including sequences with unfilled pauses as well (e.g. UP+FP+DM) and no other fluenceme (i.e. P-sequences only). These clusters of DMs, filled and unfilled pauses amount to 323 if we include sequences with other fluencemes such as repetitions. The different configurations and their distribution per language are reported in Table 6.9, where we see that DMs are much more frequent in isolation than in combination with FPs, while FPs are more frequent in combination with DMs than in isolation.

Table 6.9: Distribution of [DM+pause] clusters in face-to-face interviews

	English		French		Total	
MIX – DM with 2+ pauses	47	35.88%	18	15.25%	65	26.10%
UFL – UP+FP+DM	54	41.22%	4	3.39%	58	23.29%
FPL – FP+DM	6	4.58%	27	22.88%	33	13.25%
FPR – DM+FP	0	0%	30	25.42%	30	12.05%
UDF – UP+DM+FP	10	7.63%	17	14.41%	27	10.84%
FUR – DM+FP+UP	4	3.05%	11	9.32%	15	6.02%
UFR – DM+UP+FP	8	6.11%	4	3.39%	12	4.82%
FUL – FP+UP+DM	2	1.53%	6	5.08%	8	3.21%
FDU – FP+DM+UP	0	0%	1	0.85%	1	0.40%
Total	131	100%	118	100%	249	100%

A number of crosslinguistic differences appear from this table. The two most frequent cluster types in each language are very different, with complex clusters of both UPs and FPs (“MIX” and “UFL”) in English as opposed to simple combinations of DMs and FPs (in either order) in French, while these respective patterns take up rather low proportions in the other language. This result points to a preference for UP in English and for FP in French. The most frequent pattern in each language (UFL in English and FPR in French) is illustrated in the following examples:

(21) time to let uh (0.320) somebody else take that burden on **(0.390) uh and** my role as a senior sister is to coordinate the shifts (EN-intf-03)

(22) on est plutôt un groupe euh anglo-saxon et scandinave **donc euh** je suis le seul francophone dans le groupe
we are rather uh an Anglo-Saxon and Scandinavian group donc euh ‘so uh’ I am the only French-speaking one in the group (FR-intf-01)

Overall, we see that patterns including an unfilled pause are more frequent than without an unfilled pause, especially in English where they amount to 95% of all clusters in the table, against 52% in French. This higher attraction between the two types of pauses than between DMs and FPs is, in my opinion, yet another argument for the categorization of FPs as pauses instead of words, following the cognitive assumption that what frequently co-occurs in the speech string is categorized together in the speaker’s lexicon.⁵⁴ In addition, the higher weight of FPs in French corroborates our previous conclusion in Crible et al. (2017a) suggesting that FPs are much more common in French than in English, an observation which was supported by register information: we found that isolated FPs are very rare in the formal settings of news broadcasts and political speeches in English, whereas they do occur in a substantial quantity in French planned discourse.

⁵⁴ Recently, Tottie (2015b) found occurrences of filled pauses in computer-mediated writing, which might argue for reconsidering the categorization of FPs as pauses. An interesting research avenue in this regard would be to further investigate [DM+FP] clusters in multimodal discourse.

The most frequent DM expression in English is invariably *and* across all configurations except for two rare patterns (*actually* in DM+UP+FP; *since/if* in FP+UP+DM), followed by *so*, *but* and *you know*. In French, the basic *et* is only the most frequent in the FP+DM and UP+DM+FP patterns, with no clear preference between *donc*, *mais* or *alors* for the other patterns. Overall, the top five DM expressions in clusters with FPs do not strongly differ from the top five for isolated DMs, as opposed to our finding in Crible et al. (2017a) where we found a constraining effect of the presence of FPs, restricting DMs to generic conjunctions. The specific settings of interviews might be responsible for this difference.

Isolated and clustered DMs are further compared in terms of position preferences, as reported in Table 6.10. Compared to the distribution of isolated DMs (cf. Table 6.6), we see that [DM+FP] clusters follow the same ranking (i.e. more frequent in clause-initial, pre-field and turn-medial positions) albeit in different proportions.

Table 6.10: Proportions of [DM+FP] clusters across positions

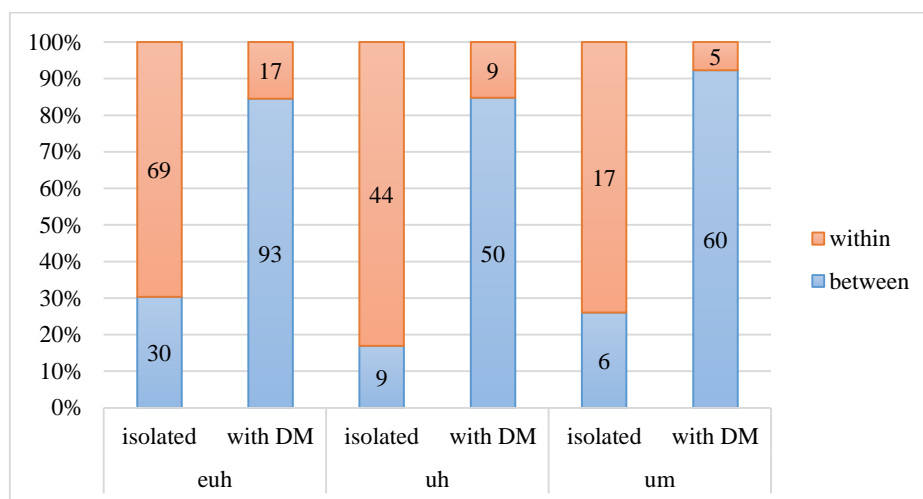
		English	French	Total
Micro	Initial	86.26%	82.2%	84.3%
	Medial	6.87%	8.47%	7.63%
	Final	2.29%	8.47%	5.22%
	independent	4.58%	0.85%	2.81%
Macro	pre-field	70.99%	62.71%	67.07%
	Left	6.87%	13.56%	10.04%
	Middle	1.53%	0%	0.80%
	Right	13.74%	16.1%	14.86%
	post-field	2.29%	6.78%	4.42%
	independent	4.58%	0.85%	2.81%
Turn	Initial	2.29%	1.69%	2.01%
	Medial	97.71%	96.61%	97.19%
	Final	0%	1.69%	0.80%

The major increase concerns French turn-medial DMs (from 69% to 97%), leaving very few occurrences for turn-initial and turn-final clusters. As in Crible et al. (2017a), the *donc euh* ‘so uh’ cluster is still prominent in final position, thus confirming Degand’s (2014) finding, according to which *donc euh* is emerging in spoken French as a “typically (turn) final pattern which appears to express a specific meaning, namely that of a conclusive relation that the addressee is invited to infer” (2014: 169), as illustrated in Example (23).

- (23) <VAL_4> je ne connais pas beaucoup d’hommes provenant de Bastogne **donc euh**
 <VAL_2> mm (2.980) et dernière question quant à cette opinion
 <VAL_4> *I don’t know many men from Bastogne* **donc euh** ‘so uh’
 <VAL_2> mm (2.980) and last question about this opinion (FR-intf-02)

With *donc euh*, the speaker is inviting the interviewer to understand that his previous statement about “men from Bastogne” might be of limited value, and that she can move on to another topic. These turn-final cases are, however, very rare in the interviews data, as opposed to the bulk of turn-medial and utterance-initial clusters. As for the position of FPs, the medial “within” position becomes much less frequent in clusters than in isolation, regardless of the FP form, as can be seen in Figure 6.7.

Figure 6.7: Patterns of position for FPs with and without a DM



Compared to Figure 6.6 where we saw that the “within” position was preferred by *euh*, *uh* and *um* in isolation, we see here that the proportions of “within” vs. “between” are systematically reversed (i.e. more “between”) when FPs cluster with DMs, which seems to show that DMs attract FPs to their positional preferences, and not the reverse, against what Beliaio & Lacheret (2013) suggested about the relative independence of prosodic hesitations from DMs. With respect to the fluency-as-frequency hypothesis, it is tempting to suggest that DM+FP clusters would be more disruptive within utterances than at boundaries given that they interrupt their host unit in a sequence of at least two fluencemes (three or more if UPs are involved), as in Example (24).

- (24) c’est certain que l’environnement est un élément **euh (0.393) je dirais (1.610)** très important
it is certain that environment is an element uh (0.393) je dirais ‘I would say’ (1.610)
 very important (FR-intf-01)

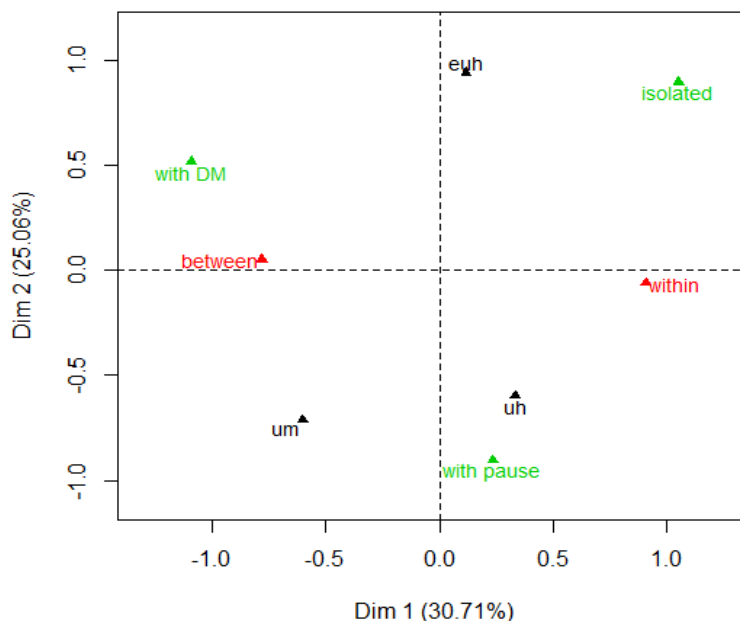
The “lexical search” meaning of the DM *je dirais* coupled with the FP and two UPs might suggest a rather disfluent reading of this example. However, in the absence of perceptive validation, more strategic interpretations cannot be excluded, for instance as an emphatic function (to stress the evaluative “very important”). To go any further, the analysis would need to take into account pause duration in these clusters, in addition to finer information on discourse segmentation mentioned earlier. Overall, we are able to confirm an impact of clustering on the position of DMs and FPs, which allows us to suggest (i) the stronger association between FPs and UPs than between FPs and DMs and (ii) the attraction of FPs to the syntactic behavior of DMs and, by extension, the relative independence of DMs from FPs.

6.3.3 Modeling the clustering of DMs and pauses

It now remains to summarize these results in a statistical model of the clustering of DMs and pauses. A binomial logistic regression was computed to predict the presence or absence of DM in the vicinity of a FP, with language, position and presence of a UP as input factors. The model is quite robust ($C = 0.795$, $r^2 = 0.339$) and returns significant positive effects of French and UPs on [DM+FP] clusters, whereas the “within” position negatively affects the combination of DMs and FPs. In other words, DMs and FPs often cluster together in French, with unfilled pauses and between utterances.

These results can be visualized in the multiple correspondence analysis (MCA) in Figure 6.8. MCA represents the attraction between levels of variables by their proximity on the graph: the closer the values, the more frequently they occur together in the data. The percentages on the axes correspond to the amount of variance in the data covered by the graph. Here, 55% of the variance is explained, which is substantial. The most interesting observations from this figure are the attraction of *uh* to unfilled pauses and the “within” position (bottom-right quadrant), and the attraction of [DM+FP] (“with DM”) to the “between” position (top-left).

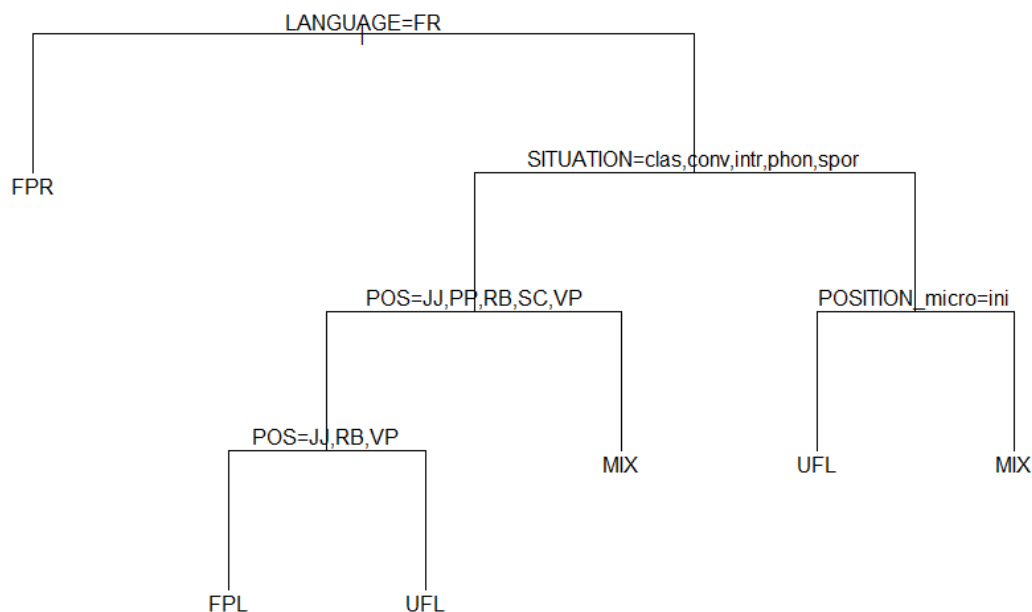
Figure 6.8: Multiple correspondence analysis of the clustering of FPs with DMs



Both the logistic regression and the MCA were computed solely on the face-to-face interview data with a limited set of variables. The classification tree in Figure 6.9 addresses this limitation by including all [DM+pause] clusters from *DisFrEn* and additional variables such as register, POS-tag and functional domain of the DM. As explained in Section 5.3.4, classification trees are used to predict an outcome (here, the type of DM+pause cluster) based on a number of factors. Values on top of each node are associated to the branch to their left. Classification trees

are not able to account for rare values and only report main significant associations after “pruning”.

Figure 6.9: Pruned classification tree of [DM+pause] configurations



We see that French DMs are clearly associated with the “FPR” (DM+FP) configuration (cf. the *donc euh* pattern), while English DMs seem to be further affected by register variation, grammatical class (POS) and position. For instance, the “UFL” (UP+FP+DM) configuration is attracted to the initial (“ini”) position in English face-to-face interviews as well as to adjectives (“JJ”), adverbs (“RB”) and verb phrases (“VP”) in all other registers, while the “MIX” (e.g. UP+DM+FP+UP) can be found in two types of patterns illustrated below:

- (25) he lost it in the fight then for this right-hand touchline **(0.527)** and **uh (0.467)** he was uh a little bit fortunate perhaps to get away with it (EN-spor-05)
- (26) and these funding councils **then uh (0.620)** **uh** essentially approve our programmes (EN-intf-05)

In the monologic broadcast settings of sports commentaries of Example (25), the MIX cluster can be interpreted as a stalling device to avoid silence while the commentator is simultaneously following the players’ actions. Utterance-medially, however, the similar pattern in Example (26) could be tentatively interpreted as more disfluent, although its post-thematic position is quite typical of speech production processes taking place right after the beginning of an utterance to allow time for macro-planning activities (cf. Section 2.1.2).

To conclude, this section strove to disentangle the parameters involved in the clustering of DMs with filled and unfilled pauses, which constitute the most frequent fluencemes and sequences in *DisFrEn*. The absence of a systematic functional annotation of the FPs forbids us from suggesting more qualitative interpretations of (dis)fluency beyond formal observations of

clustering preferences.⁵⁵ In the remainder of this thesis, functional variables of DMs will be systematically integrated in the analysis of fluenceme sequences in order to bridge this gap between formal and functional features.

6.4 The scope of semantic-pragmatic pairs

One major functional feature of DMs which is cognitively relevant is the divide between objective-subjective or semantic-pragmatic pairs of discourse relations. This distinction is drawn in many models of DM functions and was also found to be experimentally correlated with differences in processing, namely a higher cognitive load for subjective or pragmatic relations (cf. Sections 3.2.1 and 3.2.2). In *DisFrEn*, at least four DM functions are concerned with this ambivalence, viz. *cause-motivation*, *consequence-conclusion*, *condition-relevance* and *contrast/concession-opposition*.⁵⁶ Another related distinction is that of scope (local vs. global), which refers to the distance and size of the segments connected by a DM. I propose an equivalence between semantic-local and pragmatic-global functions and scopes whereby a discourse relation connecting facts (“content” relation in Sweetser’s (1990) terms) should apply to adjacent units, whereas an epistemic or speech-act relation can connect longer and more distant stretches of talk. In this section, I intend to investigate the link between the semantic-pragmatic divide and the notion of DM scope by looking for any corpus-based evidence of cognitive differences in the production of these discourse relations. In particular, I expect the distribution of semantic-pragmatic pairs to differ in terms of syntactic position and sequence types.

Starting with positional preferences, it can be expected that objective or ideational relations of cause, consequence, contrast/concession and condition occur more frequently in integrated slots within the dependency structure than their rhetorical equivalents which might be more attracted to peripheral positions, thus reflecting their higher discourse scope. In Chapter 5, the integration of syntactic and functional variables already showed some interesting effects when comparing the four domains: ideational relations were found to be negatively associated to the pre- and post-field positions and strongly attracted to left- and right-integrated slots; rhetorical functions were associated to the middle field and right-integrated slots and significantly absent from the left-integrated and post-field positions (cf. Section 5.3.4). However, this previous analysis at domain-level encompasses many different functions which do not all belong to the semantic-pragmatic pair. Therefore, this previous association plot (Figure 5.14) was replicated on the subset of the four above-mentioned pairs of relations (amounting to 2,863 DMs), revealing a number of differences. When comparing the ideational and rhetorical domains only applied to [cause], [consequence], [contrast] and [condition]⁵⁷

⁵⁵ See Bolly & Crible (2015, forthc.) on the functional annotation of “pragmatic markers” (including filled pauses) using an extended multimodal version of the present taxonomy.

⁵⁶ The *opposition* function is operationalized as the rhetorical equivalent of both *contrast* and *concession*. Other pairs could be proposed in the present taxonomy (e.g. *alternative-reformulation*; *temporal-enumeration*) although their mapping is not as straightforward as the others and less well-established in the literature. The first element of each pair is always the semantic one.

⁵⁷ The [relation] notation refers to the whole semantic-pragmatic pair, e.g. [cause] includes both *cause* and *motivation*.

relations, the ideational domain is no longer significantly different from the rhetorical one as far as the right-integrated position is concerned, while the rhetorical domain is now positively associated with the pre- and post-field positions. This result confirms our expectation of the higher discourse scope of rhetorical relations which do not function intra- but inter-sententially. Two opposite patterns are illustrated in the following examples:

(27) **if** they're successful there then they can go on to a university (EN-intf-06)

(28) <BB_9> I (0.370) w- worked as an audit manager there for about four years (0.580) and at one stage the m- (0.120) the partner that was in charge of marketing went on a sabbatical (0.190) to South Africa (0.670) and she said to me would I like to take over the marketing function because I'd been doing networking events (0.500) actively going out and trying to gain business (0.320) um and so I said yes
 <BB_1> **so** it was more circumstances then that led you to have a bit of (0.150) a career change (EN-intf-08)

In Example (27), the segment introduced by *if* is a subclause directly followed by the main *then*-clause in an objective, factual relation of condition. In (28), speaker <BB_9> is explaining how she got her new position at her job, a situation which the interviewer <BB_1> summarizes by “so it was more circumstances”. Here, the DM takes as its left context the full previous turn or narrative unit and connects it in a *conclusion* relation (Based on what you say, I can conclude that it was more circumstances). In this example, higher scope (larger size of the left context) coincides with a rhetorical or pragmatic reading of the discourse relation. In addition, this example illustrates the turn-initial use of rhetorical DMs. In the data, 6.42% of rhetorical relations are turn-initial against 2.49% of ideational relations, a significant difference ($z = -5.101$, $p < 0.001$) which indicates the higher discourse scope of rhetorical DMs, connecting (hierarchically) larger units. This interpretation of scope through the lens of position and unit type (subclause, utterance, turn of speech) recalls the approach taken by the Val.Es.Co Research Group (especially Pons Bordería & Estellés Arguedas 2009; Estellés Arguedas & Pons Bordería 2014) regarding the strong link (even restrictions) between functions, scopes, positions and unit types.

However, the associations of semantic relations to local scope and pragmatic relations to global scope do not seem to hold when we zoom in on the specific relations within the ideational and rhetorical subsets. Table 6.11 reports on their proportions across macro-syntactic positions (with the exclusion of four occurrences of the interrupted and independent positions for readability purposes). We see that, within each pair, the proportions of macro-syntactic positions are strikingly similar between the semantic and pragmatic equivalents, with stable preferences for right-integrated position in [cause], pre-field slot in [consequence] and [contrast] and left-integrated position in [condition]. It is these differences between relation types that are responsible for the observed patterns at domain-level, and not the semantic-pragmatic distinction itself.

Table 6.11: Proportions of macro-syntactic positions for the semantic-pragmatic relations

	Pre-field	Left	Middle	Right	Post-field
cause	5.26%	17.00%	0%	77.73%	0%
motivation	7.75%	8.53%	0%	83.33%	0.39%
consequence	82.18%	6.08%	2.31%	7.34%	2.10%
conclusion	82.16%	2.35%	0.59%	4.71%	10.20%
concession	78.53%	8.15%	1.09%	10.33%	1.90%
contrast	75.97%	4.65%	0%	17.83%	1.55%
opposition	87.68%	1.10%	0.92%	4.23%	6.07%
condition	0.45%	72.20%	0%	27.35%	0%
relevance	0%	73.79%	0%	26.21%	0%
Total	59.78%	13.43%	0.80%	22.32%	3.67%

The only notable differences within each pair are: the substantial proportions of post-field position for *conclusion* and *opposition* compared to their semantic equivalents (cf. the *donc euh* pattern, Example (23)); the slightly (yet significantly) higher proportion of pre-field slots in *opposition* compared to both *concession* and *contrast* which, in turn, show more occurrences of the right-integrated slots (especially *contrast*).

Overall, information about positional preferences does not suffice to establish differences in scope between semantic and pragmatic discourse relations but it makes it possible to distinguish between different types of relations. Another potential formal correlate of scope could be found in the types of fluencemes with which semantic and pragmatic relations tend to combine. If pragmatic relations trigger a processing disadvantage, as experimentally shown (e.g. Traxler et al. 1997; Canestrelli et al. 2013), this should be reflected in spoken language by the clustering of cues either signaling some production difficulty or warning the hearer of a more complex content to come. At domain level, an extended association plot revealed no significant differences across sequence types, as opposed to the distribution across positions. However, at function level, some interesting contrasts can be noted from Table 6.12.

Table 6.12: Proportions of sequence types for the semantic-pragmatic relations

	DMs only	+ pauses	+ repet.	+ interrupt.	+ substit.	mixed
cause	50.61%	38.06%	4.45%	4.86%	1.21%	0.81%
motivation	53.28%	26.25%	10.04%	4.25%	3.47%	2.70%
consequence	43.10%	46.03%	5.86%	1.46%	2.30%	1.26%
conclusion	37.25%	48.63%	7.45%	1.96%	0.98%	3.73%
concession	51.09%	39.40%	5.43%	2.17%	0.82%	1.09%
contrast	37.21%	51.16%	6.20%	0%	5.43%	0%
opposition	39.74%	46.34%	7.14%	3.30%	1.47%	2.01%
condition	41.70%	37.22%	10.76%	2.69%	4.48%	3.14%
relevance	48.54%	36.89%	9.71%	4.85%	0%	0%

Starting with a binary comparison between sequences of DMs only and clusters with pauses, we see that some relation pairs show the same clustering tendencies regardless of the domain, especially for [cause] and [condition] which both prefer isolated contexts (DMs only). This absence of effect for these two relation types might indicate that pragmatic cause (*motivation*) and pragmatic condition (*relevance*) do not have a different scope from their semantic equivalents, namely a local scope between adjacent units. This generalized local scope is probably due to the very frequent use of subordinating conjunctions to express these relations, as in Examples (29)-(32).

- (29) but it's now winter so we don't play any more **because** it gets dark too early (EN-conv-02)
- (30) in a sense it's a nice position to be in **because** I'm asking everybody to express what they think (EN-conv-03)
- (31) in the West End they s- they usually do about four pounds **if** you go before five o'clock or something like the Empire (EN-conv-02)
- (32) **if** there is such a thing as Jacksonian democracy that's what they're trying to do (EN-clas-02)

In all these examples, the related arguments are short (single clauses) and adjacent even when the relation is rhetorical as in Examples (30) and (32). It would appear that *motivation* and *relevance* only involve a different degree of speaker's subjectivity and involvement towards their speech but no effect in terms of discourse structure and scope. The situation is somewhat more complex for *contrast* and *concession*, which show different clustering tendencies. While *concession* has a majority of D-sequences (DMs only), as in Example (33), the ideational *contrast* relation favors sequences including pauses, as in Example (34), and resembles in this the distribution of its rhetorical equivalent *opposition* (35).

- (33) there's a soloist and drummer and they all get together **but** they all have fights (EN-conv-02)
- (34) growth in Germany has been sustained by reunification (0.580) **but** elsewhere in Europe activity has slowed (EN-poli-02)
- (35) I met him some time ago and he may not remember (0.260) **but anyway** do send him my best regards won't you (EN-phon-07)

Although the connected segments in (34) and (35) are short and adjacent, the DM "but" appears clustered with pauses, as opposed to the isolated "but" in (33) expressing *concession*. Our expectations of finding more disfluencies around pragmatic relations is therefore only confirmed for the *concession-opposition* pair. It should be noted that the semantic-pragmatic distinction is rather challenging to apply to the *concession* function given that this relation type always involves, by definition, an expectation which is denied and therefore requires some epistemic inference to be made by the hearer. The denied expectation in (33) might be reformulated as: They all get together so you might think that they do not fight but in reality they do fight. The semantic-pragmatic divide for *concession* is therefore more scalar than binary except for clear-cut speech-act relations as in (35).

Lastly, for the [consequence] relations, Table 6.12 shows no difference between D- and P-sequences for the semantic *consequence* and a preference for pauses in *conclusion*, although the proportion of P-sequences is actually similar for the two relations (46% and 48%). Zooming in on the particular DMs expressing [consequence], it appears that this finding can be partly explained by DM-specific preferences. Among the most frequent DMs expressing semantic *consequence*, *so* (123) and *alors* (62) rank first in each language, followed by *and* (101), *then* (25), *so that* and *therefore* (12) in English and *donc* (55), *et* (44) and *pour que* (11) in French. While the dedicated DMs *donc* and *so* follow the expected pattern (i.e. more sequences of DMs only in the semantic use), this is not the case for *and* (55 with pauses vs. 39 with DMs only) or *alors* (30 vs. 22), for instance, which might explain the similar proportions of D- and P-sequences for *consequence* and the proportions of *consequence* and *conclusion* for P-sequences noted from Table 6.12. Once more, we see that the high variability of discourse phenomena, and in particular of DMs, requires some flexibility in analytical levels and degrees of granularity.

Regarding the less frequent types of sequences (repetitions, interruptions, etc.), the differences between *consequence* and *conclusion* are not significant and do not necessarily map the expectation (e.g. slightly more substitutions in the semantic relation). Similarly, S- and Z-sequences (substitutions and mixed) are more frequent in *condition* than *relevance*. Only the proportions of repetitions in *cause* vs. *motivation* show a preference for the pragmatic relation, although the frequencies remain rather low. All in all, semantic-pragmatic (or ideational-rhetorical) pairs cannot be reliably and systematically distinguished by looking at the fluencemes in their direct co-text, with the notable exception of the [consequence] relation when expressed by *so* and *donc* and only when considering their combination with pauses.

To summarize the analyses in this section, a logistic regression model was computed on this subset of relations to predict the domain (ideational vs. rhetorical) including, as input factors, the sequence type, whether or not the DM is clustered (with DMs or other fluencemes), its macro-syntactic position and metadata (register and language). The model performs rather moderately ($C = 0.661$; $r^2 = 0.105$), yet returns a number of significant effects:

- positive effects increasing the probability of a pragmatic-rhetorical use: when the DM is clustered with another DM (compared to isolated); when the DM occurs in the post-field position (compared to the pre-field); in French (compared to English);
- negative effects increasing the probability of a semantic-ideational use: when the DM occurs in left- or right-integrated positions (compared to pre-field); in conversations, face-to-face interviews, news broadcasts, political speeches and sports commentaries (compared to classroom lessons).

It appears that sequence type is not significant in this model. It might be the case that some levels of these variables interact in meaningful patterns, yet the frequencies are too low and the model would be too complex to be reliably interpreted. Overall, this analysis invites us to reconsider the parallel between the notion of scope and the semantic-pragmatic distinction. There might be more disfluencies near major discourse boundaries, as shown by Greene & Capella (1986) and Roberts & Kirsner (2000), but not always near pragmatic relations (not at domain level, not systematically at function level), which means that pragmatic relations do not

constitute higher-level discourse boundaries nor do they necessarily display a higher scope, as illustrated in the examples of this section. Such clustering effects might be more telling by comparing sequential functions, which target the structure of discourse, with the other domains in the taxonomy (see Section 6.6).

To conclude, the high variation and low occurrences for the subset of semantic-pragmatic pairs do not allow us to confirm a link between pragmatic relations and higher discourse scope solely based on the information of position and clustering preferences, unless we use a fine-grained, DM-specific level of analysis (cf. the distribution of *so / donc*). The endeavor to find formal correlates (such as position and fluenceme sequences) to the semantic-pragmatic distinction proves unsuccessful with the annotations available in *DisFrEn*, which further illustrates how challenging it is to annotate this distinction in corpus data (cf. Crible & Degand under review; Zufferey & Degand in press). In this regard, full discourse segmentation as well as annotation of connected segments (as in the PDTB 2.0, Prasad et al. 2008) should provide a useful syntactic basis, although these undertakings are not without their own challenges. In the end, corpus data remains rather silent to considerations of scope and would require external validation in the form of production (elicitation) or perception (response times) experiments. The analyses in this section did not aim at interpreting the relative (dis)fluency of the investigated functions but rather at finding formal correlates to notions of cognitive processing. I will now turn to fluency interpretations in the remaining sections of this chapter.

6.5 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.

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 thesis (especially Section 6.2), 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 6.13 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 6.13: Relative frequency (ptw) of PDFs per language and register

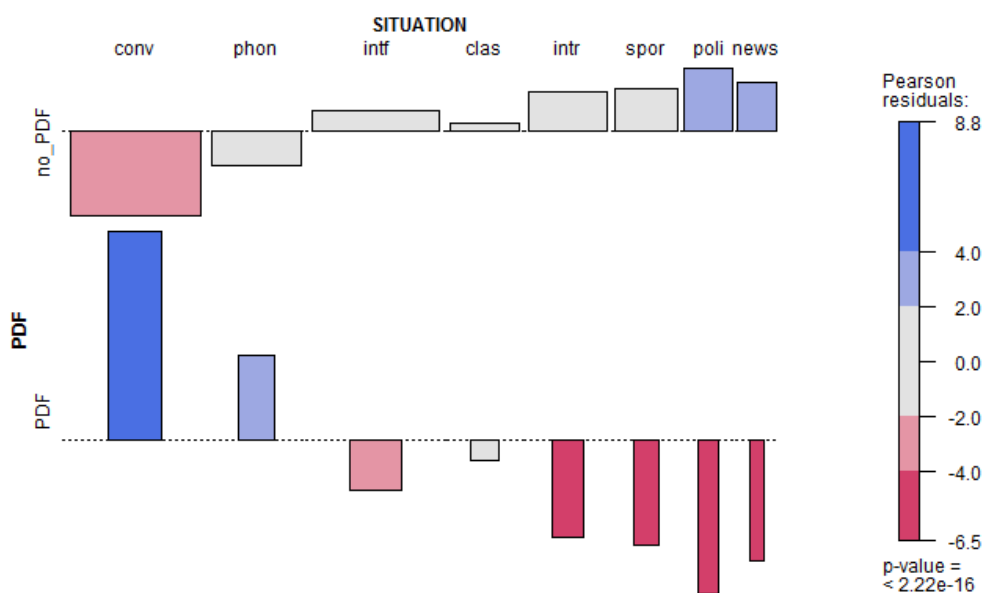
	Monitoring		Punctuation		Reformulation		Total PDFs
	EN	FR	EN	FR	EN	FR	
conversation	3.78	13.54	2.12	4.19	1.77	4.36	14.87
phone	4.31	8.40	2.26	6.19	2.46	3.10	12.58
interview	4.28	6.93	0.59	2.44	0.76	1.88	8.52
classroom	3.71	3.49	0.95	3.49	1.80	1.34	7.00
radio	2.17	3.56	0.57	1.54	1.03	0.95	4.89
sports	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 a 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 their distribution in 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 6.10 represents the extended association plot run on this data. It returns the following significant differences:

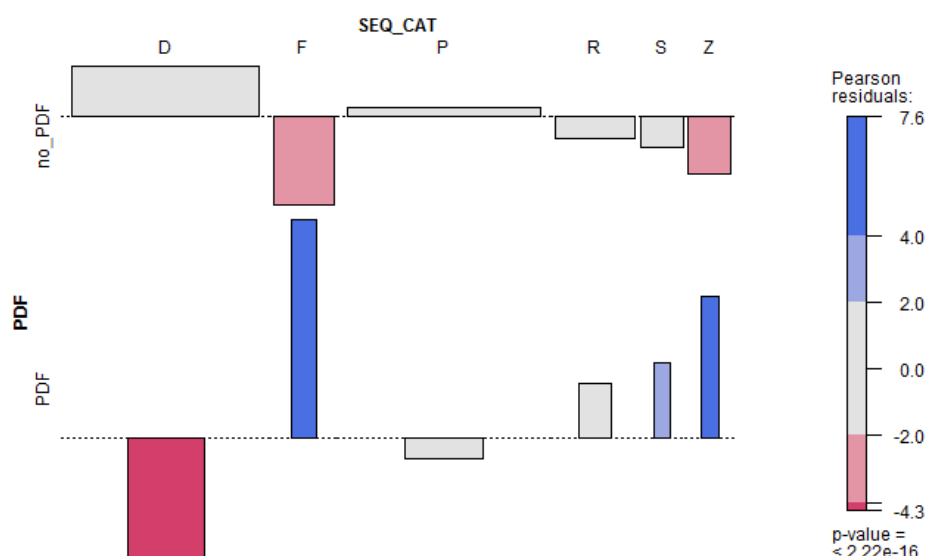
- PDFs are strongly and positively associated to conversations (42% PDFs vs. 26% non-PDFs) and, to a lesser extent, to phone calls (17% vs. 12%);
- PDFs are not significantly different from non-PDFs in classroom lessons (7% vs. 8%) and negatively associated to all the other registers (e.g. 0.32% vs.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

Figure 6.10: Extended association plot of PDFs and non-PDFs across registers

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 6.11), where we see 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 specificity of these sequence types to *monitoring*, *punctuating* and *reformulation*.

Figure 6.11: Extended association plot of PDFs and non-PDFs across sequence types

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 6.11).

- (36) 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)
- (37) <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)
- (38) 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 (36), 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 (37), <ICE_10> substitutes “interior design” by “furnish” after the truncation of “company” and various pauses (Z-sequence). Lastly, in (38), 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 (37) 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 7 to account for such distinctions).

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 (i) mixed sequences of multiple compound and simple fluencemes both embedded and peripheral (especially with *reformulation*, Example (39)), (ii) single compound fluencemes with simple fluencemes in both positions (for all three functions, Example (40)), (iii) single compound fluencemes with peripheral simple fluencemes (especially with *punctuation*, Example (41)) and (iv) single compound fluencemes with embedded simple fluencemes (especially with *reformulation*, Example (42)).

- (39) 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)
- (40) the local councillors **etcetera have have have uh (0.450) you know have** supported us all the way through (EN-intf-02)
- (41) 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)
- (42) <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 each of the three PDFs are related to disfluency each in their own way, either by introducing a nuance or correction (39), calling for cooperation and help during lexical access trouble (40), stalling for planning and maintaining the floor during a very long pause (41) or rephrasing with a different syntactic construction (42). 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.

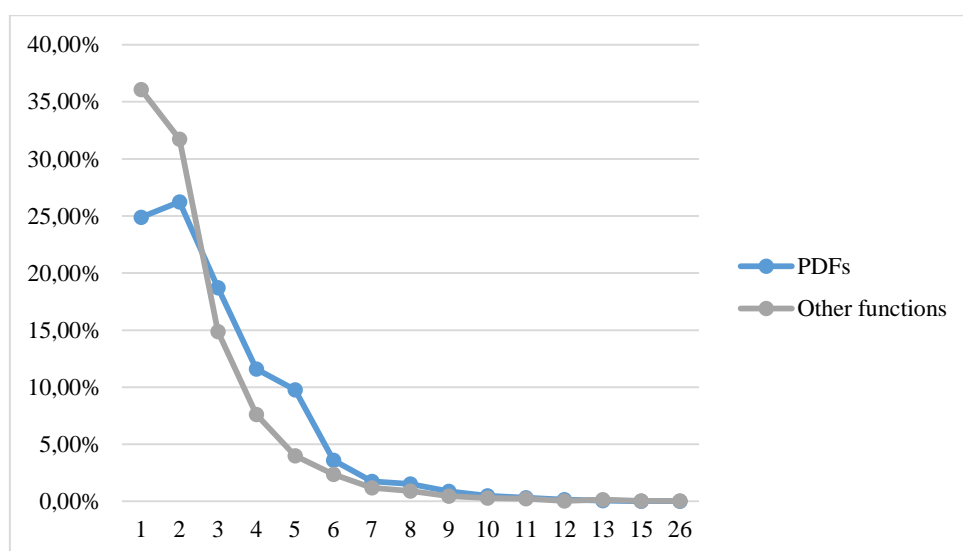
The last variable in the equation is sequence length, which was repeatedly identified as a reliable indicator of disfluency, especially for medium-size sequences, in combination with structural complexity (cf. Section 6.1.3). Figure 6.12 represents the curve of sequence length, measured by number of fluenceme tokens, across proportions of 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 after 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's residuals (computed with an extended association plot) show that five-token sequences are a specificity 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 (43).

- (43) <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 of the DM argues in favor of a non-relational reading as signaling 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 (43), provides yet another validation of their categorization as “potentially disfluent”.

Figure 6.12: Length of sequences in fluenceme tokens in PDFs and non-PDFs



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:

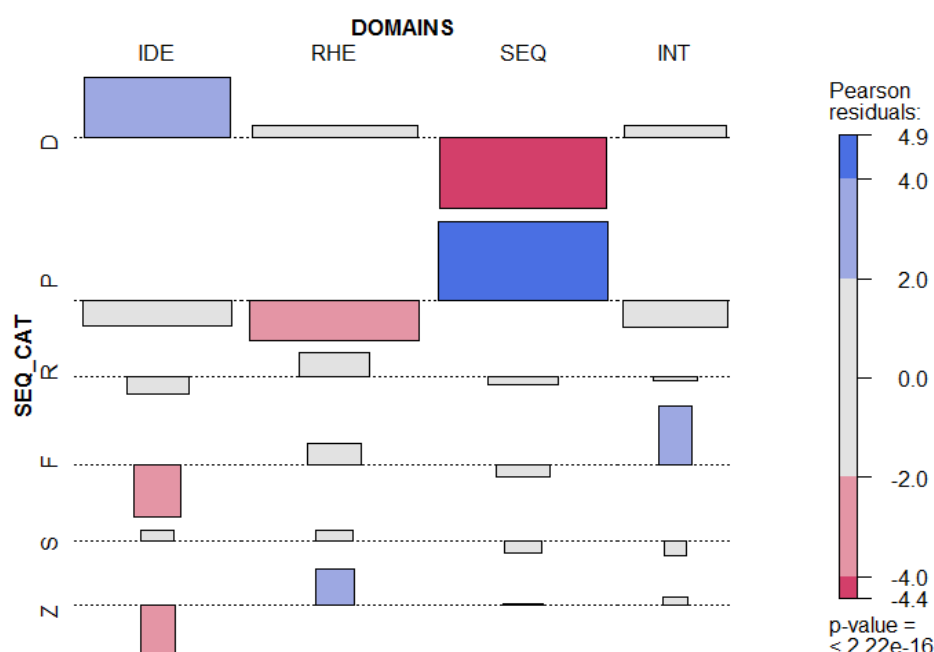
- PDFs are significantly attracted to the conversational register (compared to classroom lessons) and negatively associated with all other settings except face-to-face interviews (which are not significantly different between PDFs and non-PDFs);
- sequences of interruptions (F), mixed fluencemes (Z) 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.

6.6 Towards a cognitive-functional scale of (dis)fluency?

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 6.13 reports on the mapping between sequence type and functional domains. Only single-tagged DMs are included in this analysis ($N = 8,393$; cf. Section 5.3.2).

Figure 6.13: Extended association plot of functional domains by sequence type



We see that each domain has one favorite sequence type and one (or two) dispreferred category. 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.

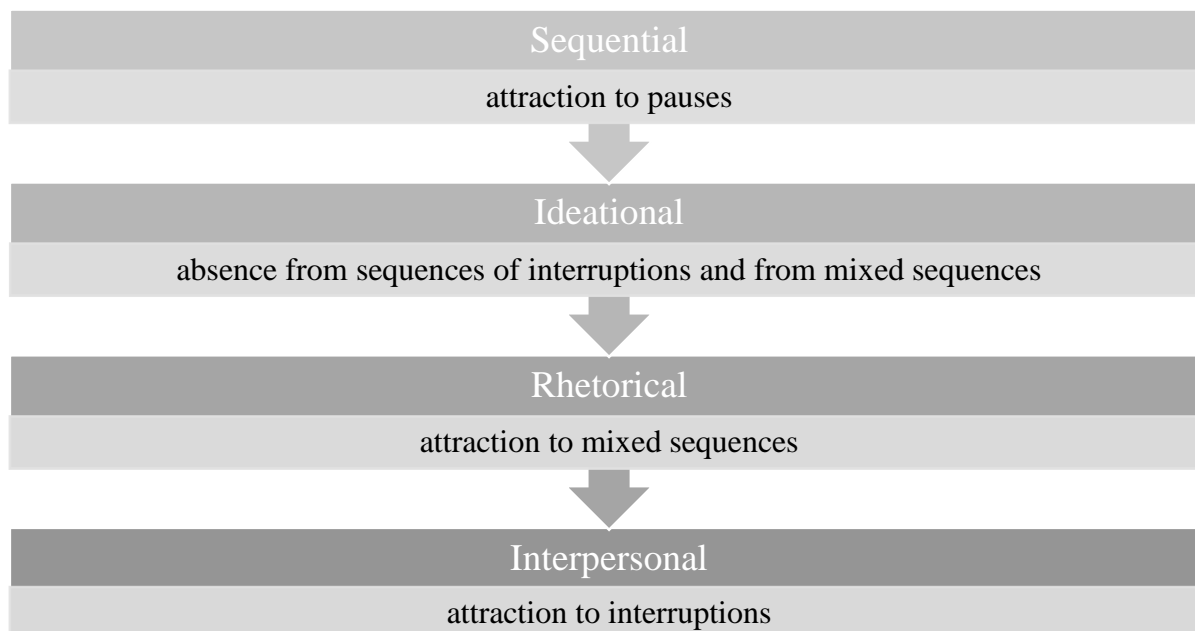
This graph also suggests a relatively high degree of fluency for ideational (“IDE”) DMs, which are quite frequently well-integrated in the utterance (isolated, D-sequences) and negatively drawn to sequences of interruptions (F) and mixed fluencemes (Z). By contrast, the rhetorical (“RHE”) and interpersonal (“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.

- (44) I know exactly the words people are trying to find **but** I’m trying not to prompt them (EN-conv-03)
- (45) 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)
- (46) 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)
- (47) 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 (44) 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 (45) 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 (46) is prosodically independent (unfilled pauses at both sides) and marks a major discourse boundary between two points of the academic lecture. Lastly in (47), 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 (47), the rest of the interrupted utterance, thus relying on the participant’s cooperation to compensate their own incompleteness – whether this incompleteness 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 6.14).

Figure 6.14: DM domains on the scale of (dis)fluency



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 5.2.2.3).⁵⁸ 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 to interpersonal DMs, initial (pre-field) position to sequential DMs and medial (middle-field) position to rhetorical DMs. By integrating domains, positions and sequence types, we can therefore confirm or not the previous interpretations of (dis)fluency.

For readability purposes, Table 6.14 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 (see Appendix 9 for the full table). Starting with the initial position, we see that it is consistently the most frequent slot across sequence types, which is explained by the initiality of DMs on the whole.

⁵⁸ Unlike in written English and written 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.

Table 6.14: 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

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. 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 (48), while rhetorical DMs are mostly medial in sequences with DMs only (D), as in (49).

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

(49) 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 (48) is similar to Example (40) above and 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 (49). 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, yet their distribution is more balanced with the initial position in more disfluent sequences, where they represent up to 50% of occurrences, especially in sequences with interruptions as in (50) and (51).

(50) 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)

(51) 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 (50), “you know” is utterance-initial, following a false-start on “me” and an unfilled pause. In (51), “hein” (‘right’) is utterance-final and follows a truncation (“bat-” for “battue”).⁵⁹ In light of these parallel examples, it appears that the effect of interpersonal DMs in F-sequences

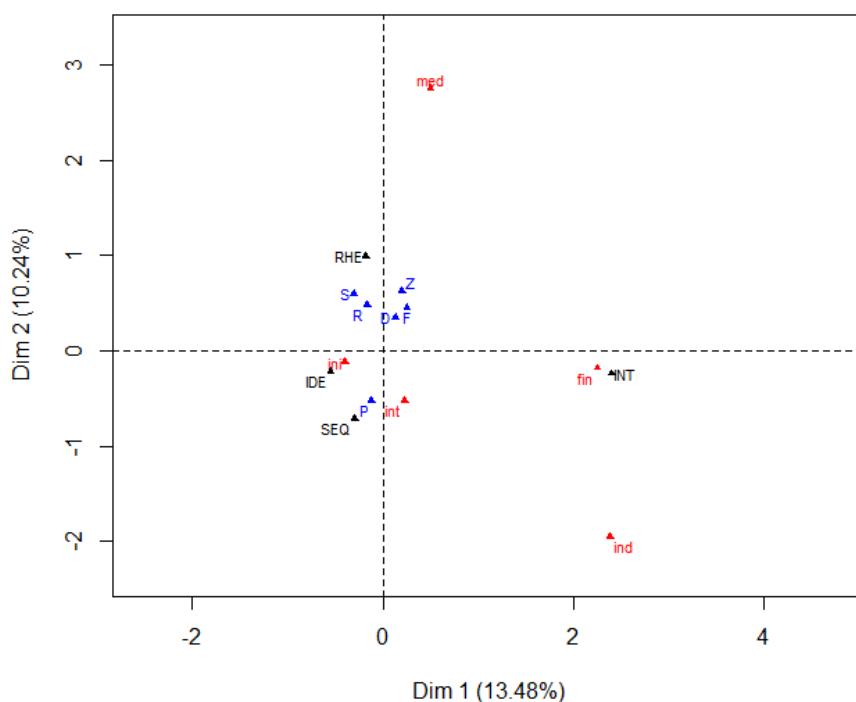
⁵⁹ The prosodic contour of the DM is often necessary to distinguish final from initial position of these DMs which are more flexible and not entitled to a specific syntactic position, as opposed to conjunctions, for instance.

does not fundamentally differ depending on the initial vs. final position, thus qualifying our expectation regarding the disfluency of final interpersonal DMs.

Overall, we can conclude from Table 6.14 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 their lower proportion of medial DMs. 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 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 and 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 6.13 and 6.14 does not hold against the inclusion of an additional variable, let alone against confrontation to 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 6.15.

Figure 6.15: 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.

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 6.15. 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 thesis, where few major crosslinguistic effects have been found.

Table 6.15: 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 R-sequences; conversations and phone calls	initial, medial and independent positions; political and sports registers

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, the sequential domain, which shows converging evidence of fluency, includes the *punctuation* function, which is one of the “Potentially Disfluent Functions” analyzed in the previous section. Likewise, 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.

Replicating the statistical analyses in this section to more fine-grained variables such as functions (instead of domains) and the detailed internal structure of sequences (instead of sequence types) is quantitatively not feasible given the large number of observed levels. Besides, as already mentioned, 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 fluncemes 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 could vouch 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 should therefore remain general and prospective, suggesting coarse-grained – yet statistically significant – trends and opening up avenues for further investigation.

6.7 Summary and interim discussion: the “silence” of corpora

This chapter tested a number of hypotheses put forward in Chapters 2 and 3 regarding the distribution of fluncemes, the relationship between discourse markers and other fluncemes 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. Paradigmatic annotation of fluncemes in the subcorpus of interviews (Section 6.1) established an overall flunceme rate of 20% (in number of words tagged as fluncemes), which is considerably higher than what previous corpus studies reported – although it is largely explained by the intrinsic ambivalence of the fluncemes presently defined as well as by the prominent weight of unfilled pauses and DMs, usually excluded or highly restricted in most typologies. In the data, fluncemes associated to “covert repair” (Levelt 1983) or “forward-looking disfluencies” (Ginzburg et al. 2014) were found to be considerably more frequent than other, 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, fluncemes 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 flunceme 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 fluncemes than not. The more complex the internal structure of a sequence, the less frequent it is in the corpus, yet close analysis of examples forbids to equate low frequency with disfluency at this stage. In fact, it is rather the combination

of compound and simple fluencemes which is a reliable indicator of fluency, especially in medium-size sequences, thus qualifying the fluency-as-frequency hypothesis.

Focusing on DM-based sequences across all registers in *DisFrEn* (Section 6.2), 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).

In the focus analysis on DMs and pauses (Section 6.3, based on Crible et al. 2017a), the independence of DMs from pauses was established by showing a greater attraction of filled pauses (FPs) to DMs than the reverse. However, FPs seem on the whole to be much more attracted to unfilled pauses than to DMs, which confirms their categorization as pauses, against some authors (e.g. Tottie 2011, 2015a) who propose to grant them lexical status. The following patterns were identified: *uh* clustered with an unfilled pause in utterance-internal position; FP clustered with a DM in utterance-initial position; the DM+FP sequence in French, which often corresponds to the emerging construction *donc euh* (Degand 2014).

In an attempt to draw a parallel between the semantic-pragmatic divide and a scale of local vs. global discourse scope (Section 6.4), I was not able to confirm a systematic link between pragmatic relations and higher scope, except when dealing with a very fine-grained, DM-specific level of analysis (cf. the *consequence-conclusion* divide for *so* and French *donc*). However, first leads towards a more reliable semantic-pragmatic distinction were identified such as the attraction of pragmatic relations to co-occurring DMs in the post-field position and to French, as opposed to semantic relations which prefer left- and right-integrated slots. More work (including corpus annotation) on the notion of scope is needed to further this line of research and could highly benefit from discourse segmentation models.

Zooming in even more on a subset of DM functions, the analysis of three "Potentially Disfluent Functions" (PDFs) in Section 6.5 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.

Lastly, a cognitive-functional scale of (dis)fluency was proposed (Section 6.6) 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. However, such an endeavor would forbid any statistical modeling, given the high number of attested levels, and should be entirely based on example analyses, which are themselves constrained by the analyst's subjectivity, so that any corpus-based scale of (dis)fluency can only aim for coarse-grained tendencies, restrictions and favorable conditions for more or less fluent sequences, and should stay clear of undue generalizations.

The analyses in this chapter 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 specificity of this research compared to the bulk of fluency studies was also to bring register variation to center stage with a large 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, this chapter has 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 7: From qualitative repair categories to a formal scale of fluency

Introduction to the chapter

The present approach to (dis)fluency as illustrated in the previous chapters is both formal and flexible: fluencemes are mainly identified and annotated on formal and structural grounds, without any consideration for their potential impact on the relative fluency or disfluency of the utterance, discourse or speaker. As a result, the same labels and macro-labels have been assigned to items and structures which are formally similar but may be functionally and perceptively very different, in line with the overarching assumption of the ambivalence of fluencemes. This flexibility and ambivalence is, in principle, true for all fluencemes in our typology and connects this approach with previous research on the productivity of non-standard and non-linear structures of speech, whether explicitly termed “disfluencies” or not, as in Auer & Pfänder’s (2007) stylistic analysis of “multiple retractions” as a rhetorical feature of French, or Du Bois’s (2014) notion of “resonances” in the framework of dialogic syntax, for instance (cf. Chapter 2, Section 2.1.3). Following this line of research, the present chapter aims at bridging the gap between formal annotation (*how*) and pragmatic interpretation (*why*) of (dis)fluent devices, in a complementary, more qualitative approach to fluency.

In Chapter 6, 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 concretely, in addition to the word-level tagging of fluencemes and fine-grained annotation of DMs, sequences of fluencemes are here classified first formally according to the relation between their different parts (e.g. immediate replacement of the *reparandum* by the *reparans*), then 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) a lot of them actually head down there head down to the Barbican and walk (EN-intf-02)
- (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, discourse markers and propositional substitutions), they differ in a number of other features such as the type and size of the unit under consideration (the verbal phrase “head down there” in (1) vs. the full utterance “they all want to come and have a go” in (2)) or the position of the DM (sentence-internal “actually” in (1), inter-sentential “and” in (2)). More crucially, these two examples illustrate the functional ambivalence of fluencemes, from corrective to non-corrective,

stagnating or progressing, disfluent or fluent: the repeated words introduce a replacement of the pronoun “there” by a more specific proper noun 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 fluencemes with additional layers of analysis, taking a different perspective on the internal structure of sequences and 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. Levelt successfully showed that these different types of repair are expressed by different forms, in meaningful clusters of cues which are designed to help the listener interpret the utterance. The analysis carried out in the following sections strives to identify such clusters of repair types and formats in English and French and to relate them to the formal and functional annotations in *DisFrEn* by means of a conversation-analytic approach to repair and the fluencemes which express repair. Particular attention will be paid to modified repetitions (whether co-occurring with DMs or not) since they can potentially be used in all types of overt repair as defined by Levelt (1983). Modified repetitions appear to be a particularly suited starting point to test whether the ambivalence of fluencemes shows form-function correlates. In doing so, I hope to complement the tentative scale which has been sketched so far (cf. Section 6.6) and against which fluencemes could be “diagnosed”, from very strategic and fluent to very disruptive and disfluent uses, thus converging evidence from the findings of the previous chapters.

In this chapter, I will first situate my own study within the literature on repair and reformulation in order to put forward my research questions and hypotheses (Section 1). Since this analysis required additional coding and variables, its specific methodology will be laid out in Section 2. Results will be presented in Section 3, and discussed in light of our previous findings in Section 4.

7.1 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 the notion of non-linearity, especially in connection with modified repetitions (henceforth RMs) and DMs.

7.1.1 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). Each of them will be briefly reviewed in this section, starting chronologically with Charolles & Coltier (1986), who 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 1998). The remainder of their paper reports on an elicitation experiment with pupils which shows that paraphrastic reformulations are difficult to acquire and use adequately (in terms of maintained textual coherence, choice and variety of the markers).

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).

Finally, De Gaulmyn (1987) differs quite neatly from the two previous references by taking into account the specificity 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 subtraction), 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; Ciabbarri 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 and RMs, 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 (see Section 7.4.4).

7.1.2 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 Ciabbarri (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 following.

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 qualifies this criterion and restricts it to non paraphrastic reformulation, whereas paraphrastic uses 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, are detrimental to the use of her categories in a bottom-up approach such as the present one.

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 den 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 strikes as particularly abstract and not directly related to the concept of fluency, which makes it difficult to adapt to the aims of this chapter.

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 Ciabbarri (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, Ciabbarri (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 thesis, might be too focused on the types of markers themselves rather than the types of reformulations. In particular, Ciabbarri’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 form-function patterns of reformulations, where repair types and DM functions need to remain independent variables.

The last – and by far most relevant – 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 taking a bath 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 incompleteness 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. In the perspective of classifying the range of functions of modified repetitions on a formal scale of (dis)fluency, 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 circular with the 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 chapter is concerned.

7.1.3 From reformulation to repair: Levelt's (1983) typology of repair

As explained in Section 2.2.1, 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, which lies at the core of this chapter, 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. Ciabbari (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 (cf. Section 2.2.1). I will follow Levelt and focus on overt repairs only, of which he distinguishes four categories:

1) 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.

2) 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 of qualification. 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.

3) 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.

4) 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 are therefore clearly fluent cases of repairs. The operational definition of this repair type, as well as all others above and their subtypes, will be detailed and illustrated in Section 7.2 below.

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 qualify 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*.⁶⁰

Levelt's (1983) model, and in particular his repair typology, has been directly replicated in a number of studies (e.g. Brédart 1990; 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 (cf. Sections 2.1.3, 2.2.1). My research questions and hypotheses are strongly based on the definitions and assumptions made by Levelt (1983) on repair types, their associated formal features and overall functional ambivalence.

7.1.4 Research questions and hypotheses

In this chapter, I take modified repetitions and their co-occurring discourse markers as starting point to try and find formal correlates 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-

⁶⁰ This use of "editing term" refers to Levelt's (1989) terminology, as detailed in Section 2.2.1: 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 4.3.2).

repairs are strong disruptions of the syntactic, lexical and/or phonetic structure and occupy the disfluent end of the scale; A-repairs are moderate qualifications 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 will be combined with the formal variables identified by Levelt (1983) as well as the existing annotations of fluencemes and DM functions in the corpus, in order to answer the following questions:

1) What is the proportion of fluent vs. disfluent structures? In particular, how often are modified repetitions involved in typically fluent (R) or typically disfluent (E, D) repairs? Similarly, 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 great polysemy of DMs, I do not start from a list of lexemes but from a group of functions in order to remain consistent with the bottom-up, onomasiological approach adopted in this thesis, 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. Furthermore, I expect that RMs and DMs do not tend to co-occur frequently, since their signaling function would be redundant and competitive 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).

2) Are repair types associated with particular formal variables? Each variable (moment of interruption, distance from the occasion, items in the editing phase, presence of certain fluencemes) will be cross-tabulated with the others to find any meaningful pattern of co-variation, with a particular emphasis on repair types. Levelt (1983) and Fox et al. (1996) suggest the following hypotheses: 2a) detection of repair should occur sooner (i.e. the distance between the occasion and the interruption should be shorter) when the source of the repair is an error than when it is an issue of appropriateness or a fluent list construction; 2b) within-word interruptions should only target erroneous words (repairs of lexical or phonetic error when distance is null) and not "neutral" words.

3) 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.

I will now turn to the methodology, detailing the coding procedure and defining each variable and how they may differ from Levelt's (1983) original study, when applicable.

7.2 Identification and coding scheme

While this analysis is strongly based on the definitions and assumptions in Levelt (1983), a number of revisions and precisions were necessary in order to make the typology fully operational. The relationship between the methodology in this chapter and the existing annotations of fluencemes will also be detailed in this section, as well as the exact materials, procedure and post-treatment.

7.2.1 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) are rather related to covert repairs where the cause of the repair is internalized and cannot be reliably identified, as in (10).

- (10) students can leave at the age of sixteen and **(0.280)** perhaps go to a college where they can do more vocational training **(0.660)** **or or** indeed look **for a for a** job (EN-intf-06)

Based on contextual information and the occurring fluencemes, it is hardly possible to decide whether the speaker in this example is deciding on the upcoming content (conceptual planning), searching for his words (lexical planning), or deliberately stressing some phrases by pausing or repeating before them (intersubjective strategy). In the same vein, given their low propositional content, DMs cannot be considered to replace other DMs but only to co-occur with them, following the annotation of fluencemes, nor can they be considered as false-starts because of their flexible syntactic status (optionality and mobility).

The identification of repairs is not entirely fluenceme-based: a few repair occurrences were identified when no fluencemes were annotated. Although quite rare, these cases include content-level mappings which are not formally marked according to the criteria in the fluenceme annotation protocol. Consider the following example:

- (11) they're built to be (0.650) uh PSVs they're buses basically on the road (EN-intf-02)

In Example (11), "PSVs" is the specific terminological equivalent of "buses", which corresponds to a case of AL-repair. However, formally, only the clitic pronoun "they're" is repeated in each utterance, which is excluded from our annotation of fluencemes (see Crible et al. 2016, Appendix 2: 384). Repairs which only partly overlap with fluenceme annotations, and those which do not include any annotated fluenceme, are marked separately in order to be easily retrieved if necessary. 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 (see Section 7.2.7 below for an assessment of coding consistency). I believe that this independence between the two analytical levels is

beneficial for the analysis since it avoids circularity, as I have also argued for the integration of functional and formal variables (Crible 2016; see also Crible & Degand under review for the independence of annotation levels within functional variables).

Finally, this analysis excludes other-repairs and cross-turn self-repairs such as:

- (12) <VAL_2> je crois que dans la perception des gens qui ont prononcé cette opinion
c'est plus euh un français (0.353) **du temps passé**
*I think that for the people who had this opinion it's more uh the French
language (0.353) of the old times*
- <VAL_3> ah si c'est du temps passé je dirais je ne plutôt pas d'accord
ah if it's the old times then I would say I rather disagree
- <VAL_2> mm (0.587) **enfin du temps passé entendons-nous un français
standard rigoureux**
*mm (0.587) well the old times let's be clear a standard rigorous French
language (FR-intf-01)*

In this example from the subcorpus of face-to-face interviews, the interviewer qualifies the expression “du temps passé” after the interviewee has reacted, in a different turn where she repeats the inappropriate expression, uses an editing term “entendons-nous” (literally ‘let’s understand each other’) and then introduces the repair. Although it does qualify as repair, its distance from the original utterance, separated by someone else’s turn, excludes it from being selected.

All repairs were identified manually through careful reading of the transcripts on the EXMARaLDA interface, making use of the audio when necessary. The repair sequences were then extracted and copied onto an Excel sheet, where the repair type and other variables were manually coded, following the instructions detailed in the following sections.

7.2.2 Repair category

The first variable in this coding procedure concerns the type of repair contained in a sequence identified according to the criteria presented in the previous section. If an utterance contains several repairs, it is decomposed into as many sequences as needed so that a repair category applies only to the relevant elements. Consider the following example:

- (13) we are a class five (0.870) uh pa- class five passenger vessel uh (0.370) which means
[we're based we're we're we're (0.247) limited] to [inland (1.060) uh w- inshore
waters] (EN-intf-02)

In this excerpt, two repair sequences (marked by square brackets) are intertwined: first, “based” is replaced by “limited” after the contraction “we’re” is repeated three times, resulting in a lexical error (EL-repair); next, “inland” is substituted by “inshore” after the truncation of “waters”, which counts as a second lexical error. The first sequence under consideration is therefore “we’re based we’re we’re we’re (0.247) limited” while the second only covers “inland (1.060) uh w- inshore waters”.

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 7.1 below.

Table 7.1: 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
Apppr.	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”

D-repairs

The first category of repairs (D-repairs) covers problems related to linearization or ordering of messages, either superficially with insertions of linguistic material, or more structurally when utterances are re-arranged and re-started in a different way. D-repairs triggering full re-arrangement of utterances necessarily involve a start-over with a new beginning, as in (14).

- (14) you stay in the water but **you do (0.360) we cruise** out un- underneath the Ho (EN-intf-02)

This initiality criterion helps distinguish D-repairs from ES-repairs (see below). Initiality does not necessarily apply to D-repairs characterized by insertions, which can be restricted to one or two words, up to full constituents as in (15).

- (15) **he led his** he he was (0.110) playing bowls up on the Ho (0.620) when he when he uh when he saw the uh invasion and **led his** (0.170) led his ships out to fight them (EN-intf-02)

E-repairs

The category of E-repairs is mostly unchanged from the original typology, except for the addition of criteria and biases implemented for better consistency. Lexical errors (EL) correspond to word substitutions which target an erroneous expression. Hesitation with AL-repairs (see below) can arise on the limit between error and inappropriateness, which is why a systematic bias for EL has been introduced to resolve such hesitations: this disfluent bias favors

the coarse-grained distinction between fluent and disfluent uses rather than finer distinctions of intermediate levels. Clear cases of EL remain the majority, as in (16).

- (16) but that ended uh (0.180) as we know in the uh formally at the **end of (0.170) beginning of** the nineteenth century (EN-intf-05)

Syntactic errors (ES-repairs) cover any intra-sentential restructuring of grammatical elements and constructions. They mostly consist in changes of prepositions or other function-words as well as verb phrases, and must not imply a complete restart of the utterance from its beginning (in which case they are considered to be D-repairs). This non-initiality is well illustrated in Example (17), where the repair only retraces back to the complementizer *that*.

- (17) they're quite impressed **that (0.500) they have you know that you do know** what you're doing (EN-intf-03)

Phonetic errors (EF-repairs) correspond to misarticulations identified as such by the speaker (whereas Levelt also includes unnoticed occurrences). In my view and in line with the annotation of fluencemes, this self-identification is a preliminary requisite for an occurrence to be qualified as repair, since the *reparans* slot would be empty otherwise. EF-repairs cover addition, omission or modification of a phoneme, as in (18).

- (18) a lot of the (0.860) people that are going to be buying from us (0.220) are going through a **dispre- (0.100) distressed** period of time (EN-intf-08)

A-repairs

The third category is that of A-repairs, and its internal structure shows two main changes from Levelt's (1983) original version: first, the subcategory originally labeled "coherence" (AC-repairs) has been removed from the revised typology after the first round of coding because it appeared to be too rare and too problematic to code reliably, given that it involves considerations of thematicity and semantic fields (see Section 7.2.7 below for the full analysis of intra-annotator reliability). The second change within the Appropriateness category is the possibility to assign no subcategory but only the main A-repair label, when the occasion for repair is related to appropriateness in general but does not correspond more specifically to an issue of ambiguity (AA) or precision (AL), as in Example (19).

- (19) on ne parle plus comme cela non non (0.890) me semble pas en tout cas
we don't talk like that anymore no no (0.890) I don't think so at least (FR-intf-03)

In this example, the speaker qualifies his assertion with an epistemic stance, in order to better match the utterance with the intended message. Apart from these two changes, AA and AL remain the same as their original definitions (cf. Section 7.1.3) and are respectively represented in Examples (1), repeated here as (20), and (21).

- (20) a lot of them actually **head down there head down to the Barbican** and walk (EN-intf-02)

- (21) I'm working with a law firm (0.490) to actively (1.320) bring in new clients [...] to also go out into the local community and find **people (0.180) suppliers** that we can actually (0.180) take back to our clients (EN-intf-08)

We can see that, compared to repairs of lexical errors (EL), appropriateness repairs do not negate the truth of the original utterance but merely qualify it: “Barbican” is a more specific referent for “there”; “suppliers” is the more technical term for “people” in this context. These qualifications are weaker than pure corrections, which motivates the difference in degree of fluency between A- and E- repairs, intermediate for the former, lower for the latter (cf. the disfluent bias above).

R-repairs

Lastly, R-repairs no longer correspond to a “rest” category which Levelt (1983) used for complex cases (which are now always disambiguated to fit one of the categories presently defined), but refer here to “resonance” and consist of productive strategies of recycling. The term is borrowed from Du Bois (2014) and corresponds to Auer & Pfänder’s (2007) “retractions”: R-repairs must involve both repeated elements and modified elements combined in a recycled structure which progresses in the syntagmatic axis, as opposed to other types of repairs which are more stagnating, replacing one constituent by another. This progression is paramount in the definition of R-repairs and might create an enumeration, parallelism, contrast or any other “fluent” effect, as in Example (22).

- (22) we have to have (0.980) um life saving floats we have to have (0.280) life buoys we have to have (0.470) bilge pumps (EN-intf-02)

Subtypes

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, cf. Example (11) above) or not, when the degree of precision is at stake but involves terms of an equal level of specialization, as in (23) where the speaker defines more precisely what he means by “correctes”.

- (23) 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 or “LocLin” (i.e. insertion of one or two words, related to the ordering of words in an utterance) and global linearity issue or “GloLin” (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 (24)-(26) below.

(24) it's more of the Liverpool **acc-** but I can certainly tell the difference (EN-intf-03)

(25) 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)

(26) 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 (cf. Example (22)). Parallel R-repairs express a stronger sense of contrast or mirroring between two or several elements, as in exclusive alternatives (Example (27)).

(27) 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)

To conclude on this revised typology, I would like to point out the partial overlap between some of these categories and the fluencemes that typically express them: this mapping is circular for false-starts and insertions, which are defined as criterial for the subtypes of D-repairs, but does not affect the two fluencemes at stake in this chapter, namely modified repetitions and discourse markers, which are not restricted to any type or subtype presented in this section. I also want to stress the fact that, while this analytical grid for repairs has a larger coverage than the original proposal by Levelt (1983), most modifications correspond to additional types or subtypes (for instance the “LocLin” 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, especially focusing on modified repetitions.

7.2.3 Formal variables

In addition to the qualitative typology of repair types, Levelt (1983) identified a number of formal variables which are relevant to the study of repair and the association between form and function of the repair. Following the same corpus-based methodology as previously, I will define the levels of each of these variables and specify the operational criteria for their application to my corpus data.

7.2.3.1 Interruption point

Levelt's (1983) most investigated variable is the interruption point, that is the type of unit being uttered when the *reparandum* ends. He originally identified three units hierarchically ordered from largest to smallest: constituent or phrase boundary, word boundary, and within-word phonological boundary. To account for the larger scope of R-repairs, typically repeating full

clauses, I propose to refine these structural possibilities by further distinguishing clauses and full utterances. The list of possible values is provided in Table 7.2, where they are defined and illustrated by authentic examples from the corpus.

Table 7.2: Definition and examples of boundaries at the interruption point

Boundary	Definition	Example
within	incomplete word, syllable or phoneme	it's ha- roughly half half
word	smallest unit of independent use	so the uh it's been it's been fun
phrase	complete intra-sentential constituent	we then go through there through the Barbican
clause	incl. relative, complement and subordinate clauses	we can argue between ourselves whether it's you that pays or or me that pays
unit	full utterance (semantic and syntactic completion)	is it going to look like dad it is going to look like mum

The coding of this variable is rather straightforward, apart from two precisions which need to be added. First, the present definition of words is simplistic and does not take lexicological issues of compound or multi-word expressions into consideration: an interruption after *credit* will be tagged as word-boundary, even if the intended unit was *credit card*, an assumption which is not consistently available for the analyst. The within-word boundary is reserved to word fragments, following the annotation of truncations.

The second precision concerns phrases, which can sometimes be mistaken for word boundaries when the interruption occurs after a word which could be considered to end its host-phrase. In most cases, either the intonation contour of the *reparandum* or the rest of the repair (e.g. presence of an adjective in the repair) indicate whether the original utterance was complete or not, as in the example below.

(28) 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)

In (28) the substitution of “abdication” by “abolition” is then completed by the complement “of slavery”, thus retrospectively rendering the original phrase incomplete and coded as word-boundary. In other words, simple noun phrases (e.g. “the boy”) are tagged as phrase-boundary unless a more complete phrase can be reconstructed from the *reparans*. No other particular instructions are required for this variable which is involved in a number of Levelt's (1983) original hypotheses (cf. hypothesis 2b in Section 7.1.4).

7.2.3.2 *Occasion for the repair*

This variable documents the particular lexeme which either triggered the repair (that is, the erroneous or inappropriate word) or the last word before the interruption when the repair cannot be related to a particular word but to the whole structure or utterance instead. In case of a word-to-word equivalence between *reparandum* and *reparans*, we trace the occasion of repair back to the first problematic word in the phrase, as in (29).

- (29) you'll find that many of the houses (0.260) have medieval (0.590) um (0.910) **areas under them tunnels** (0.330) in which the wine was stored (EN-intf-05)

In this example, two options are possible: the occasion for repair can either be the last word in the phrase (“them”) or the first problematic word (“areas”) which is the syntactic equivalent of the *reparans* “tunnels”. In my view, the repair operates between the two nouns, with “areas” being specified by the more accurate term “tunnels”, so that the prepositional phrase “under them” is not considered a part of the *reparandum* but only “neutral words” in Levelt’s terms. However, when no equivalence of syntactic class can be found, the last word of the phrase is considered to be the occasion for repair, as in “the other tour stays you stay” (EN-intf-02), where “tour” is the occasion.

The value for this variable only repeats the orthographic transcription of the lexeme identified as the occasion for repair. This information is not involved in any analysis on its own, but is used as the basis for the measurement of distance, as explained in the next subsection.

7.2.3.3 *Distance between interruption and occasion*

This numeric variable counts the number of words (including truncations) between the occasion for the repair and the editing term (if any) or end of the *reparandum*. The value is null when the repair starts immediately after the occasion for repair, or after some editing terms. Levelt (1983) originally measured this variable in number of syllables, but this level of precision is not necessary for the present analysis.

To sum up the variables describing the internal structure of the repair, the interruption point, occasion for repair and distance are illustrated with the following example.

- (30) a blacksmith was asked to (0.420) um put **shoes on the devil (0.520) horse shoes on the devil** and (0.400) because he was so nervous about doing it he accidentally put a nail (EN-intf-08)

Here, the interruption point occurs at the end of a whole unit (“a blacksmith was asked to put shoes on the devil”); the occasion for repair is the noun “shoes” which is later specified as “horse shoes”, amounting to a distance of three words (“on the devil”) between the occasion and its repair (after a silent pause in the editing phase).

7.2.3.4 Way of restarting

This last formal variable is directly borrowed from Levelt (1983) and concerns the *reparans* rather than the *reparandum* as opposed to the last three variables. It contains four possible levels:

Instant replacement. The speaker retraces just to the occasion for repair which is substituted. This way of restarting does not necessarily imply a null distance between the occasion and the interruption but allows for the presence of neutral words or editing terms, as in: “anybody that has more than 21000 pounds currently or 22000 pounds I think it is” (EN-intf-04).

Anticipatory retracing. The speaker retraces and repeats some words before the occasion for repair, from single articles or pronouns to much longer stretches of words. In other words, some material from the *reparandum* will be repeated in the *reparans* as in “they started as lawyers or they started as accountants” (EN-intf-08). The number of graphical units being repeated before the occasion of repair is counted (here, “they started as”, 3 words). Levelt (1983) calls this information the *span of retracing* and originally measures it in syllables, but as already mentioned, here it is words that are counted.

Pre-specification. The speaker repeats the occasion for repair after the insertion of new material, either a few words or a longer stretch of words. The difference with anticipatory retracing is that it is the occasion for repair itself that is repeated, with some new material inserted before it in the *reparans*: “encourage them really to after a while to go home” (EN-intf-03). Pre-specification does not necessarily imply a linearity problem but can also be related to a repair for appropriateness or even lexical error when the inserted material brings up a strong semantic change, as in “loss of independence or fear of loss of independence” (EN-intf-04).

New material. The speaker replaces an interrupted structure by a new one with little or nothing in common. This type of restarting is necessarily initial, as opposed to instant replacement which is mostly utterance-medial.

These four ways of restarting are further illustrated with built examples in Table 7.3.

Table 7.3: Variations of repair according to the way of restarting

instant replacement “INSTANT”	we’re limited to inland inshore waters
anticipatory retracing “ANTICIP”	we’re limited to inland to inshore waters
pre-specification “PRESPE”	we’re limited to inland mainly inland waters
new material “NEW”	we’re limited to / the bus must be on inland waters

From this table, we can see that instant replacement involves no addition of new material nor repetition of previous material but directly substitutes one word by another. Anticipatory retracing involves the repetition of one word “to” from before the *reparandum* “inland”. Pre-specification repeats the occasion for repair itself “inland” and inserts new material in the *reparans* “mainly”. New material is the most different format of all four since nothing is in common between the *reparandum* and the *reparans*.

Some of these options are to a certain extent associated to a specific repair (sub)type: for instance, *new material* and *pre-specification* in Table 7.3 respectively correspond to issues of false-start and local linearity. However, this association is not systematic, as I have said above and as illustrated by the examples of *instant replacement* and *anticipatory retracing* in this table which are rather cases of lexical errors (EL-repairs). The relationship between type of repair and way of restarting is particularly telling of the types of cues which are given by the speaker to instruct the hearer on how to process the repair. This formal information will be especially interesting to differentiate A-repairs from E-repairs, taking us closer to a formal scale of fluency.

7.2.4 Relation to annotated fluencemes

The variables presented so far are additional information coded in the existing *DisFrEn* dataset. Although independent, this new analysis still makes use of the annotations detailed in Chapter 4, in order to find recurrent mappings between repair type, formal variables and the actual fluencemes in the sequence.

7.2.4.1 Discourse markers

In line with the general approach of this thesis, the present analysis pays particular attention to DMs, their presence, position and function, focusing on those occurring in the editing phase. First, a binary value indicates whether some DMs occur in the sequence containing the repair. If so, the specific lexeme(s) is/are specified, in their order of appearance if there are more than one. Each DM is then 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.

Lastly, focusing on DMs in the editing phase (including when they are the first word of the new segment, see above), 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 nor retrieved for this analysis.

7.2.4.2 Other fluencemes

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, excluding peripheral RMs which are either considered as a separate repair sequence, or discarded when they are combined with truncations or phonetic variants, as in “the end **of a of an** era”.

In addition to insertions, interruptions, RMs and DMs, all other fluencemes included in the editing phase are grouped in an “other” category mostly containing pauses and identical repetitions. 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).

7.2.5 Other information

All sequences containing a repair are extracted from the transcript and copied onto the Excel sheet with the following information: text ID code; sequence ID code from the original annotations, if any; beginning time of the sequence; full unit containing the repair sequence; fluenceme labels and sequence summary from the original annotations, if any. Additional comments can be added in a separate column in order to retrieve interesting examples, complex cases, contextual information as well as recurring observations such as the combination of *anticipatory retracing* with *pre-specification* when the speaker repeats a few words and inserts new material at the same time in the repair, or when the value for repair type has been changed after the second round of annotations (see Section 7.2.7 below).

7.2.6 Procedure and post-treatment

The coding scheme presented in the previous sections was manually applied to the subcorpus of face-to-face interviews in English and French (17,000 and 18,000 words in each language, respectively, cf. Chapter 4). 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 selection criteria presented in Section 7.2.1. Access to prosody turned out to be particularly useful for the coding of repair type, while the other variables were plainly based on the transcription and the available fluenceme labels.

The coding itself was carried out directly in Excel, and checked for typing errors or inconsistencies. The presence of false-starts (FS) and modified repetitions (RM) was semi-automatically retrieved from the fluenceme labels contained in each occurrence of repair. No further post-treatment was required at this stage. In a separate table, all occurrences of DMs expressing either *reformulation*, *specification* or *enumeration* and retrieved from the subcorpus of face-to-face interviews were listed (cf. hypothesis (1) in Section 7.1.4). This filtered list was matched with the first table of repair sequences, in order to identify the DMs which occur in

the editing phase of a repair. Nothing else was coded in this DM table apart from the repair type of the sequence and the existing annotations (of function, position, etc.). DMs expressing other functions than the selected three were not included in this second table.

7.2.7 Coding consistency and intra-annotator agreement

All coded variables were first checked to make sure that no repair occurrence had been overlooked in the transcripts, and that there were no major inconsistencies, following the criteria and biases defined in the coding scheme. Then, a second round of blind coding was carried out (after a few days' interval) in order to provide a measure of intra-annotator reliability. The conditions of the two coding phases are not strictly identical since the second round only made use of the Excel sheet and not the full context nor the audio. Furthermore, this re-coding only targets the repair category and not the other variables, which are both less subjective and less central to the analysis. The information from these other variables was hidden (except orthographic transcription and fluenceme labels) so as not to influence the re-coding of the repair category. The values of the two coding phases were then formatted to be compared, first in Excel with a relative measure of agreement (the proportion of values which are the same or different across the two rounds) and later with a statistical measure of agreement, i.e. Cohen's kappa. For each case of disagreement, a final "gold-value" was established and then implemented in the dataset (with an additional comment indicating that the value has been revised, cf. Section 7.2.5).

Starting with a global assessment measure, intra-annotator consistency 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 ($\kappa = 0.779$ and $\kappa = 0.74$, respectively, cf. Section 4.2.2.3), which vouches for the replicability of the analysis. A closer meta-analysis of the disagreements allows one to identify problematic categories. Table 7.4 reports on the number and percentage of agreements and disagreements for each repair type.

Table 7.4: Intra-coder consistency for each repair type

Repair category	Agreements	Disagreements	Total	Agreements %
Delay (D)	104	21	125	83%
Syntactic errors (ES)	58	18	76	76%
Lexical errors (EL)	63	12	75	84%
Resonance (R)	63	3	66	95%
Level of precision (AL)	18	10	28	64%
Ambiguity (AA)	10	5	15	66%
Generic appropriateness (A)	6	6	12	50%
Phonetic errors (EF)	6	1	7	86%
Terms coherence (AC)	0	2	2	0%
Total	328	78	406	81%

First, it should be noted that this table is computed against the total number of labels assigned: in cases of disagreements, each type is counted separately, thus doubling the actual number of disagreements (from 39 to 78). At first glance, all repair types show substantial proportions of agreement, apart from rare values such as AC and A. 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. This D-ES pair is responsible for most disagreements of the ES category (14/18). Zooming in on these cases, a qualitative analysis of problematic sequences reveals three recurrent patterns:

- insertions usually around the verb phrase, as in “they get off [...] and **walk (0.520) go walk** because...” (EN-intf-02) or “people are (0.207) coming from all sorts of walks of life to **ha- come and have** a chat” (EN-intf-02), which can be interpreted either as problems of syntactic structure or linearity issues. These cases have been systematically resolved as D-repairs and further sub-coded as “LocLin”;
- hesitations between replacement and re-start, as in “the amount of people in Plymouth that uh (0.940) [...] the majority of tourists coming to Plymouth” (EN-intf-02), where, despite the initiality criterion, the repair shows strong connections with the original utterance which tend to be seen as a change of construction at the lexico-syntactic level instead of a full re-start;
- replacements at the beginning of an utterance or subclause, especially when just the first word is substituted, as in “**he they** appear regularly” (EN-intf-06). These cases are resolved as ES-repairs when they affect function-words such as pronouns.

Other problems stem from A-repairs and their subtypes, which show lower proportions of agreement than the other categories. The low frequency of A-repairs is responsible both for the absence of stronger criteria and biases (for lack of observations to be trained on) and for the difficulty to interpret the disagreements, apart from a tentative recommendation to further define the limit between generic appropriateness (A) and precision (AL), especially for cases such as “it’s ha- roughly half half” (EN-intf-02). This particular example is resolved as A-repair because the insertion adds a nuance, while AL-repairs typically replace one term by another. AC-repairs were removed altogether from the revised typology since the only two potential cases were disagreed upon in the second round, making this category entirely unreliable at least in its present state of definition. A last observation for A-repairs concerns the notable absence of hesitations between A and E types (only 5 cases of disagreement involve an EL and a type of A), which could have been expected from the conceptual similarity between inappropriateness and error.

The remainder of the disagreements are not related to a particular category but rather to the lack of (audio) context during the second round of annotation (which benefitted neither from the full transcription nor the prosody), as in Example (31).

- (31) in fact **yes only the other day** a a mum whose little one was going to be six (EN-intf-03)

Here, the transcription conventions did not make it explicit that “yes” was the truncated form of “yesterday” which is repaired by “only the other day”, making this sequence a case of lexical (EL) repair.

In light of these observations, all remaining cases of disagreement were settled, following the general principle of coherence with the definitions, criteria and biases. The resulting gold-standard values are reported in Table 7.5.

Table 7.5: Final values for repair type and number of (dis)agreements

Gold-value for repair type	Agreements	Disagreements	Total
Delay (D)	104	9	113
Phonetic errors (EF)	6	0	6
Lexical errors (EL)	63	9	72
Syntactic errors (ES)	58	10	68
Generic appropriateness (A)	6	4	10
Ambiguity (AA)	10	4	14
Level of precision (AL)	18	3	21
Resonance (R)	63	0	63
Total	328	39	367

This table shows that, after revision of the problematic cases, all categories are more frequently involved in agreement than disagreement, which tends to show that the gold-standard values are often the values from the first coding phase (cf. the lack of context and prosody in the second round). The fact remains that A-repairs are more problematic than the other repair types, relatively to their total number of occurrences. EF- and R-repairs strike as very reliable categories since their rare cases of disagreement were due to the technical format of the second coding phase.

All in all, the kappa-score and the fine-grained analysis of intra-annotator agreement reveal that the coding of repair, which is the most interpretive and qualitative variable in this coding scheme, is quite robust and replicable. The analysis of formal variables and their association to repair types should be all the more interesting since its qualitative basis is rooted in a reliable methodology which manages to handle the variation and complexity of authentic spoken data.

7.3 Results

7.3.1 Distribution of repair categories

I will start the systematic analysis of repair types and their associated variables with general considerations on the distribution of the different (sub)categories, in order to identify tendencies of use for overt repairs and possible crosslinguistic differences.

Table 7.6 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 qualifies 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 (Section 6.1.1). This difference can be explained by the fact that the analysis in the present chapter only targets overt repairs, as opposed to the larger 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.

Zooming in on the specific types of repairs, it appears that false-starts (i.e. D-repairs for linearization) are the most frequent overall with 22% of all occurrences, closely followed by lexical errors (especially in English where their frequency is very similar to false-starts) and syntactic errors (especially in French). List constructions show a smaller yet substantial frequency of 13% in the two languages (cf. the frequency of S-sequences in interviews, Figure 6.6, Section 6.2). All the other (sub)types amount to less than 5% each (except English AL-repairs), the least frequent one being phonetic errors.

Table 7.6: Frequency and proportions of repair categories and subtypes by language

Repair category	EN	FR	Total	EN %	FR %	Total %
Error (E)	61	85	146	36.97%	42.08%	39.78%
Lexical (EL)	33	39	72	20.00%	19.31%	19.62%
Syntactic (ES)	24	44	68	14.55%	21.78%	18.53%
Phonetic (EF)	4	2	6	2.42%	0.99%	1.63%
Delay (D)	51	62	113	30.91%	30.69%	30.79%
false-start	32	49	81	19.39%	24.26%	22.07%
global linearity	11	6	17	6.67%	2.97%	4.63%
local linearity	8	7	15	4.85%	3.47%	4.09%
Resonance (R)	27	36	63	16.36%	17.82%	17.17%
list	21	26	47	12.73%	12.87%	12.81%
parallel	6	10	16	3.64%	4.95%	4.36%
Appropriateness (A)	26	19	45	15.76%	9.41%	12.26%
Level of precision (AL)	13	8	21	7.88%	3.96%	5.72%
terminology	7	1	8	4.24%	0.50%	2.18%
N/A	6	7	13	3.64%	3.47%	3.54%
Ambiguity (AA)	9	5	14	5.45%	2.48%	3.81%
Generic (A)	4	6	10	2.42%	2.97%	2.72%
Total	165	202	367	100%	100%	100%

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” (EL+A= 32%) and those related to fluent strategies (R= 17%). This focus of monitoring on form rather than content is, in my view, evidence of the specificity of unplanned speech where speakers have to order complex information into the linear phonological 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 spatial, a-temporal nature of writing (cf. Section 2.1.3). 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.

Turning to A-repairs, their low frequency could be due to methodological issues, namely their weak definition and problematic identification, as was shown in Section 7.2.7: it could be the case that more prescriptive criteria would have helped to identify more occurrences. As a result, A-repairs are not easily interpretable in the analysis because of their low frequency and low reliability, in addition to their intermediary position on the fluency-disfluency scale. Further findings will indeed show that A-repairs are not always (formally or otherwise) distinguishable from other repair types (see below).

Finally, from a contrastive perspective, Table 7.6 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 and 6 regarding the distribution and clustering of DMs and fluncemes.

7.3.2 Repair category and formal correlates

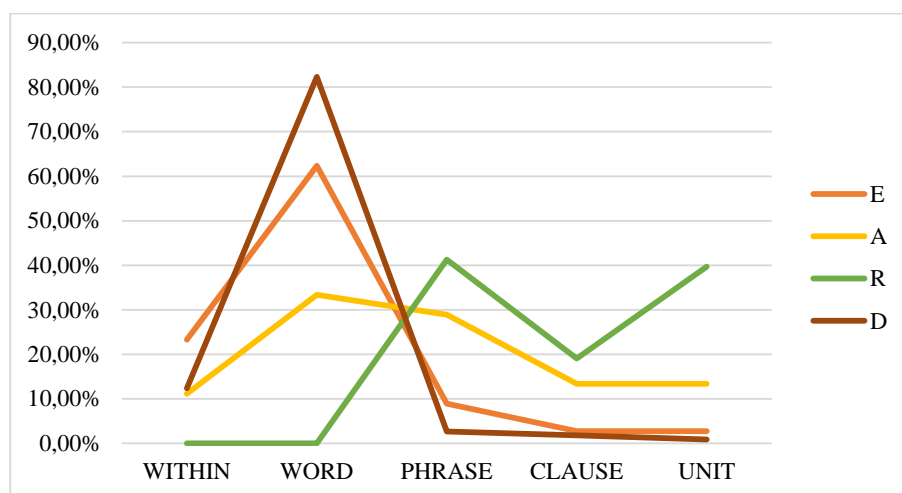
The main hypothesis in this analysis states that different types of repairs are expressed in different forms so as to help the listener predict what is coming next and how to integrate it with the original utterance. This hearer-oriented hypothesis is inspired by Levelt’s (1983)

results on the association between repair type and formal variables, but also by more recent, especially experimental literature on the predictive role of disfluencies (e.g. Arnold et al. 2003, Corley 2010) and my own corpus findings regarding the association of form-function patterns in DMs and fluencemes. In the following sections, I will systematically look for relations between the coded variables and try to identify recurrent patterns.

7.3.2.1 Repair category and moment of interruption

Levelt's (1983) Main Interruption Rule ("stop the flow of speech immediately upon detecting the occasion of repair", 1983: 56) predicts that, in case of error repairs, speakers will interrupt at any point of their utterance without regard for linguistic structure in order to stop the course of inferential mechanisms as soon as possible. The interruption should therefore occur sooner in E-repairs, which would result in a higher frequency of within-word boundaries for this category than for the others. Figure 7.1 shows that my data only partially confirms this hypothesis.

Figure 7.1: Proportions of interrupted units by repair category



A number of interesting conclusions can be drawn from this graph. First, within-word interruptions confirm hypothesis (2a) in that they appear to take up larger proportions for erroneous words and actually follow the cline of disfluency: the more disfluent the repair (when D and E are considered equally disfluent), the higher proportion of within-word interruptions. However, we see that the moment of interruption for A-repairs can also occur at a within-word boundary, albeit rarely, especially for the generic A-repairs as in "it's ha- roughly half half" (EN-intf-02) or the following example:

- (32) c'est devenu assez **dur** **puisqu** **maint-** **enfin** (0.170) **dur** (0.453) **relativement**
 puisqu maintenant euh je suis responsable
it's become quite difficult since n- well (0.170) difficult (0.453) relatively since now uh
I'm responsible (FR-intf-06)

In (32), the speaker uses the DM "enfin" to introduce a backward-looking move and retrospectively qualifies the adjective "dur" with a postponed adverb "relativement" after two

neutral words, one of them being truncated: this example testifies to the structural possibility of within-word interruptions for non-erroneous words, and therefore denies the monopoly of E-repairs on this format, against the expectation of hypothesis (2b). Looking at the specific repair types, only EF-repairs are more frequently interrupted within words than at word boundary (4 out of 6 cases), although the small frequencies forbid any stronger conclusion.

The second major observation from Figure 7.1 concerns the overwhelming frequency of word boundaries for each repair category except R-repairs (and AL-repairs, within the A category, which are interrupted at phrase boundary in 10 cases out of 21). The tendency to stop at word boundary and not after more complete constituents is telling of the disfluent character of D, E and A-repairs altogether: when there is something wrong, regardless of the degree or nature of the problem, speakers tend to interrupt their utterance sooner rather than later and not to take syntax into account. The notable difference in this regard is the R-repair category, which shows no occurrence of interruptions at boundaries smaller than the phrase. R-repairs show the largest raw frequency and proportion of interruptions at phrase, clause and unit boundary (41%, 19% and 40%, respectively) compared to E, D and A-repairs. This result was to be expected from the very definition of the R category, since R-repairs build on longer, more complete units which carry independent meanings (as members of an enumeration or a parallel construction). More importantly, it confirms the overarching hypothesis that fluent and disfluent strategies are associated with formal correlates, at least in terms of the moment of interruption.

Further evidence of this divide between R-repairs and the other three categories can be found in the (quasi-)absence of interruptions at larger constituent boundaries for D, E and A-repairs, especially for clause and unit: while a few occurrences at phrase boundary can be found in types A (especially AL) and E (especially EL), they are very rare in the “structural” repairs (D and ES: less than 5%). Occurrences of clause and unit boundaries are even scarcer with anecdotal cases (1 or 2 in each repair type) such as Example (33).

- (33) on ne peut pas dire que on parle sans accent ou sinon vous **ne sauriez pas parler**
(1.000) ne pourriez pas parler plutôt
we cannot say that we speak without an accent otherwise you would not be able to
speak (1.000) could not speak rather (FR-intf-01)

In this example of EL-repair, the speaker replaces a modal verb (in a typical Belgian French use) by another after the completion of the whole unit and a long pause, in a rather distant backward-looking move. In the data, such examples are rare, especially in English, and mostly correspond to cases where the speakers change their mind, hence not only repairing the linguistic output but also the idea behind it. Overall, D- and E-repairs are strongly associated with word boundaries and almost never interrupted after larger constituents, while R repairs are exclusively associated to larger constituent boundaries (especially phrase and unit). A-repairs seem to confirm their intermediary place on the fluency-disfluency scale by showing a more varied profile, with some occurrences at every moment of interruption. This variable therefore appears to be highly relevant to the present endeavor of building of scale of (dis)fluency on formal grounds.

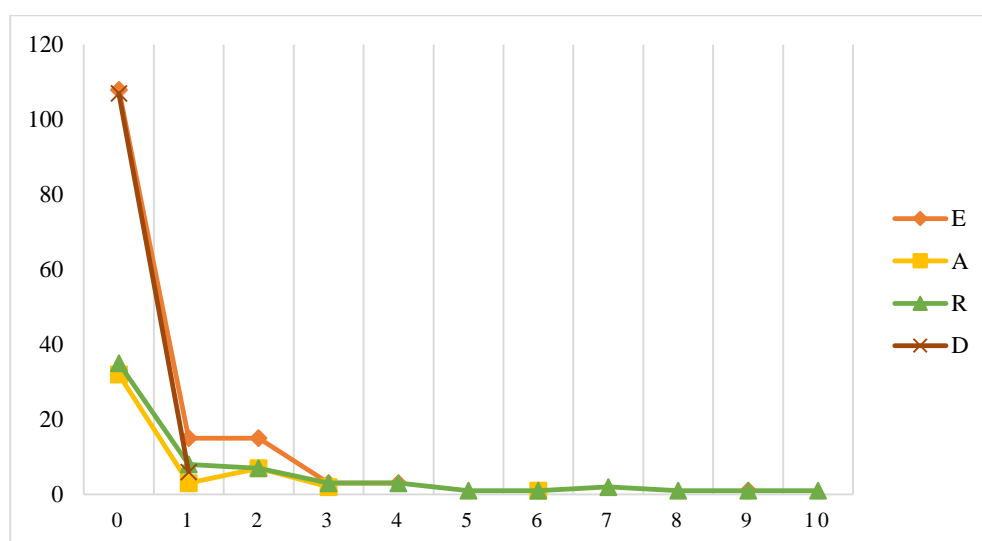
7.3.2.2 Repair category and distance

The numeric variable of distance is related to the moment of interruption since it measures the latency between occasion for the repair and the interrupted unit (either word or larger). A natural hypothesis (2a) inspired by Levelt's (1983) assumptions is that this latency for trouble detection should be shorter for erroneous words where the discrepancy between message and intention is more striking as well as more detrimental for the hearer. E-repairs should therefore stand apart from the other categories and be associated to short distances, as opposed to less disfluent repairs (A and R) where longer distances are expected. D-repairs will not be discussed any further in this section since their very definition implies immediate repair with no or little connection with the original utterance: only 6 cases out of 113 show a distance of 1 word, as in Example (34).

- (34) and its first (0.180) appearance in the world was largely as a (0.320) trading city which was which involv which involved in **trading cloth (0.560) creating making cloth and then trading cloth** (EN-intf-05)

These exceptions only correspond to cases of insertions, leaving the majority of D-repairs (especially the false-start subtype) with null distances. Figure 7.2 shows the distribution of each repair category by distance in words.

Figure 7.2: Distance (in words) between occasion and interruption by repair category



Overall, the general trend across all repair categories is a decreasing curb from null distance to 10 words between the occasion and the interruption. Another common result is the overwhelming frequency of null distance (282/367), especially for D-repairs as mentioned above. We can see another peak, albeit much smaller, around one and two words (15% of the total once combined), and almost no occurrence after 4 words (10 cases in total).

Looking at the repair categories, two further patterns emerge from this graph. First, contrary to hypothesis (2a), E- and A-repairs cannot be distinguished on the basis of this variable: they both can be interrupted after one to three words, as opposed to the expected

difference between disfluent (null or short distance) and intermediary repairs (longer distances). Examples (30) (reproduced below as 35) and (36) illustrate this shared format.

(35) a blacksmith was asked to (0.420) um put **shoes on the devil (0.520) horse shoes on the devil** and (0.400) because he was so nervous (EN-intf-08)

(36) j'ai eu mes **dix-neuf ans en Allemagne (0.410) mes vingt ans en Allemagne** plutôt
I turned nineteen in Germany (0.410) twenty in Germany rather (FR-intf-03)

In (36), the distance between the occasion for the lexical error “dix-neuf” and the interruption (unfilled pause after the full utterance) is of three words (“ans en Allemagne”) which is the same as the A-repair (generic appropriateness) in (35) with “on the devil”. These cases of delayed interruption are, however, not the majority of E- and A-repairs, which rather share a preference for null distance. Still, this structural possibility of short – but not null – distance distinguishes them from D-repairs (almost only null distance) and R-repairs (often longer distance) as we will now come to see.

The other observation is in fact this specificity of fluent R-repairs to allow for the structural possibility of leaving long distances between the occasion and the interruption. As we can see on the graph, it is the only possible category after the four-word distance (with three exceptions). Although these cases are rare, they amount to 16% of all R-repairs once combined (5- to 10-word distances). Example (37) illustrates such structures, here with a 9-word distance in an enumeration of two members, “dirigeants” and “hommes politiques”.

(37) il est évident que les (1.330) **les dirigeants français (1.867)** maîtrisent mieux la langue
 (1.740) que les dirigeants belges (1.520) et que **les hommes politiques français**
 (0.620) maîtrisent mieux la langue que les hommes politiques belges
it is obvious that the (1.330) the French leaders (1.867) master the language better
(1.740) than the Belgian leaders (1.520) and that the French politicians (0.620) master
the language better than the Belgian politicians (FR-intf-02)

We can conclude that, while small distances cannot help in distinguishing different types of repair, long distances are yet another specific feature of R-repairs. From a cognitive perspective, this result can be interpreted in terms of the seriousness of the problem: erroneous and inappropriate words or structures are easily detected by the speaker's monitoring system and should be replaced as soon as possible for the hearer's comprehension, whereas fluent structures creating long structural resonances do not require such urgency. By contrast, the longer the list member, the more fluent it might sound, since it shows some degree of planning or at least an efficient use of resources stored in short-term memory.

7.3.2.3 Repair category and way of restarting

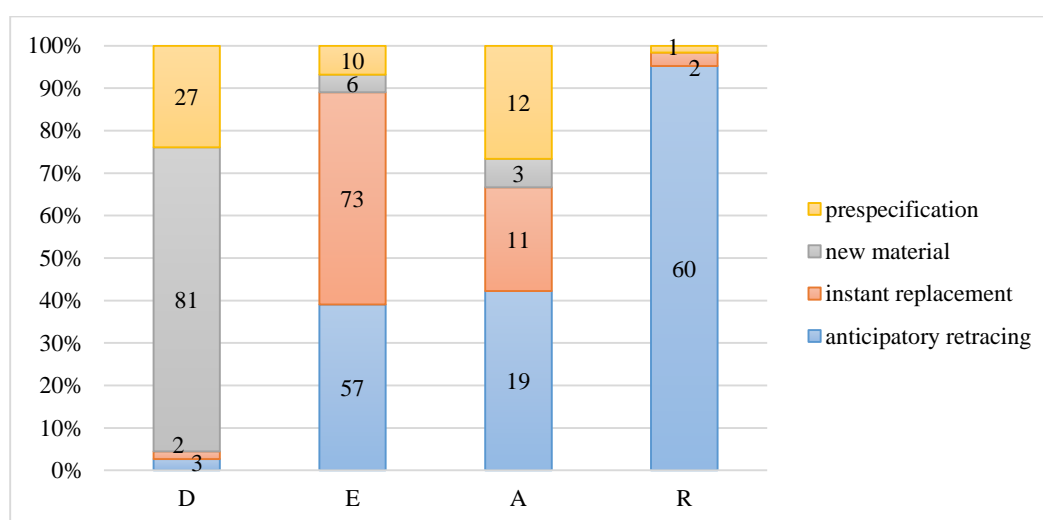
Moving from the *reparandum* to the *reparans*, no specific hypotheses were made beyond Levelt's (1983) assumption that “the way in which a repair is made is very different for A- and E-cases” (1983: 84). Bearing in mind the fact that some repair types (namely false-starts and insertions) are intrinsically associated with particular ways of restarting, Figure 7.3 shows the variation in ways of restarting among the four repair categories. In this graph, we can see that

each category differs greatly from the others, with one (sometimes two) preferred way(s) of restarting. It thus appears at first glance that the format of the repair is not random, either between the categories or within each category. Anticipatory retracing is the most frequent format in general (139 occurrences) and in particular it strikes as the prominent choice for R-repairs with only 3 exceptions in 63 cases. This strong connection is partly induced by the definition of the category, but it is not the only possible configuration, as shown by Examples (38) and (39) which are respectively cases of instant replacement (“you” by “me”) and pre-specification (with the insertion of “not”).

(38) we can argue between ourselves whether it’s **you that pays or or me that pays** (EN-intf-04)

(39) does the nurse **smile at me or not smile at me** (EN-intf-03)

Figure 7.3: Ways of restarting across repair categories



Anticipatory retracing is also very frequent in E-repairs – especially lexical errors (EL) as in (40) – and A-repairs – especially level of precision (AL) as in (41). Once again, these two specific types related to “finding the right word” present similar behaviors (cf. the occurrence of interruptions at phrase boundary, Section 7.3.2.1 above).

(40) whether we look to people to fund the cost themselves (0.340) or **whether the state** (0.220) no wrong **whether the taxpayer** funds it (EN-intf-04)

(41) there’s a lot of people that **know the machines know the original DUKWs** (EN-intf-02)

It might be considered that anticipatory retracing is a more helpful way of restarting for the hearer since it retraces back to where the repair should be integrated in the original utterance, as opposed to instant replacement or new material which do not provide such instructions. In this perspective, it is surprising to see that the same hearer-oriented strategy is used both in fluent R-repairs and in repairs related to lexical or terminological inadequacy (EL and AL), these two groups being on very different ends of the (dis)fluency scale: way of restarting alone cannot help in distinguishing fluent from disfluent repairs since they all tend to be well

integrated in the utterance (see Section 7.3.3 for a similar result on the presence of modified repetitions).

E- and A-repairs also share a higher degree of internal variation than the other two categories (D, R) which have clear preferences, viz. new material and anticipatory retracing, respectively. However, when zooming in on the repair subtypes, it appears that only R-repairs (both “lists” and “parallels”) present a uniform preference for anticipatory retracing, while all other categories show divergences in their subtypes, as can be seen in Table 7.7.

Table 7.7: Preferred ways of restarting and their proportion by subtype of repair

Repair	Format	Example	%
A	Generic	Pre-spec. “it’s ha roughly half half”	70%
	Precision	Anticip. “know the machines know the original DUKWs”	62%
	Ambiguity	Instant “you think it’s (0.300) care homes are good”	50%
E	Lexical	Anticip. “for businesses across the UK well across England”	61%
	Syntactic	Instant very more f- much more focused	62%
D	False-start	New “we have to uh (0.360) the vehicles are built”	99%
	Linearity	Pre-spec. “from there we g- then go back down”	84%

This table shows the most frequent way of restarting for each subtype of A, E and D repairs, a representative example and their respective proportion (e.g. pre-specification takes up 70% of all generic appropriateness repairs). We can see the heterogeneity mentioned above: within one repair category, each subtype has a different preferred format. This internal variation is particularly striking for A-repairs and E-repairs where the largest proportions (70% and 62%) remain smaller than in D-repairs (84%, 99%).

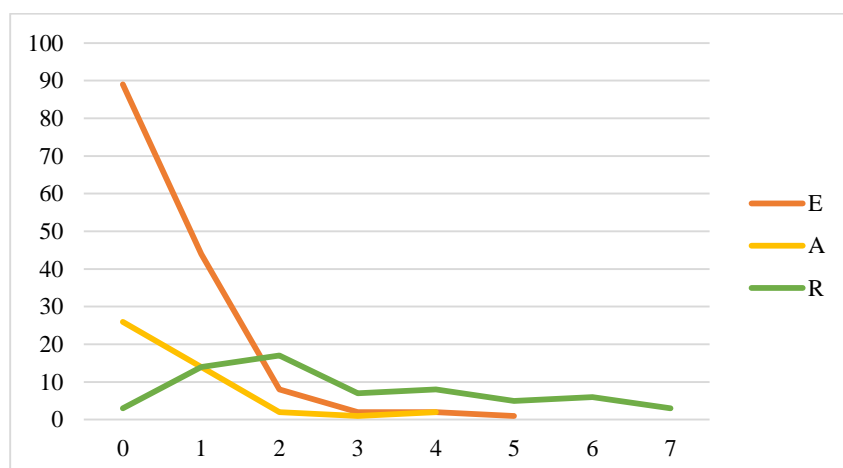
New material strikes as a specificity of D-repairs, in particular the false-start subtype: while almost absent from the other three categories, this way of restarting takes up 99% of false-starts (72% of all D-repairs). We can conclude from the evidence of D, E and A repairs that the way of restarting is strongly related to specific repair subtypes and not larger categories, which argues for a more fine-grained level of analysis (as in Table 7.7, as opposed to Figure 7.3) for this particular variable.

In the case of anticipatory retracing, the span of retracing (i.e. the number of words being repeated before the *reparans*) further allows us to refine previous observations and distinguish repair types which are seemingly similar. Figure 7.4 shows the tendencies for span of retracing across repair categories, with the exception of D-repairs which are never restarted by anticipatory retracing. We see that, overall, longer spans are decreasingly frequent except for R-repairs where the modal value (i.e. most frequent value) is two words and where longer spans (up to seven words) show substantial frequencies compared to the other categories which rarely retrace back to more than two words. In fact, E and A mostly retrace to only one word before the *reparandum*: qualitative analysis of the data shows that the retraced words are often function-words (prepositions, articles, pronouns), which indicates that speakers tend to restart at a constituent boundary (albeit local), as in Examples (42) and (43).

- (42) have very little to do with (0.120) uh (0.610) uh knowing **what's (0.239) what a baby is** about let alone a premature baby (EN-intf-03)
- (43) je ne suis pas sûr **que bien parler (0.450) que que que la notion de bien parler** c'est-à-dire de s'exprimer correctement (0.460) pour être entendu et compris (0.760) ait quoi que ce soit à voir avec l'accent
I am not sure that speaking well (0.450) that that that the notion of speaking well that is to express oneself correctly (0.460) in order to be heard and understood (0.760) has anything to do with accent (FR-intf-02)

The ES-repair in (42) restarts a *what*-complement and changes the subject-verb structure; the AL in (43) retraces to the French complementizer *que* (which is itself repeated three times) before specifying “*bien parler*” by “*la notion de bien parler*”, where the insertion is not related to a linearity issue but to a specification of the referent.

Figure 7.4: Span of retracing (in words) across E, A and R-repairs



The major information from Figure 7.4 is that fluent (R) and disfluent (EL, AL) repairs can now be formally distinguished by their respective tendency to either repeat long stretches of words (thus signalling that nothing is wrong with them), or repeat only function-words at the beginning of local phrases or subclauses. This functional specificity of AL and EL to retrace to local constituents stands in sharp contrast with the diversity of units being repeated in R-repairs, from sentence-internal phrases (cf. Example (27) repeated below as (44) for convenience) to clauses or even full utterances (45).

- (44) 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)
- (45) **it's not so much done through advertising (0.240) it's not so much done through PR** (EN-intf-08)

As a final comment on span of retracing, I would like to note that types and subtypes of repairs within a category do not differ much in terms of span length (as opposed to way of restarting), which argues for a flexible analysis with varying levels of precision depending on the variable at stake, thus promoting the use of sub-labels and macro-labels to zoom in and out of the data

when necessary, in line with the general approach to DMs and fluencemes in the rest of this thesis.

7.3.2.4 *Synthesis of formal variables*

In this section, I will try to synthesize the results obtained so far from individual variables to try and uncover preferred, most recurrent formats for different repair types, starting with the integration of information on the *reparandum* and on the repair format. Table 7.8 shows interesting associations which partly refer to the previous findings on the formal correlates of repair.

Table 7.8: Cross-tabulation of moment of interruption and way of restarting

	Anticipatory	Instant	New material	Pre-specification	Total
Within	18	17	7	11	53
Word	35	61	77	26	199
Phrase	45	3	0	7	55
Clause	14	5	2	3	24
Unit	27	2	4	3	36
Total	139	88	90	50	367

First, restarting by anticipatory retracing (i.e. repeating a few words of the *reparandum* before the repair) is a structural possibility for each interrupted unit but appears as the preferred format for interruptions at phrase, clause and unit boundary. In addition, information on span of retracing shows particularly long spans for phrase and unit boundary. My interpretation of this result is that the larger and more complete the unit, the more the speaker tends to re-integrate it in the original utterance with some anticipatory retracing: the speaker needs to signal that the on-going unit, although potentially complete and fluent, should actually be replaced by (or at least related to) an upcoming repair. Example (46) illustrates this relation between length and integration.

(46) we have **some of the poorest wards some of the poorest constituencies** (EN-intf-05)

In this case of AL-repair, the speaker retraces four words (“some of the poorest”) back to the beginning of the object phrase before substituting “wards” with the more technical term “constituencies”. It could be argued that other versions of this utterance with shorter spans of retracing (e.g. *we have some of the poorest wards poorest constituencies*, span = 1) would not be much more difficult to interpret for the listener, yet the speaker chose to make a bigger effort and repeat more words from the previous segment to better situate the new one. This confirms that integrated repairs, which are typical of R, EL and AL repairs (cf. previous section), contribute to hearer-oriented strategies and tilt the scale towards its fluent end.

On the other hand, more disruptive moments of interruption (within-word and word boundary) are more clearly “irregular” and thus do not require extra signaling of the trouble, which is evidenced by their stronger association to less conservative ways of restarting, namely

instant replacement and new material (although anticipatory retracing can occur for these levels as well). Example (47) shows such a case of disruption.

- (47) je trouve que là où **il y a** (0.967) on massacre quand même assez fort le français...
I think that the place where there is (0.967) they butcher French quite badly... (FR-intf-01)

In this syntactic error (ES-repair), the presentational structure “il y a” is interrupted (word unit) and instantly replaced by an impersonal verb phrase “on massacre” without further notice of the substitution apart from the unfilled pause. The original utterance (before the pause) is therefore left incomplete and without transition. We can now safely say that interruptions at small, incomplete constituents (word and within-word) constitute major disturbances in the syntax, often lead to complete start-overs and are associated with the group of “structural” repairs (D+ES, cf. Section 7.3.2.1), a cluster of evidence which would support their disfluent diagnosis.

Regarding the distance between interruption and *reparans*, immediate repairs are the most frequent in all formats of restarting, with a general decreasing frequency as the distance increases, across all ways of restarting. Longer distances only occur in anticipatory retracing, and within this format, mainly with short spans of retracing. In fact, it is to be expected that long distances and long spans are mutually exclusive: either the substitution targets a nearby word, in which case it needs to be integrated in previous context (short distance, long span, as in Example (48)); or it targets an initial element which introduces some stable material (long distance, short span, as in Example (49)).

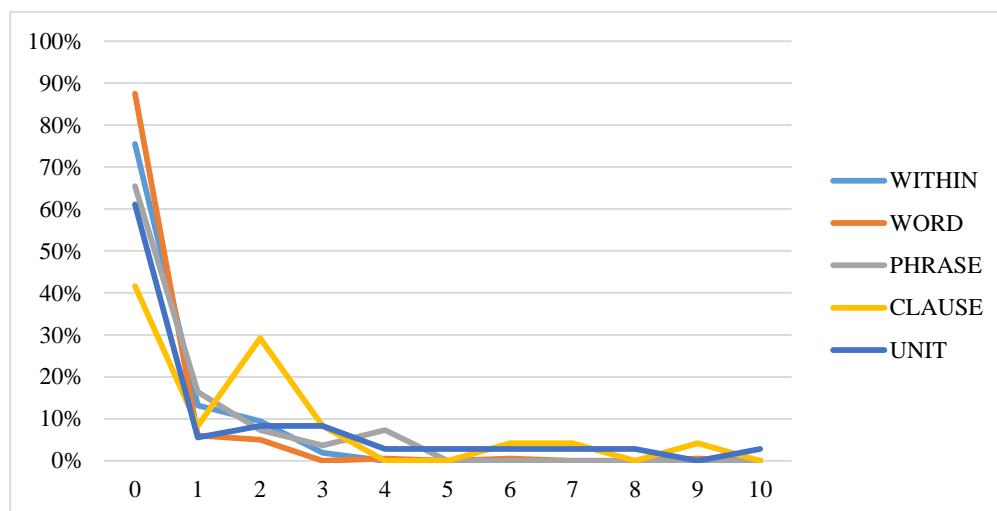
- (48) so (0.560) you have to be able to **listen** (0.550) and you have to be able to **give** them information (EN-intf-03)
- (49) the **hairstresser** comes in to our own hairdressing salon (0.470) the **manicurist** comes in the **reflexologist** comes in (EN-intf-04)

Most of these cases are R-repairs, which points to the association between long repairs (either long distance or long span) and fluency: contrary to what could be expected, repairs including lengthy repetitions are not so much related to stagnation and hesitation as to efficient use of resources available for all the participants in the local context (cf. Section 7.3.2.2; cf. Section 6.1.2 on sequence length). There is little experimental evidence in the literature on the perceptive or cognitive effect of non-identical repetitions (for the processing of identical repetitions, see Fox Tree 1995, 2001; MacGregor et al. 2009). However, Ejzenberg (2000) corroborates the fluent role of resonances: “In psycholinguistic terms, redundancy and repetition allow a speaker to set up a paradigm and slot in new information where the frame for the new information stands ready, rather than having to be newly formulated” (2000: 299), observing that such a strategy is not mastered by low-fluency learners, as Rabab’ah & Abuseileek (2012) have also shown.

Another interesting mapping of variables, only related to the *reparandum* this time, is that of moment of interruption and distance between occasion and interruption. Levelt (1983) dealt with this question at length and noticed a number of tendencies, most notably that delayed interruptions (distance \neq 0) occur more frequently at larger constituent boundaries than

immediate interruptions. He concludes that attention for trouble is enhanced towards the end of constituents, regardless of the delay between *reparandum* and interruption. The situation in the present corpus is represented in Figure 7.5.

Figure 7.5: Proportions of distance (in words) by moment of interruption



This figure shows that immediate interruption is the most frequent format regardless of the ongoing unit, with an overall proportion of 75%, although we can see internal variation depending on the unit type: clauses show a lower proportion of immediate interruption (42%), while all the other units prefer this format in at least 60% of the time. Hypothesis (2b) is confirmed by the larger proportions of immediate repairs in within-word and word boundaries (more than 70%), while longer distances (more than three words) are only possible for larger units (phrase, clause and unit). In addition, within-word interruptions are not entirely restricted to erroneous words, since the data shows 13 occurrences (25%) where the truncated word is not the occasion for repair. Considerations of repair type are not particularly relevant to further distinguish these patterns.

To sum up, recurrent patterns of repair types with their prototypical format emerge from the synthesis of formal and functional variables: “structural” repairs (D, ES) tend to interrupt local constituents and be repaired in start-overs (50); fluent repairs (R) are, by definition, more conservative (i.e. maintain large stretches of linguistic material in the different parts of the repair) and more respectful of syntactic boundaries (51); “lexical-search” repairs (EL, AL) are somewhat intermediary, sharing with R-repairs a tendency to be well integrated in the local syntax, regardless of the degree of lexical inadequacy (from error to inappropriateness) which cannot be formally distinguished (52). Examples (50)-(52) illustrate each of these typical configurations.

(50) we actually have **to uh (0.360)** the vehicles are built (EN-intf-02)

(51) so **they get better at dance through that they get better at dance through (0.540)**
being at school as well (EN-intf-06)⁶¹

(52) **il y a des fautes euh il y a des expressions** qui sont mal utilisées
there are mistakes uh there are expressions which are wrongly used (FR-intf-01)

I will now turn to the relation between the variables defined in Levelt's (1983) framework and my own approach to the annotation of fluencemes.

7.3.3 RMs in repair

The following sections will strive to answer the first research question mentioned in Section 7.1.4, namely the extent to which fluencemes, especially modified repetitions (RMs) and discourse markers (DMs), are involved in different types of repair. I will start with RMs to see their proportion in fluent and disfluent repairs. The subcorpus of face-to-face interviews contains a total of 4,439 sequences of fluencemes: 182 of these sequences include an RM, of which 147 are involved in a repair as coded in the present analysis, amounting to a very high coverage of 81% of all RMs.⁶² The remaining 20% of RMs not involved in one of the four repair types defined above mostly correspond to cases of completed truncations, as in the following example:

(53) was it quite difficult to integrate **in the s- in the sort** of village life (EN-intf-06)

Here, the repetition of “in the” does not introduce an overt repair in the sense of a modification of previous linguistic material, but completes the truncated word “sort” (actually part of the DM “sort of”) in a covert repair which cannot be interpreted in terms of repair types. In fact, in light of such examples, the theoretical choice which we made in the annotation protocol (Crible et al. 2016, Appendix 2: 384) to annotate truncations with modified (RM+TR) instead of identical repetitions (RI+TR) is debatable and should perhaps be reconsidered in a revised version of the protocol: the absence of modification in cases such as (53) vouches for its categorization as identical repetition of (incomplete) words, hence labeled as RI+TR, while RMs should be reserved for actual change of form or content, in line with the opposition between covert (RI) and overt (RM) repair.

Table 7.9 below represents the occurrences of RMs in the different repair types: “No RM” corresponds to the absence of RM, “RM” to presence and “N/A” to the rare cases where no fluencemes were originally annotated in the repair sequence. We can see that each repair category is associated with a specific pattern regarding the presence or absence of RMs. The most extreme cases are D-repairs on the one hand, with a large preference for no RMs (84%), and R-repairs on the other showing the reversed proportion (85% of RMs included) with a few exceptions of no annotation. In A-repairs, there is no difference between absence and presence, even when looking at the specific types. Cases of “zero annotation” strike as most frequent in

⁶¹ This example could be mistaken for a case of ambiguity (AA) between “through that” and “through being at school” but the (audio) context makes it clear that it is rather a case of enumeration (R), the pronoun “that” actually referring to a previous topic (taking after-hours dance lessons).

⁶² This figure shows the proportion of RM sequences, as extracted from the original annotations, which were later coded as involving one of the four repair types. If we start from repair sequences, only 40% involve a RM.

this category even in raw frequencies, considering the small number of total occurrences: these A-repairs are not marked formally by any linguistic deviance but present an equivalence of content, in accordance with the selection criteria (Section 7.2.1) and as shown in Example (54).

- (54) I'm responsible for the development of research activity (0.600) **knowledge exchange which means working closely with local industry** (EN-intf-04)

Here, the speaker introduces the definition “knowledge exchange” by a dedicated verb “which means” in a relative clause, thus marking the semantic equivalence without any cue of substitution or formal resonance. There are 11 of these cases (A-repairs with “no annotation”) in the corpus.

Table 7.9: Proportions of modified repetitions across repair types

	No RM	RM	N/A
Delay	84%	13%	3%
Error	52%	42%	6%
lexical (EL)	40%	50%	10%
syntactic (ES)	62%	35%	3%
phonetic (EF)	83%	17%	0%
Approp.	38%	38%	24%
precision (AL)	29%	38%	33%
ambiguity (AA)	50%	29%	21%
generic (A)	40%	50%	10%
Resonance	0%	86%	14%
Total	51%	40%	9%

The picture is more complex for E-repairs, with only a slight majority of no RMs, while the EL and ES types show the same reversal as for the way of restarting: ES-repairs, associated with instant replacement, tend not to involve RMs as in Example (55); the difference is small for EL-repairs but slightly in favor of the presence of RMs, which is consistent with their preferred format of anticipatory retracing (cf. Example 52).

- (55) at the other end of the spectrum **we're I'm** responsible also for innovation (EN-intf-05)

These results suggest that RMs are quite frequent in very fluent repairs (R), quite rare in very disfluent repairs which once more correspond to the “structural” group (D, ES), and equally absent or present in the intermediary repairs (EL and AL): the presence of RM is therefore entirely relevant to the formal scale of (dis)fluency and corroborates the information from other variables discussed in the previous sections. However, RMs cannot be used independently to distinguish repair types, especially between fluent and intermediary repairs, since their presence is a structural possibility for all categories (only their absence excludes R-repairs). Still, it is striking that, in my data, the biggest proportion of RMs is involved in fluent R-repairs, in more than one third of all RM occurrences. In my view, this not only illustrates the high ambivalence

of RMs but also validates the formal approach to all fluencemes: if the annotation had only covered “disfluent” fluencemes (provided an operational definition of such a concept could be designed), it would have missed a lot of data (“silence” in computational terms) which are structurally identical, albeit functionally different.

In the next step, we can integrate the information at the fluenceme level (i.e. presence or absence of an RM) and that of the formal variables detailed above, in order to overcome the limitations of individual variables which particularly failed to distinguish fluent and intermediary repairs on a number of occasions (namely integration with anticipatory retracing, interruption at phrase boundary, and presence of an RM). Most of the formal variables, such as span of retracing and distance between occasion and interruption, present no major effect. However, the moment of interruption is more relevant to the present analysis, as can be seen in Table 7.10 which reports on the proportion of RMs included in the different repair types across interrupted units (figures in this table therefore only show cases where RMs are present, leaving out “no RMs” and “no annotation”). Proportions are completed by raw frequencies (in brackets) since the former could otherwise be misleading, especially in the case of delay repairs (few occurrences, large proportions).

Table 7.10: Distribution of RMs in interrupted units by repair type

	Within	Word	Phrase	Clause	Unit	Total
Delay	20% (3)	60% (9)	13% (2)	7% (1)	0% (0)	100% (15)
Error	30% (18)	49% (30)	16% (10)	3% (2)	2% (1)	100% (61)
Approp.	12% (2)	41% (7)	23% (4)	18% (3)	6% (1)	100% (17)
Resonance	0% (0)	0% (0)	39% (21)	17% (9)	44% (24)	100% (54)
Total	16% (23)	31% (46)	25% (37)	10% (15)	18% (26)	100% (147)

This table shows a sharp contrast between incomplete (within-word and word) and complete units (clause and unit), while phrase boundaries stand in an intermediary position (cf. highlighted cells). This contrast opposes E-repairs on the one hand to R-repairs on the other: when the interruption affects a small unit, the presence of an RM is almost always linked to E-repairs, as in Example (56); when the interruption affects a larger constituent, the presence of an RM is almost always linked to R-repairs (57).

(56) I was thinking about **loss of independ (0.360) or fear of loss of independence** (EN-intf-04)

(57) **they’re the important things because they’re the normal (0.460) things** (EN-intf-03)

In other words, RMs can be involved in disfluent repairs but these are easily identifiable by the moment of interruption: the ambivalence of RMs can therefore be resolved by this formal variable. Interruptions at phrase boundary remain somewhat problematic in this disambiguation endeavor since both E- and R-repairs show a number of RMs at this level. All in all, the interaction between moment of interruption and presence of an RM shows that the strong associations identified between repair types and their preferred unit can be re-used and mapped

onto the behavior of RMs: if we assume that R-repairs are more fluent than E-repairs, we can deduce the degree of fluency of a particular occurrence of RM by looking at the unit size, without resorting to a qualitative coding of repair type. We could easily reformulate these conclusions in basic formulas, for instance:

- if RM= “YES” + moment of interruption= “UNIT”, then the degree of fluency is high;
- if RM= “YES” + moment of interruption= “WORD”, then the degree of fluency is low.

These generalizations are obviously submitted to precautions related to the small sample size and the resulting absence of statistical modelling, however I believe that they are meaningful and would be interesting to test on other fluencemes apart from RMs.

The final step for this investigation of RMs in repairs is to focus on the fluencemes of false-starts (FS), in order to see whether they are the counterpart of RMs, that is, whether FS are involved in repairs where RMs are not, and vice versa. From the corpus-based analysis of sequence types, I claim that false-starts (along with truncations, with which they form the category of F-sequences) are more disfluent than other fluencemes (cf. Sections 6.2, 6.5). As a result, they should be strongly associated to disfluent repairs, bearing in mind that one subtype of D-repairs is circularly defined as false-starts. Table 7.11 compares the proportions of RM and FS fluencemes across the different repair types in order to test the hypothesis of their repulsion.

Table 7.11: Proportions (and frequency) of RM and FS fluencemes included across repair types

	FS included %	RM included %
Delay	69% (78)	13% (15)
Error	25% (36)	42% (61)
lexical (EL)	12% (9)	50% (36)
syntactic (ES)	40% (27)	35% (24)
phonetic (EF)	0% (0)	17% (1)
Appropriateness	9% (4)	38% (17)
precision (AL)	5% (1)	38% (8)
ambiguity (AA)	21% (3)	29% (4)
generic (A)	0% (0)	50% (5)
Resonance	2% (1)	86% (54)
Total	32% (119)	40% (147)

Three major differences can be observed:

- FS are much less frequent in lexical error (EL) repairs than RMs which take up almost half of the occurrences;
- FS are almost absent in appropriateness (A) repairs, as opposed to RMs where they frequently occur;
- FS are completely absent from R-repairs, with only one occurrence of FS (which actually co-occurs with a RM).

The only similarity between FS and RM is in syntactic errors (ES-repairs) where they occur in equal (raw and relative) frequencies. Although both frequent options of ES, they almost never co-occur (with one exception) which means that even when both fluencemes can express the same repair type, they very rarely combine with each other. The high frequency of the false-start fluenceme in both D and ES-repairs confirms the non-ambivalent, disfluent nature of false-starts and constitutes another sign of the conceptual and formal proximity of these two “structural” repair types.

In the same line of reasoning, a closer investigation of the D-repairs which do not contain a false-start reveals that they often correspond to cases of truncation (and a few cases of “zero annotation”) as in Example (24) reported here as (58), which is another validation of the grouping of FS and TR into the same macro-label (namely F-sequences).

(58) it’s more of the Liverpool **acc-** but I can certainly tell the difference (EN-intf-03)

All in all, it appears that false-start fluencemes are much more restricted than RMs in terms of possible repair types, since they mostly occur in the two types of structural repairs. We can conclude from this comparative analysis that FS is indeed the counterpart of RM, since they each prefer different repair types and rarely co-occur, even in ES-repairs, which is the only category where they are both frequent. Another conclusion is that the assumed ambivalence of fluencemes is not valid for false-starts which stand clearly on the disfluent end of the scale, as already suggested by the analyses of register and DM functions in Chapter 6.

7.3.4 DMs in repair

In the interviews subcorpus, 1,917 sequences of fluencemes contain one or several DM tokens (909 in English, 1008 in French), of which 134 are involved in a repair sequence. Out of these 134 DMs, 85 occur in the editing phase of a repair sequence, i.e. between the *reparandum* and the *reparans*. All in all, few DM sequences are involved in repair (7%) (only 4% in their editing phase), especially compared to the proportion of RMs involving repairs (81%).⁶³ A similarly low frequency was already observed in Pallaud et al. (2013a), who found that only 10% of “disfluent interruptions” include DMs (cf. Section 3.3.1). 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 Section 6.1).

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*. In other words, 35% of overt repairs include DMs, which is quite similar to the proportion of repairs involving RMs (40%, see Table 7.9). In other words, the occurrence of DMs does not imply that of repairs (only 7% of all DMs), but repairs, when they occur, do seem to contain DMs, although only in 1/3 of the time. This could be interpreted as evidence of their high fluency since they are not often involved in structural and/or lexical errors or ambiguities. However, such a conclusion would overlook the fluent uses of R-repairs and the intermediary degree of EL and AL-repairs: their tendency to cluster with pauses and, to

⁶³ As for RMs, these proportions report on the number of DMs involved in one of the four repair types, and not the amount of repairs involving one or several DMs (cf. Footnote 62).

a lesser extent, identical repetitions (as shown in the previous chapter) 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.

7.3.4.1 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 7.12 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 7.12: Distribution of DMs across repair types and positions in the repair (if any)

	Presence of the DM		Position of the DM		
	NO	YES	EP	Periphery	Repair
Delay	57% (64)	43% (49)	31	15	3
Error	70% (102)	30% (44)	25	15	4
phonetic (EF)	83% (5)	17% (1)	0	1	0
lexical (EL)	69% (50)	31% (22)	17	4	1
syntactic (ES)	69% (47)	31% (21)	8	10	3
Appropriateness	60% (27)	40% (18)	11	6	1
generic (A)	40% (4)	60% (6)	4	1	1
ambiguity (AA)	57% (8)	43% (6)	2	4	0
precision (AL)	71% (15)	29% (6)	5	1	0
Resonance	70% (44)	30% (19)	16	2	1
Total	65% (237)	35% (130)	83	38	9

We can see that, when a DM is present in the repair, it is mostly located in the 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 (59) and (60).

(59) 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)

(60) <BB_1> are you responsible for (0.230) organising that or somebody in your department

<BB_4> **well it** as individual nurses we are allocated to care for babies (EN-intf-03)

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: (i) initial or preliminary commitment to abide to the “temporal imperative” of speech even though the utterance is not entirely planned; (ii) suspension of speech, usually after the first (function) word; (iii) hiatus (such as the filled pause “uh” in Example (59), absent in Example (60)) and (iv) restart (with new material in the two examples). 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 7.12 that each repair type occurs more frequently without a DM, DMs still strike as 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.

(61) find somebody in the hospital first (0.370) to see **if you know because** it's easier (EN-intf-03)

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

(63) 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.

7.3.4.2 DM lexemes and functions

A first observation of the DM lexemes in the corpus confirms the contrastive hypothesis (hypothesis 3) 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.

Table 7.13 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 (cf. Section 7.2.4.1). 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 6.5). A typical example of *reformulation* in EL-repair can be found in Example (28) reproduced here as (64).

- (64) 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)

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 (65).

- (65) les Français maîtrisent bien leur langue euh les Belges maîtrisent à 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)

Table 7.13: Functions and frequent lexemes of DMs in the editing phase

Function	Nb of occ.	Lexemes FR	Lexemes EN
Reformulation	27	<i>enfin ; ou</i>	<i>or ; well</i>
Addition	14	<i>et</i>	<i>and</i>
Monitoring	12	<i>hein ; vous savez</i>	<i>you know</i>
Specification	8	<i>enfin ; c'est-à-dire</i>	<i>actually</i>
Topic-resuming	5	<i>donc</i>	<i>but ; so</i>
Punctuation	4	<i>bon</i>	-
Temporal	4	<i>alors ; quand</i>	-
Cause	3	<i>comme</i>	<i>because</i>
Opposition	3	<i>mais ; bon</i>	<i>but</i>
Alternative	3	<i>ou</i>	<i>or</i>
Emphasis	2	<i>du moins</i>	-
Motivation	2	-	<i>because</i>
Condition	2	<i>quand</i>	-
Ellipsis	2	<i>etcetera</i>	-
Hedging	1	<i>je diras</i>	-
Closing	1	<i>voilà</i>	-
Concession	1	<i>mais</i>	-
Contrast	1	-	<i>and</i>
Total	95	-	-

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 (i) their classification as “Potentially Disfluent Functions” (Section 6.5) and (ii) the corpus results in Section 6.6 which showed the association of interpersonal DMs to F-sequences, as in Example (66) where “you know” co-occurs with two truncations and a filled pause to signal trouble in lexical access (EL).

- (66) 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. *Reformulation* and *monitoring*, which are two of the “Potentially Disfluent Functions” (PDFs) I proposed in Section 3.4, 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 (67).

- (67) 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) reminds that of the *monitoring* function: in this perspective, *monitoring* and *punctuation* are similar, which supports the proposal in Crible & Degand (under review) to categorize them as two variants of the same function, one interpersonal and the other sequential.

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.

- (68) <VAL_2> pensez-vous que l’accent peut (0.327) influencer la façon dont on est
perçu
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 (FR-intf-01)

In (68), “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 (23) with “c’est-à-dire”). 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:

- (69) 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 Ciabbari (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 (cf. Example (65) above), which can be explained by the tendency of spoken language to be underspecified and to rely on context to disambiguate polysemous 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 (70) which shows no formal resonance.

(70) 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 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: reformulative DMs mostly occur in EL-repairs (cf. Example (64)); additive DMs in “lists” R-repairs (cf. Example (65)); monitoring DMs in disfluent repairs (cf. Example (66)). Occurrences of the other functions are too rare to identify similar meaningful patterns. To this first limitation of sample size, I would like to add a potential issue of circularity: to what extent does the function of the DM, when present, influence the repair category during the coding process? Repair types were shown in the previous sections to be strongly associated with formal characteristics, while DMs only occur in 35% of all repair sequences, thus limiting their potential impact on the analysis. However, all annotation layers were available during the procedure, and it might be the case that the observed patterning between DM functions and repair types is actually the result of some circularity between these two levels of analysis. Both DM functions and repair types are pragmatic, more qualitative and subjective variables which partially overlap on the conceptual level. In the next sections, I will turn to more formal variables in order to avoid such interdependence.

7.3.4.3 Functions of DMs and format of the repair

Starting with the moment of interruption, it appears that only the *reformulation* function is not restricted to any unit, although this might be due to its higher raw frequency. However, it is surprising to observe that half of the occurrences of *reformulation* occur in repairs which are interrupted at large constituent boundaries, especially at unit boundaries, as in Example (19) reproduced here as (71):

(71) on ne parle plus comme cela non non (0.890) me semble pas **en tout cas**
we don't talk like that anymore no no (0.890) I don't think so en tout cas 'at least'
 (FR-intf-03)

In this example of generic appropriateness (A-repair), the reformulation takes scope over the whole previous utterance by qualifying its epistemic strength, thus signaling a rather distant backward-looking move. It is striking that, out of the minority of interruptions at unit boundary which do not correspond to R-repairs (11 cases across D, E and A against 25 in R), seven cases

include a *reformulation*, while no occurrence of *reformulation* is involved in R-repairs: in other words, although both R-repairs and reformulative DMs share a formal attraction to large units, they are mutually exclusive, which could be seen as evidence for the independence of repair types and DM functions discussed above. Apart from these large constituent boundaries, most repairs containing a reformulative DM are interrupted at word boundary, introducing the repair as soon as possible.

Similarly, additive DMs mostly occur in repairs interrupted at unit or clause boundary (11 out of 14 cases), which is explained by the relation between *addition* and R-repairs. *Monitoring* and *specification* are also dependent on their associated repair type and respectively occur in utterances interrupted at local (cf. D- and ES-repairs at within and word boundary) and intermediary units (cf. A-repairs at word and phrase boundary). However, DMs are not entirely bound by their repair type and might be responsible for formal differences: interruption at unit boundary, generally rare in the corpus (10% of all repair occurrences, second to least frequent format) becomes the second most frequent format when a DM is present in the editing phase thanks to its association with the *reformulation* and *addition* functions. It would thus seem that the presence of a DM has an impact on the frequency of this repair format, which might be related to the typically initial position of DMs and their role as connectors of full clauses instead of local constituents (cf. their initiality and independence from other fluencemes, Sections 5.2.1 and 6.3.2).

With respect to the way of restarting, a similar co-variation of DM functions and their typical repair type can be observed, for instance with the diversity of repair formats for *reformulation* (cf. its occurrence in D, E and A repairs) or the restriction of *addition* to anticipatory retracing (cf. its connection to R-repairs). Anticipatory retracing contrasts with restarts by new material in terms of the types and number of different functions which occur in each of them: DMs in anticipatory retracing mainly express functions which are conceptually related to a repair type, such as *reformulation* (EL), *monitoring* (D+ES), *alternative* or *addition* (R); DMs in new material, on the other hand, are much more diverse and can express almost any function, such as *closing boundary* or *causal relation* as in Example (72).

- (72) je crois pas **que** (0.260) **et comme** cette politique est exactement euh (0.933) en sens inverse de ce qu'elle devrait être
I don't think that (0.260) *and comme* 'since' *this policy is exactly uh* (0.933) *the opposite of what it should be* (FR-intf-01)

The tendency illustrated above shows that DMs are often the first word of a new utterance after an interruption (typically restarted with *new material*), where they signal the end of the problematic construction: this strategy to start over with DMs (regardless of their function) is further evidence of their positive role on fluency, being part of the solution rather than the problem (Auer 2005). This last result echoes Merlo & Mansur's (2004) study where DMs ("lexical pauses" in their terms) are shown to occur pervasively in different discourse parts to introduce new moves or steps in the description. The signaling function of DMs therefore vouches for their ambivalent – if not fluent – role in the otherwise disfluent context of interruptions, as in Example (72), and reflects their deep connection with issues of linearity (ordering utterances, announcing upcoming material).

7.3.4.4 DMs and RMs

Mapping the occurrence of DMs and 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 7.14 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 7.14: Presence and position of DMs and RMs

	N/A	No RM	RM	Total
No DM	10% (24)	48% (113)	42% (100)	100% (237)
DM	6% (9)	58% (80)	36% (49)	100% (138)
Editing phase	8% (7)	61% (51)	30% (25)	100% (83)
Repair	14% (2)	57% (8)	29% (4)	100% (14)
Periphery	0% (0)	51% (21)	49% (20)	100% (41)
Total	9% (33)	51% (193)	40% (149)	100% (375)

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 30% 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 (73).

- (73) 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 thrice 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 58%, which is more than both the proportion of co-occurrence (36%) and that of joint absence (no DM, no RM, 48%). 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 80 DMs in the editing phase) as in Example (74).

- (74) 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 (75) and (76).

(75) the mums remember you **and** the dads remember you (EN-intf-03)

(76) is it going to look like dad is it going to look like mum (EN-intf-03)

In (75), 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 (76), 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. This final result suggests to shift the focus of the next section to the other fluencemes occurring in the repair, and in the editing phase in particular.

7.3.5 Other elements in the repair and editing phase

So far, RMs have been identified as a major structural component in the large majority of repairs, while DMs are typical of the editing phase, although to a smaller extent. Other fluencemes have been found in the data to be recurrent elements in the repair structure, namely truncations (TR), lexical and parenthetical insertions (IL, IP) (cf. Section 7.2.4.2). These fluencemes often combine with an RM and can be expected to show strong associations with specific types of repair based on their definition: insertions are conceptually related to the “linearity” subtypes of D-repairs (“local” for IL, “global” for IP); truncations could be motivated by the urgency to interrupt erroneous words, hence being frequent in EL or ES.

In the interviews, truncations (TRs) are the most frequent fluencemes and occur in 64 repairs without any restriction in terms of repair types (except AL), although they are mostly involved in disfluent repairs such as EL, ES and D, as expected. The presence of TRs can therefore safely be related to the disfluent end of the scale, with very few exceptions (only three R-repairs contain a truncation, out of the 63 occurrences). Lexical insertions (ILs) also occur in almost all repair types (except EF) and appear to be related to structural repairs in particular (31

cases of D and ES out of 54): either the inserted material introduces an initial re-start (“LocLin” subtype of D) or targets an utterance-internal element (ES), typically verbs as in Example (77).

(77) I know my nephew sometimes’ll **to speak to** me in the Liverpool accent (EN-intf-03)

ILs often co-occur with TRs (15 cases), especially in disfluent repairs, which supports the claim that speakers tend to signal their trouble with combinations of cues. Lastly, parenthetical insertions (IPs) are quite rare in the data (cf. their low frequency in the whole *DisFrEn* corpus) with only 10 occurrences, of which six are involved in the “GloLin” subtype of D-repairs. The distribution of TR, IL and IP in the different repair types is not highly informative insofar as the very definition of these fluencemes is criterial for the internal structure of specific repair types (especially structural issues), one possible exception being truncations which equally occur in lexical and syntactic repairs.

Regarding the elements in the editing phase, 21 types of fluenceme sequences were produced in the data, in addition to DMs and empty editing phase. The most frequent fluencemes are unfilled pauses (UPs), either alone or clustered with other fluencemes: 143 repairs (out of 367) contain UPs, including 28 clusters with filled pauses, across all repair types. Pauses are therefore more frequent than DMs in the editing phase (and in other positions as well), which confirms their highly ambivalent role, either as punctuating and structuring devices or as hesitation symptoms: their high frequency and wide distribution confirms them as the most flexible fluencemes in the typology, above RMs and DMs, and calls for their in-depth investigation as an independent object of study, in line with current (experimental) prosodic research (e.g. Candéa 2000; Bosker et al. 2014; Lundholm 2015) and the early research on fluency (cf. Section 1.2).

Other notable elements in the editing phase are identical repetitions (RI), which are the fourth most frequent fluenceme in *DisFrEn* overall, and here involved in 21 repairs, mostly disfluent and structural repairs (D and E; none in R). This negative reading of RIs should however not be generalized to all their contexts of appearance since RIs do not exclusively occur in overt repairs: RIs in repairs only amount to 5% of all the 449 occurrences in the interviews subcorpus, which means that this fluenceme, much like DMs, is more associated to covert than overt repair (cf. Section 7.3.4). The cognitive-functional motivations of covert repairs are very diverse and mostly related to planning strategies and anticipatory effects (cf. Section 7.3.4.1): RIs can therefore be categorized as “forward-looking disfluencies” (Ginzburg et al. 2014), which do not look back on problematic utterances (as in disfluent overt repairs) but announce upcoming material.

Coordinating conjunctions were identified as recurrent elements in the editing phases of R-repairs, which is the case for 20 occurrences mostly represented by English *or* and French *et* (‘and’), in utterance-internal uses, which disqualifies them from the DM category. Example (78) illustrates three of these cases in one utterance.

(78) I can certainly tell the difference between (0.420) somebody who’s truly Liverpoolian (0.330) **and** somebody who has a Cheshire accent **or or** a north Wales accent (EN-intf-03)

Finally, explicit editing terms (ET in our fluenceme typology, not to be confused with the editing phase in a repair) are quite rare (10 occurrences, including seven in EL-repairs) and very varied in forms, with short dedicated expressions such as “wrong” or French “plutôt” (‘rather’) and longer free phrases as in “that’s what I was looking for” (EN-intf-02) or “je ne sais pas” (‘I don’t know’).

To conclude, a large number of fluenceme types can be involved in the segments of a repair and in the editing phase. Some of them are quite specific to one repair type and its associated degree of fluency: identical repetitions (RI) in D- and E-repairs; coordinating conjunctions (CC) in R-repairs; explicit editing terms (ET) in EL-repairs; insertions (IL, IP) in structural repairs. By contrast, the most frequent ones (pauses, DMs, truncations) are more flexible and scattered across all repair types, although previous sections in this and other chapters have suggested cognitive-functional interpretations based on formal variables, clustering tendencies and register variation.

7.4 Summary and interim discussion: low quantity, high quality?

I will proceed to the summary of the main results of this chapter, opening to more theoretical conclusions before discussing methodological issues raised by the present analysis. In response to hypothesis (3), a first conclusion of this chapter is, once more, the similarity of English and French texts in terms of the distribution and format of repairs, which echoes the relative absence of major crosslinguistic differences observed in the previous chapters of this thesis (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 linearity and 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.

In response to hypothesis (2), the attempt to build a formal scale of fluency where different degrees of fluency are associated with formal, more objective features was successfully met by a number of cross-tabulations which converge in identifying the following three major patterns: structural repairs (D, ES) typically interrupt short units and introduce start-overs with fresh material (low fluency); lexical repairs (EL, A) are more integrated in the original utterance and cannot be formally divided into error-correction and appropriateness-adjustment (intermediate fluency); resonances (R) are strongly related to larger segments, long

distances and high integration in the utterance (high fluency). Not all variables in the coding scheme are equally relevant to distinguish repairs (sub)types:

- *moment of interruption* stands out as highly relevant for the basic distinction of these three degrees of fluency;
- *distance* between occasion and interruption cannot distinguish E- from A-repairs, but clearly differentiates R-repairs from the other types. This association between long distances and fluent repairs echoes Auer's (2005) conclusion on the use of delayed self-repairs by skilled speakers as a strategy to handle long and complex turns (as in interviews) in order to cope with linearization issues;
- *way of restarting* cannot distinguish intermediate from fluent repairs (EL, A, R), unless the specific repair type is taken into account;
- *span of (anticipatory) retracing* is one of the rare variables that differentiates fluent and intermediate repairs.

All in all, by mapping several variables together, it appears that repairs taking scope over large constituents (typically, R, also EL and A) tend to be more conservative and more integrated in the linguistic structure, whereas repairs interrupting short units often lead to complete or partial re-starts. There seems to be a relation between length (both long distances and long spans of retracing) and high fluency, which I tentatively connected with the notion of short-term memory: by recycling large stretches of words, either before (long span of retracing) or after the occasion for repair (long distance), speakers manifest their efficient use of stored material while lowering the cognitive demands on the hearer's part, since the latter does not have to process new material. Lengthy repetitions therefore seem to contribute to hearer-oriented strategies, along with other features of repairs such as their integration in the original utterances and the structural or semantic resonances between the two segments. It appears that speakers resort to specific formats to help their interlocutor interpret the repair, and that these cues are only used in repairs of high or intermediate fluency.

Another conclusion is therefore that the more trouble the speaker is experiencing, the less available s/he is to accommodate the hearer's need, which might be seen as a sort of vicious circle: when the problem is serious (usually regarding the structure of utterances), both the speaker and the hearer need to attend to their own needs (of production and interpretation, respectively); when the problem is less serious (related to finding the right word) or when there is no problem at all (fluent resonances), the speaker can and does provide their interlocutor with converging cues on how to interpret the on-going repair. The present hearer-oriented view of repair situates this study in the lineage of the French classics on reformulation (especially Charolles & Coltier 1986) as well as Cuenca's (2003) definition of reformulation. The strong role of repetition within repair and reformulation was also present in De Gaulmyn (1987) long before the recent experimental and corpus-based work discussed in Section 7.3.2.4.

In response to hypothesis (1) regarding the relation between repair and fluencemes, RMs and DMs appear to have reversed proportions of involvement in repairs overall, with most repairs (80%) containing an RM, while only 22% include at least one DM in the editing phase (up to 35% all positions combined). 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 presence of different fluencemes in repairs was then treated as a formal variable in itself, in order to answer the following question: is it possible to diagnose the relative fluency of a sequence along the tripartite scale (structural, lexical, resonance) based on the types and combinations of fluencemes it contains? The occurrence of RMs revealed to be a structural possibility for all repair types and therefore not sufficient evidence of a particular fluency degree. This being said, the overall frequency of RMs across repair types decreases with the degree of fluency, which could indicate its stronger relation to fluent than disfluent repairs. By contrasting RMs and false-starts in terms of frequency by repair type and (absence of) co-occurrence, the fluent interpretation of RMs was confirmed: RMs build the foundations of conservative repairs, whereas false-starts are major disruptions in the syntax.

By contrast, 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 false-start), 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. Although the association of DM function and repair type might be partly circular, independent effects have also been identified where the presence of a DM has an impact on the preferred format of the repair (cf. the frequency of "unit" boundaries).

Apart from RMs and DMs, other fluencemes were found to frequently occur in repairs, especially truncations (in the *reparandum*) and lexical insertions (in the *reparans*). Focusing on the editing phase, pauses strike as the most frequent fluenceme, before DMs and identical repetitions, which mirrors the overall frequency of these fluencemes in the whole subcorpus of interviews. Pauses, like DMs, are not restricted to any repair type, which confirms the high ambivalence of these two fluencemes. Moreover, 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. Identical repetitions, by contrast, tend to occur in disfluent repairs, although the majority of occurrences in the corpus are actually not involved in an overt repair, a characteristic which they share with DMs. Apart from a few other cases of explicit editing terms and coordinating conjunctions, one third of all repairs include no fluenceme in their editing phase. These results strikingly differ from the classic references in the field which state the necessary presence of a marker between the two segments of a reformulation. They also partly invalidate the “one marker for one reformulation type” claim (Charolles & Coltier 1986; Rossari 1994): the more frequent a fluenceme in the editing phase, the less specific it is to a particular repair type or format (cf. the ambivalence of pauses or, to a lesser extent, the occurrence of reformulative DMs in different types and formats of repair).

Taking a step back from the formal 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. Apart from the different degrees of fluency on the scale, which I have repeatedly shown to be associated to formal features of the repairs, three patterns were confirmed:

- F-sequences, comprising the truncations and false-starts fluencemes (cf. Section 7.3.3) and their association to interpersonal DMs (discussed in Section 7.3.4.2);
- “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 (under review).

What these results also confirm is the intuition that the notion of reformulation, as defined in formal (Section 7.1.1) or contrastive linguistics (Section 7.1.2), is narrower than the notion of repair in the present approach, which also includes issues of structure or linearization, as well as modified repetitions for 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. The focus on RMs and DMs in this chapter allowed me to clarify the relation and partial overlap between the notions of repair, reformulation and repetition.

This chapter also involves the notion of (non-)linearity: RMs are the perfect example of fluencemes which can combine backward (error correction) and forward scope (list construction) while contributing to the speaker’s linearization effort, for instance with

insertions in RM+IL patterns. I have tried to show how these different directions or moves correspond to more or less fluent strategies which are themselves associated with formal correlates, thus resolving the ambiguity or flexibility of non-linear structures. This endeavor is particularly relevant and promising for computational linguistics: while many tools already exist to automatically detect “disfluencies” (e.g. Heeman & Allen 1999; Christodoulides et al. 2014; Moniz et al. 2015), insights from the present quantitative-qualitative study could refine their classification by specifying not only the type of fluenceme but also its diagnosis on the (dis)fluency scale, paving the way for the automatic detection of “fluencies” as well.

Two elements of methodological discussion should be addressed before turning to the general conclusion of the thesis (Chapter 8), namely the qualitative nature of the coding procedure, and the interdependence of the variables. The coding scheme used throughout this chapter might strike as particularly more qualitative than the corpus-based approach adopted so far: while the functional annotation of DMs is already challenging and arguably subjective, the identification of repair types relies heavily on deep interpretation of the speaker’s motives and intentions, as well as some normative evaluation of the degree of “error” involved in a repaired utterance (cf. also the approach to DM co-occurrence undertaken in Section 5.4). This strong involvement of the researcher also concerns to a smaller extent more formal variables in this analysis such as the occasion for repair (where did the problem come from?) or the interruption point (is the on-going unit complete or not?). I would like to suggest that the main difference between the method described in Chapter 4 and the procedure detailed in Section 7.2 of this chapter corresponds to the difference between corpus annotation and conversation 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 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 conversation analysis (as developed by Schegloff et al. 1977), 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, as explained by Turner (1971):

As a solution to the vexed problem of the relation between the shared cultural knowledge (members’ knowledge) that the sociologist possesses and the analytic apparatus that it is his responsibility to produce, I propose the following: the sociologist inevitably trades on his members’ knowledge in recognizing the activities that participants to interaction are engaged in; for example, it is by virtue of my status as a competent member that I can recurrently locate in my transcripts instances of “the same” activity. (1971: 177)

In the context of coding repair types, this reference to world knowledge and to one’s experience as a member of a linguistic community is heavily relied upon by the analyst. 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, sometimes with very few occurrences to interpret, so much so that this analysis, 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.

The second element of this methodological discussion is the potential circularity or inter-dependence of the variables involved in the analysis. One goal of this study was to avoid the pitfalls of previous approaches to reformulation where definitions of reformulation types are entirely based on the type of marker involved (Rossari 1994) or the functions of the marker (Ciabarra 2013). To do so, I kept as distinct levels of analysis the formal and functional annotation of fluencemes on the one hand, and the qualitative coding of repair type on the other. This endeavour was already present in the previous chapters where formal variables (identification of fluencemes, syntactic characteristics of DMs) did not overlap with functional variables. In Crible & Degand (under review), we argue that such independence of variables is beneficial even within a single set of (functional) variables such as domains and functions, since it allows for more reliability and more explanatory power, and may uncover unexpected associations. In the present chapter, however, I have noted on several occasions the circularity of some definitions, such as the potential inter-dependence of repair type and DM functions. In the perspective of building a formal scale of fluency, where cognitive-functional interpretations are related to objective features of the linguistic material, this potential problem of circularity is paramount to bear in mind when assessing the validity and explanatory power of the proposed model. Although convincing patterns of formal correlates have been identified for the majority of problematic areas, I do not have the data or tools to objectively measure the degree of circularity in my analysis, and it is only fair to assume that it is certainly not null.

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 conversation-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, and managed to tie the different goals and parts of this thesis together in a coherent and convincing way in spite of its limitations.

Chapter 8: Conclusion

8.1 Summary of the main findings

The present usage-based contrastive study of discourse markers and (dis)fluency across registers pursued a three-fold objective: (i) 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); (ii) to uncover fluent and disfluent uses of DMs based on the converging evidence of their linguistic features, their contextual variation and the types of fluencemes with which they tend to cluster (Chapter 6); (iii) to suggest a formal scale of (dis)fluency through which the format (*how*) of repair sequences can help disambiguate their degree of fluency (*why*), thus coming to terms with the ambivalence of fluencemes (Chapter 7). This first section of the concluding chapter summarizes the main results of this thesis.

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.

As for the effect of language preferences, 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 (see below) 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’.

All in all, the following schemas 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 large, bottom-up 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 multifunctionality 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, connectivity 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.

The quantitative-qualitative analysis of DM co-occurrence offered to bridge the gap between top-down categories and bottom-up annotation and revealed that the phenomenon of discourse-level co-occurrence, which concerns one DM in five, seems to target complementarity rather than redundancy of meanings, a result which points to the role of underspecification, as in the frequently co-occurring DMs *and* or *quoi* ‘right’.

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 large 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. Concretely, 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 suggest 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 7. However, these schemas 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, where the linearity of the linguistic product competes with the non-linearity of the production and reception processes. Three major patterns of different degrees of (dis)fluency were identified and mapped with specific formats of the repair sequence:

- structural repairs, usually interrupting local constituents and resolved by start-overs (low fluency);
- lexical-search repairs, usually well-integrated in the local structure (intermediary fluency);
- resonance repairs, systematically marked by a deep integration in higher-order syntactic units and the repetition of large stretches of previously uttered material (high fluency).

The analyses in Chapter 7 also revealed the positive role of lengthy repetitions (either long-distance or long-retracing repetitions), which constitute a conservative strategy for speakers to anchor new information in recently stored material, thus relieving the cognitive load of production and interpretation (cf. Auer 2005 on the strategic use of delayed self-repair). Mapping Levelt's model with the fluencemes annotated in *DisFrEn* allowed us to confirm the strong ambivalence of modified repetitions, as opposed to more clearly disfluent fluencemes such as false-starts, which do not provide any interpretative clue to the hearer but constitute mere disruptions in the linear flow of words. 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 three 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 fact that spoken utterances are not meant to be linear, that is, a monotonous flow of words where every position is equally salient and informative, but rather a dynamic, time-sensitive and co-built object. 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 thesis 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, linear) and rhythm (i.e. temporal, non-linear) in spoken discourse and (dis)fluency in fact reconciles time with space, in accordance with the “spacetime continuum” metaphor with which this thesis started.

8.2 General discussion

The results summarized above raise a number of theoretical and methodological issues. The starting assumption of the present approach to (dis)fluency – and of the collaborative project to which this thesis strove to contribute – 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, 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 6 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 schema, 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 thesis.

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 large 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, as shown by the repair sequences investigated in Chapter 7, where DMs occurred in low frequencies scattered across a variety of repair types. Some DM features were also found to be only remotely relevant to analyses of fluency, such as the objective-subjective divide in discourse relations or the role of position in the cognitive-functional scale of fluency, which further indicates that the behavior of DMs cannot be fully related to considerations of fluency and disfluency. Instead, distinctions such as objective-subjective or initial-final are more meaningful in investigations of register or crosslinguistic variation, as well as other dimensions of DM use such as emotionality (Romano & Cuenca 2013), persuasion (Hosman & Siltanen 2011) or intersubjectivity (House 2013). In this sense, the present study of DMs is also restricted, although a complete portrait of DMs in all their potential ramifications is probably infeasible, as attested by the fragmentation of the state of the art.

As a last point of discussion, I would like to come back to the notion of (non-)linearity and its status in the present research. As developed in Chapter 2, linearity, coupled with temporality, offers an accurate metaphorical representation of the production and interpretation of discourse which pinpoints the specificity of speech as opposed to writing. This notion was

therefore used to describe the nature of spoken language as well as to define the present approach to (dis)fluent devices. A non-linear definition of fluency motivates the inclusion of ambivalent, multifunctional markers beyond the usual restriction to “disfluencies” or unintentional accidents of language performance found in previous works. I hope that this study has shown that discourse markers and fluencemes in general constitute “tricks” to manage the linearity and temporality of the speech channel through forward- and backward-looking operations such as recycling a previous utterance, relating two utterances or announcing an upcoming change or new discourse boundary, thus restituting some spatiality to cope with the time pressure of production and interpretation of coherent, fluent discourse.

8.3 Implications and research avenues

Although this study of native (dis)fluency is more fundamental 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 (Chapter 3), 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 what was already investigated in the present work.

The functional taxonomy specifically designed for *DisFrEn* has already been applied to other data and languages such as spoken Slovene or Belgian French Sign Language (cf. the references in Section 4.2.1). 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 should make use of the 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. Ciabbarri 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 large 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 specificities of learner language probably forbid 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 could 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 socioeconomic 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 perceptive light on the corpus-based patterns identified.

Lastly, I would like to encourage a more theoretical line of research investigating the compatibility of the Construction Grammar framework (Fillmore & Kay 1993; Goldberg 1995; Croft 2001) to deal with patterns or schemas of discourse markers and fluencemes (see Langacker (2005) on the comparison between schemas and constructions). In this theory, discourse-level constructions, although acknowledged in principle, have only recently started to draw the attention of linguists (Fischer 2010; Gras 2011; Fischer & Alm 2013; Aijmer 2016) and might need to be further conceptualized before they can be used as a reference framework for large-scale analyses. Nevertheless, I believe that a stronger theoretical background, placing patterns of discourse markers and fluencemes at the same level of cognitive and processing reality as more central (i.e. grammatical) units of language, would pave the way for a more systematic cognitive-pragmatic approach to DMs overcoming the limitations of the present proposal and opening up to a wealth of theoretical, empirical and applied directions of research.

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

- Aijmer, K. 1988. "Now may we have a word on this": The use of *now* as a discourse particle". In M.K. Ossi & I.M. Rissanen (eds), *Corpus Linguistics, Hard and Soft: Proceedings of the Eighth International Conference on English Language Research on Computerized Corpora*, Amsterdam, Rodopi: 15-34.
- Aijmer, K. 2002. *English Discourse Particles: Evidence from a Corpus*. Philadelphia: John Benjamins.
- Aijmer, K. 2016. Pragmatic markers as constructions. The case of *anyway*. In G. Kaltenböck, E. Keizer & A. Lohmann (eds), *Outside the Clause: Form and Function of Extra-Clausal Constituents*, Amsterdam, John Benjamins: 29-58.
- Aijmer, K., Foolen, A. & Simon-Vandenberg, A.-M. 2006. Pragmatic markers in translation: A methodological proposal. In K. Fischer (Ed.), *Approaches to Discourse Particles*, Amsterdam, Elsevier: 101-114.
- Aijmer, K. & Simon-Vandenberg, A.-M. 2011. Pragmatic markers. In J. Zienkowski, J.-O. Östman & J. Verschueren (eds), *Discursive Pragmatics* [Handbook of Pragmatics Highlights 8], Amsterdam, John Benjamins: 223-247.
- Altenberg, B. 2006. The function of adverbial connectors in second initial position in English and Swedish. In K. Aijmer & Simon-Vandenberg, A.-M. (eds), *Pragmatic Markers in Contrast*, Oxford, Elsevier: 11-37.
- Andersen, G. 1997. They like wanna see like how we talk and all that. The use of *like* as a discourse marker in London teenage speech. In M. Ljung (Ed.), *Corpus-based Studies in English*, Amsterdam, Rodopi: 37-48.
- Andersen, G. 2000. The role of the pragmatic marker like in utterance interpretation. In G. Andersen & T. Fretheim (eds), *Pragmatic Markers and Propositional Attitude*, Amsterdam, John Benjamins: 17-38.
- Andersen, G. 2001. *Pragmatic Markers and Sociolinguistic Variation. A Relevance-Theoretic Approach to the Language of Adolescents*. Amsterdam: John Benjamins.
- Anscombe, J.-C. & Ducrot, O. 1983. *L'Argumentation dans la Langue*. Liège-Bruxelles: Mardaga.
- Anzai, Y. 2009. Deux fonctionnements du marqueur français "tu vois" dans les dialogues spontanés: Relation entre les faits intonatifs et la structure morphosyntaxique. In *Actes d'IDP 09*: 63-76.
- Arnold, J., Fagnano, M. & Tanenhaus, M. 2003. Disfluencies signal thee, um, new information. *Journal of Psycholinguistic Research* 32(1): 25-36.
- Arnold, J. & Tanenhaus, M. 2011. Disfluency effects in comprehension: How new information can become accessible. In E. Gibson & N. Perlmutter (eds), *The Processing and Acquisition of Reference*, Cambridge, MA, MIT Press: 197-217.
- Arppe, A. & Järvi, J. 2007. Every method counts: Combining corpus-based and experimental evidence in the study of synonymy. *Corpus Linguistics and Linguistic Theory* 3(2): 131-159.
- Asher, N. & Lascarides, A. 2003. *Logics of Conversation*. Cambridge: Cambridge University Press.
- Auchlin, A. 1981. Mais heu, pis bon, ben alors voilà, quoi! Marqueurs de structuration de la conversation et complétude. *Cahiers de Linguistique Française* 2: 141-160.

- Auer, P. 1996. The pre-front field in spoken German and its relevance as a grammatical position. In *Pragmatics* 6(3): 223-259.
- Auer, P. 2005. Delayed self-repairs as a structuring device for complex turns in conversation. In A. Hakulinen & M. Selting (eds), *Syntax and Lexis in Conversation: Studies on the Use of Linguistic Resources in Talk-in-interaction*, Amsterdam, John Benjamins: 75-102.
- Auer, P. 2009. On-line syntax: Thoughts on the temporality of spoken language. *Language Sciences* 31(1): 1-13.
- Auer, P. 2015. The temporality of language in interaction. Projection and latency. In A. Deppermann & S. Günthner (eds), *Temporality in Interaction*, Amsterdam, John Benjamins: 27-56.
- Auer, P. & Pfänder, S. 2007. Multiple retractions in spoken French and spoken German. A contrastive study in oral performance styles. *Cahiers de Praxématique* 48: 57-84.
- Avanzi, M. 2009. La prosodie des verbes parenthétiques en français parlé. *Linx* 61: 131-144.
- Avanzi, M., Simon, A.-C., Goldman, J.-P. & Auchlin, A. 2010. C-PROM. Un corpus de français parlé annoté pour l'étude des proéminences. In *Actes des 23èmes journées d'étude sur la parole* (May 25-28, Mons, Belgium).
- Bachy, S., Dister, A., Francard, M., Geron, G., Giroul, V., Hambye, P., Simon, A.-C. & Wilmet, R. 2004. Conventions de transcription régissant les corpus de la banque de données VALIBEL. https://www.uclouvain.be/cps/ucl/doc/valibel/documents/conventions_valibel_2004.PDF.
- Balthasar, L. & Bert, M. 2005. La base de données "Corpus de langues parlées en interaction" (CLAPI): Genèse, état des lieux et perspectives. In M. Savelli (Ed.), *Corpus Oraux et Diversité des Approches*, *Lidil* 31.
- Barr, D. & Seyfeddinipur, M. 2010. The role of fillers in listener attributions for speaker disfluency. *Language and Cognitive Processes* 25(4): 441-455.
- Barth, D. & Kapatsinski, V. in press. Evaluating logistic mixed-effects models of corpus data. In D. Speelman & D. Geeraerts (eds), *Mixed Models and Modern Multivariate Methods in Linguistics*, Berlin, Springer.
- Bates, E. & MacWhinney, B. 1989. Functionalism and the Competition Model. In B. MacWhinney & E. Bates (eds), *The Crosslinguistic Study of Sentence Processing*, Cambridge, Cambridge University Press: 3-73.
- Bates, D., Maechler, M., Bolker, B. & Walker, S. 2014. lme4: Linear mixed-effects models using Eigen and S4. R package version 1.0-6. <http://CRAN.R-project.org/package=lme4>.
- Bazzanella, C., Bosco, C., Garcea, A., Gili Fivela, B., Miecznikowski, J. & Tini Brunozzi, F. 2007. Italian *allora*, French *alors*: Functions, convergences and divergences. *Catalan Journal of Linguistics* 6: 9-30.
- Beach, W. 1993. Transitional regularities for "casual" *Okay* usages. *Journal of Pragmatics* 19: 325-352.
- Beeching, K. 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? *Langue française* 154(2): 78-93.
- Beeching, K. in press. Just a suggestion: *just/e* in French and English. In C. Fedriani & A. Sanso (eds), *Discourse Markers, Pragmatics Markers and Modal Particles: New Perspectives*, Amsterdam, John Benjamins.

- Beeching, K. & Detges, U. (eds). 2014. *The Role of the Left and Right Periphery in Semantic Change: Crosslinguistic Investigations of Language and Language Change*. Brill: Leiden.
- Beliao, J. & Lacheret, A. 2013. Disfluency and discursive markers: When prosody and syntax plan discourse. In R. Eklund (Ed.), *Proceedings of Disfluency in Spontaneous Speech (DiSS)*: 5-8.
- Benamara, F. & Taboada, M. 2015. Mapping different rhetorical relation annotations: A proposal. In *Proceedings of the 4th Joint Conference on Lexical and Computational Semantics (*SEM), collocated with NAACL*, Denver, USA.
- Bertrand, R., Blâche, P., Espesser, R., Ferré, G., Meunier, C., Priego-Valverde, B. & Rauzy, S. 2008. Le CID — Corpus of Interactional Data — Annotation et Exploitation Multimodale de Parole Conversationnelle. *Traitement Automatique des Langues*, 49(3).
- Bertrand, R. & Chanut, C. 2005. Fonctions pragmatiques et prosodie de *enfin* en français spontané. *Revue de Sémantique et Pragmatique* 17: 41-68.
- Besser, J. & Alexandersson, J. 2007. A comprehensive disfluency model for multi-party interaction. In S. Keizer, H. Bunt & T. Paek (eds), *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*: 182-189.
- Biber, D., Conrad, S. & Reppen, R. 1999. *Corpus Linguistics: Investigating Language Structure and Use*. Philadelphia: John Benjamins.
- Blakemore, D. 1987. *Semantic Constraints on Relevance*. Blackwell, Oxford.
- Blakemore, D. 1993. The relevance of reformulations. *Language and Literature* 2(2):101-120.
- Blakemore, D. 2002. *Relevance and Linguistic Meaning. The Semantics and Pragmatics of Discourse Markers*. Cambridge: Cambridge University Press.
- Blakemore, D. & Carston, R. 2005. The pragmatics of sentential coordination with *and*. *Lingua* 115: 569-589.
- Blanche-Benveniste, C. 2003. Le recouvrement de la syntaxe et de la macro-syntaxe. In A. Scarano (Ed.), *Macro-syntaxe et Pragmatique. L'Analyse Linguistique de l'Oral*, Rome, Bulzoni: 53-75.
- Blanche-Benveniste, C., Bilger, M., Rouget, C., Van Den Eynde, K. & Mertens, P. 1990. *Le Français Parlé. Etudes Grammaticales*. Paris: CNRS.
- Blankenship, K. & Holtgraves, T. 2005. The role of different markers of linguistic powerlessness in persuasion. *Journal of Language and Social Psychology* 24(1): 3-24.
- BNC Consortium. 2007. *The British National Corpus*, version 3 (BNC XML Edition). Distributed by Oxford University Computing Services. <http://www.natcorp.ox.ac.uk/>.
- Bolden, B. 2008. "So what's up?": Using the discourse marker *so* to launch conversational business. *Research on Language and Social Interaction* 41(3): 302-337.
- Bolden, G. 2009. Implementing incipient actions: The discourse marker "so" in English conversation. *Journal of Pragmatics* 41: 974-998.
- Bolly, C. 2009. Constructionnalisation et structure informationnelle. Quand la grammaticalisation ne suffit pas pour expliquer *tu vois*. *Linx* 61: 103-130.
- Bolly, C. 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, C. & Crible, L. 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, C. & Crible, L. forthcoming. From context to functions and back again: Towards a multimodal taxonomy of pragmatic markers.
- Bolly, C., Crible, L., Degand, L. & Uygur-Distexhe, D. 2015. MDMA. Identification et annotation des marqueurs discursifs “potentiels” en contexte. *Discours* 15.
- Bolly, C., Crible, L., Degand, L. & Uygur-Distexhe, D. in press. Towards a Model for Discourse Marker Annotation in spoken French: From potential to feature-based discourse markers. In C. Fedriani & A. Sanso (eds), *Discourse Markers, Pragmatics Markers and Modal Particles: New Perspectives*, Amsterdam, John Benjamins.
- Bolly, C. & Degand, L. 2009. Quelle(s) fonction(s) pour “donc” en français oral? Du connecteur conséquentiel au marqueur de structuration du discours. *Linguisticae Investigationes* 32(1): 1-32.
- Bolly, C. & Thomas, A. 2015. Facing Nadine’s speech. Multimodal annotation of emotion in later life. In K. Jokinen & M. Vels (eds), *Proceedings of the 2nd European and the 5th Nordic Symposium on Multimodal Communication*, August 6-8, Tartu, Estonia, Linköping, Linköping Electronic Conference Proceedings 110: 23-32.
- Bortfeld, H., Leon, S., Bloom, J., Schober, M., & Brennan, S. 2001. Disfluency rates in conversation: Effects of age, relationship, topic, role and gender. *Language and Speech* 44: 123-147.
- Bosker, H., Quené, H., Sanders, T. & 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.
- Boula de Mareüil, P., Adda, G., Adda-Decker, M., Barras, C., Habert, B. & Paroubek, P. 2013. Une étude quantitative des marqueurs discursifs, disfluences et chevauchements de parole dans des interviews politiques. *TIPA Travaux Interdisciplinaires sur la Parole et le Langage* 29.
- Boula De Mareüil, P., Habert, B., Bénard, F., Adda-Decker, M., Barras, C., Adda, G. & Paroubek, P. 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, S. 1991. Word interruption in self-repairing. *Journal of Psycholinguistic Research* 20(2): 123-138.
- Brennan, S.E. & Schober, M.F. 2001. How listeners compensate for disfluencies in spontaneous speech. *Journal of Memory and Language* 44: 274-296.
- Brinton, L. 1996. *Pragmatic Markers in English. Grammaticalization and Discourse Functions*. New York: Mouton de Gruyter.
- Briz, A. & Pons Bordería, S. 2010. Unidades, marcadores discursivos y posición. In O. Loureda & E. Acín (eds), *Los Estudios sobre Marcadores del Discurso*, Madrid, Acro/Libros: 523-557.
- Briz, A. & Val.Es.Co group. 2003. Un sistema de unidades para el estudio del lenguaje coloquial. *Oralia* 6: 7-61.
- Broen, P. & Siegel, G. 1972. Variations in normal speech disfluencies. *Language and Speech* 15: 219-231.

- Brognaux, S., Roekhaut, S., Drugman, T., & Beaufort, R. 2012. Train&Align: A new online tool for automatic phonetic alignment. In *Proceedings of the Spoken Language Technology Workshop (SLT)*: 416-421.
- Brown, P. & Levinson, S.C. 1987. *Politeness: Some Universals in Language Usage*. Cambridge: Cambridge University Press.
- Brumfit, C. 1984. *Communicative Methodology in Language Teaching*. Cambridge: Cambridge University Press.
- Buyse, L. 2012. So as a multifunctional discourse marker in native and learner speech. *Journal of Pragmatics* 44(13): 1764-1782.
- Buyse, L. 2014. "We went to the restroom or something". General extenders and stuff in the speech of Dutch learners of English. In J. Romero-Trillo (Ed.), *Yearbook of Corpus Linguistics and Pragmatics: New Empirical and Theoretical Paradigms*, Berlin, Springer: 213-237.
- Bybee, J. 1985. *Morphology: A Study on the Relation between Meaning and Form*. Amsterdam: John Benjamins.
- Bybee, J. 2006. From usage to grammar: The mind's response to repetition. *Language* 82(4): 711-733.
- Bygate, M. 2009. Teaching the spoken foreign language. In K. Knapp & B. Seidhofer (eds), *Handbook of Foreign Language Communication and Learning*, Berlin, Mouton de Gruyter: 401-438.
- Candéa, M. 2000. *Contribution à l'Etude des Pauses Silencieuses et des Phénomènes Dits "d'Hésitation" en Français Oral Spontané*. PhD thesis, Université Paris III.
- Canestrelli, A., Mak, W. & Sanders, T.J.M. 2013. Causal connectives in discourse processing: How differences in subjectivity are reflected in eye movements. *Language and Cognitive Processes* 28(9): 1394-1413.
- Carter, R. & McCarthy, M. 2006. *Cambridge Grammar of English*. Cambridge: Cambridge University Press.
- Castellà, J.M. 2004. *Oralitat i Escriptura. Dues Cares de la Complexitat del Llenguatge*. Barcelona: Publicacions de l'Abadia de Montserrat.
- Chafe, W. 1992. The importance of corpus linguistics to understanding the nature of language. In J. Svartvik (Ed.), *Directions in Corpus Linguistics*, Berlin, Mouton de Gruyter: 79-97.
- Chafe, W. 1994. *Discourse, Consciousness, and Time*. Chicago: University of Chicago Press.
- Chambers, F. 1997. What do we mean by fluency? *System* 25(4): 535-544.
- Chanet, C. 2001. 1700 occurrences de la particule *quoi* en français parlé contemporain: Approche de la "distribution" et des fonctions en discours. *Marges Linguistiques* 2: 56-80.
- Charolles, M. & Coltier, D. 1986. Le contrôle de la compréhension dans une activité rédactionnelle: Eléments pour l'analyse des reformulations paraphrastiques. *Pratiques* 49: 51-66.
- Cheschire, J. 2007. Discourse variation, grammaticalisation and stuff like that. *Journal of sociolinguistics* 11(2): 155-193.
- Christodoulides, G., Avanzi, M & Goldman, J.-P. 2014. DisMo: A morphosyntactic, disfluency and multi-word unit annotator. An evaluation on a corpus of French spontaneous and read speech. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC) 2014*, 26-31 May, Reykjavik, Iceland: 3902-3907.

- Ciabarri, F. 2013. Italian reformulation markers: A study on spoken and written language. In C. Bolly & L. Degand (eds), *Across the Line of Speech and Writing Variation* [Corpora and Language in Use - Proceedings 2], Louvain-la-Neuve, Presses universitaires de Louvain: 113-128.
- Clark, H. 2002. Speaking in time. *Speech Communication* 36: 5-13.
- Clark, H. & Fox Tree, J. 2002. Using *uh* and *um* in spontaneous speaking. *Cognition* 84: 73-111.
- Clark, H. & Wasow, T. 1998. Repeating words in spontaneous speech. *Cognitive Psychology* 37: 201-242.
- Colletta, J.-M., Kunene, R.N., Venouil, A., Kaufmann, V. & Simon, J.-P. 2009. Multi-track annotation of child language and gestures. In M. Kipp, J.-C. Martin, P. Paggio & D. Heylen (eds.), *Multimodal corpora (Lecture Notes in Computer Science 5509)*, Berlin, Springer: 54-72.
- Collister, L. 2013. *Multimodality as a Sociolinguistic Resource*. PhD thesis, University of Pittsburgh.
- Condon, S. 1986. The discourse functions of OK. *Semiotica* 60: 73-101.
- Condon, S. 2001. Discourse *ok* revisited: Default organization in verbal interaction. *Journal of Pragmatics* 33: 491-513.
- Connor, U.M. & Moreno, A.I. 2005. Tertium Comparationis: A vital component in contrastive research methodology. In P. Bruthiaux, D. Atkinson, W.G. Eggington, W. Grabe & V. Ramanathan (eds), *Directions in Applied Linguistics: Essays in Honor of Robert B. Kaplan*, England, Multilingual Matters: 153-164.
- Corley, M. 2010. Making predictions from speech with repairs: Evidence from eye movements. *Language and Cognitive Processes* 25(5): 706-727.
- Corley, M., MacGregor, L. & Donaldson, D. 2007. It's the way that you, er, say it: Hesitations in speech affect language comprehension. *Cognition* 105: 658-668.
- Corley, M. & Stewart, O. 2008. Hesitation disfluencies in spontaneous speech: The meaning of *um*. *Language and Linguistics Compass* 4: 589-602.
- Crible, L. 2014. Identifying and describing discourse markers in spoken corpora. Annotation protocol v.8. Technical report, Université catholique de Louvain.
- Crible, L. 2015. Grammaticalisation du marqueur discursif complexe ou sinon dans le corpus de SMS belge: Spécificités sémantiques, graphiques et diatopiques. *Le Discours et la Langue* 7(1):181-200.
- Crible, L. in press. Towards an operational category of discourse markers: A definition and its model. In C. Fedriani & A. Sanso (eds), *Discourse Markers, Pragmatics Markers and Modal Particles: New Perspectives*, Amsterdam, John Benjamins.
- Crible, L. under review. Discourse markers and (dis)fluency in English and French: Variation and combination in the *DisFrEn* corpus.
- Crible, L. & Cuenca, M.J. under review. Discourse markers in speech: Distinctive features and corpus annotation.
- Crible, L. & Degand, L. under review. Reliability vs. granularity in discourse annotation: What is the trade-off?
- Crible, L., Degand, L. & Gilquin, G. 2017a. The clustering of discourse markers and filled pauses: A corpus-based French-English study of (dis)fluency. *Languages in Contrast* 17(1): 69-95.

- Crible, L., Dumont, A., Grosman, I. & Notarrigo, I. 2016. Annotation manual of fluency and disfluency markers in multilingual, multimodal, native and learner corpora. Version 2.0. Technical report, Université catholique de Louvain and Université de Namur.
- Crible, L., Dumont, A., Grosman, I. & Notarrigo, I. 2017b. (Dis)fluency across spoken and signed languages: Applications of an interoperable annotation scheme. Paper presented at the *International Conference on Fluency & Disfluency Across Languages and Language Varieties*, February 15-17, Louvain-la-Neuve, Belgium.
- Crible, L. & Zufferey, S. 2015. Using a unified taxonomy to annotate discourse markers in speech and writing. In H. Bunt (Ed.), *Proceedings of the 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (isa-11), IWCS 2015 Workshop*: 14-22.
- Croft, W. 1991. *Syntactic Categories and Grammatical Relations: The Cognitive Organization of Information*. Chicago: University of Chicago Press.
- Croft, W. 2001. *Radical Construction Grammar. Syntactic Theory in Typological Perspective*. Oxford: Oxford University Press.
- Crystal, D. 1987. *The Cambridge Encyclopedia of Language*. Cambridge: Cambridge University Press.
- Crystal, D. 1988. Another look at, well, you know... *English Today* 4(1): 47-49.
- Cuenca, M.J. 2003. Two ways to reformulate: A contrastive analysis of reformulation markers. *Journal of Pragmatics* 35: 1069-1093.
- Cuenca, M.J. 2013. The fuzzy boundaries between discourse marking and modal marking. In L. Degand, B. Cornillie & P. Pietrandrea (eds), *Discourse Markers and Modal Particles. Categorization and Description*, Amsterdam, John Benjamins: 191-216.
- Cuenca, M.J. & Bach, C. 2007. Contrasting the form and use of reformulation markers. *Discourse Studies* 9(2): 149-175.
- Cuenca, M.J. & Marín, M.J. 2009. Co-occurrence of discourse markers in Catalan and Spanish oral narrative. *Journal of Pragmatics* 41: 899-914.
- Cutting, J. 2008. *Pragmatics and Discourse, 2nd edition*. New York: Routledge.
- Danks, J. & End, L. 1987. Processing strategies for reading and listening. In R. Horowitz & J. Samuels (eds), *Comprehending Oral and Written Language*, San Diego, Academic Press: 271-294.
- Danlos, L., Antolinos-Basso, D., Braud, C. & Roze, C. 2012. Vers le FDTB: French Discourse Tree Bank. In *Proceedings of TALN 2012: 19ème conférence sur le Traitement Automatique des Langues Naturelles* 2: 471-478.
- Davies, A. 2003. *The Native Speaker: Myth and Reality*. Clevedon: Multilingual Matters.
- De Cock, S. 2000. Repetitive phrasal chunkiness and advanced EFL speech and writing. In C. Mair & M. Hundt (eds), *Corpus Linguistics and Linguistic Theory. Papers from the Twentieth International Conference on English Language Research on Computerized Corpora (ICAME 20)*, Amsterdam, Rodopi: 51-68.
- De Gaulmyn, M.-M. 1987. Actes de reformulation et processus de reformulation. In P. Bange (Ed.) *L'Analyse des Interactions Verbales. La Dame de Caluire: Une Consultation*, Bern, Peter Lang: 83-98.

- De Jong, N., Steinel, M., Florijn, A., Schoonen, R. & Hulstijn, J. 2012. The effect of task complexity on functional adequacy, fluency and lexical diversity in speaking performances of native and non-native speakers. In A. Housen, F. Kuiken & I. Vedder (eds), *Dimensions of L2 Performance and Proficiency. Complexity, Accuracy and Fluency in SLA*, Amsterdam, John Benjamins: 121-142.
- De Mönnink, I.M. 1997. Using corpus and experimental data: A multimethod approach. In M. Ljung (Ed.), *Corpus-Based Studies in English*, Amsterdam, Rodopi: 227-244.
- de Saussure, L. & Sthioul, B. 2002. Interprétations cumulative et distributive du connecteur *et*: Temps, argumentation, séquençement. *Cahiers de Linguistique Française* 24: 293-314.
- Defour, T., D'Hondt, U., Vandenberghe, A.-M. & Willems, D. 2010. *In fact, en fait, de fait, au fait*: A contrastive study of the synchronic correspondences and diachronic development of English and French cognates. *Neuphilologische Mitteilungen* 111(4): 433-463.
- Degand, L. 2014. "So very fast, very fast then" Discourse markers at left and right periphery in spoken French. In K. Beeching & U. Detges (eds), *The Role of the Left and Right Periphery in Semantic Change: Crosslinguistic Investigations of Language and Language Change*, Leiden, Brill: 151-178.
- Degand, L., Cornillie, B. & Pietrandrea, P. (eds). 2013. *Discourse Markers and Modal Particles. Categorization and Description*. Amsterdam: John Benjamins.
- Degand, L. & Fagard, B. 2011. *Alors* between discourse and grammar. The role of syntactic position. *Functions of Language* 18(1): 29-56.
- Degand, L. & Gilquin, G. 2013. The clustering of "fluencemes" in French and English. Paper presented at the 7th International Contrastive Linguistics Conference (ICLC 7) – 3rd Conference on Using Corpora in Contrastive and Translation Studies (UCCTS 3), July 11-13, Ghent, Belgium.
- Degand, L., Martin, L.J. & Simon, A.-C. 2014. Unités discursives de base et leur périphérie gauche dans LOCAS-F, un corpus oral multigenres annoté. In *Proceedings of CMLF 2014 – 4ème Congrès Mondial de Linguistique Française 2014, Berlin, Germany: EDP Sciences*.
- Degand, L. & Simon, A.-C. 2009. On identifying basic discourse units in speech: Theoretical and empirical issues. *Discours* 4.
- Degand, L. & Simon-Vandenberghe, A.-M. 2011. Grammaticalization and (inter)subjectification of discourse markers. *Linguistics* 49: 287-294.
- Dehé, N. & Wichmann, A. 2010. The multifunctionality of epistemic parentheticals in discourse: Prosodic cues to the semantic-pragmatic boundary. *Functions of Language* 17(1): 1-28.
- Demirşahin, I. & Zeyrek, D. 2014. Annotating discourse connectives in spoken Turkish. In *Proceedings of LAW VIII – The 8th Linguistic Annotation Workshop*: 105-109.
- Denke, A. 2009. *Nativelike Performance. Pragmatic Markers, Repair and Repetition in Native and Non-native English Speech*. Saarbrücken: Verlag Dr. Müller.
- Deppermann, A. & Günthner, S. 2015. *Temporality in Interaction*. Amsterdam: John Benjamins.
- Diessel, H. & Hilpert, M. 2016. Frequency effects in grammar. In Mark Aronoff (Ed.), *Oxford Research Encyclopedia of Linguistics*. New York: Oxford University Press.
- Diewald, G. 2006. Discourse particles and modal particles as grammatical elements. In K. Fischer (Ed.), *Approaches to Discourse Particles*, Amsterdam, Elsevier: 403-426.

- Diewald, G. 2013. "Same same but different" – Modal particles, discourse markers and the art (and purpose) of categorization. In L. Degand, B. Cornillie & P. Pietrandrea (eds), *Discourse Markers and Modal Particles. Categorization and description*, Amsterdam, John Benjamins: 19-46.
- Dister, A. 2007. *De la Transcription à l'Etiquetage Morphosyntaxique – Le Cas de la Banque de Données Textuelles Orales VALIBEL*. PhD thesis, Université catholique de Louvain.
- Dister, A., Francard, M., Hambye, P., & Simon, A.-C. 2009. Du corpus à la banque de données. Du son, des textes et des métadonnées. L'évolution de la banque de données textuelles orales VALIBEL (1989-2009). *Cahiers de Linguistique* 33(2), 113-129.
- Divjak, D. 2015. Four challenges for usage-based linguistics. In J. Daems, E. Zenner, K. Heylen, D. Speelman & H. Cuyckens (eds), *Change of Paradigms: New Paradoxes. Recontextualizing Language and Linguistics*, Berlin, Walter de Gruyter: 297-309.
- Dobrovoljc, K. 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.
- Dorgeloh, H. 2004. Conjunction in sentence and discourse: Sentence-initial *and* and discourse structure. *Journal of Pragmatics* 36: 1761-1779.
- Dostie, G. 2004. *Pragmaticalisation et Marqueurs Discursifs. Analyse Sémantique et Traitement Lexicographique*. Bruxelles: De Boeck.
- Dostie, G. 2009. Discourse markers and regional variation in French. A lexico-semantic approach. In K. Beeching, N. Armstrong & F. Gadet (eds), *Sociolinguistic Variation in Contemporary French*, Amsterdam, John Benjamins: 201-214.
- Dostie, G. 2013. Les associations de marqueurs discursifs - De la cooccurrence libre à la collocation. *Linguistik Online* 62(5).
- Du Bois, J. 1987. The discourse basis of ergativity. *Language* 64: 805-855.
- Du Bois, J. 2014. Towards a dialogic syntax. *Cognitive Linguistics* 25(3): 359-410.
- Du Bois, J.W., Chafe, W.L., Meyer, C., Thompson, S.A., Englebreton, R. & Martey, N. 2000-2005. *Santa Barbara Corpus of Spoken American English, Parts 1-4*. Philadelphia: Linguistic Data Consortium.
- Duez, D. 1991. *La Pause dans la Parole de l'Homme Politique*. Paris: Editions du CNRS.
- Dupont, M. 2015. Word order in English and French. The position of English and French adverbial connectors of contrast. *English Text Construction* 8(1): 88-124.
- Durand, J., Laks, B. & Lyche, C. 2002. La phonologie du français contemporain: Usages, variétés et structure. In C. Pusch & W. Raible (eds), *Romanistische Korpuslinguistik - Korpora und gesprochene Sprache/Romance Corpus Linguistics - Corpora and Spoken Language*, Tübingen, Gunter Narr Verlag: 93-106.
- Durand, J., Laks, B. & Lyche, C. 2009. Le projet PFC: Une source de données primaires structurées. In J. Durand, B. Laks & C. Lyche (eds), *Phonologie, Variation et Accents du Français*, Paris, Hermès: 19-61.

- Dutrey, C., Rosset, S., Adda-Decker, M., Clavel, C. & Vasilescu, I. 2014. Disfluences dans la parole spontanée conversationnelle: Détection automatique utilisant des indices lexicaux et acoustiques. In *Proceedings of the XXXe Journées d'Etude sur la Parole (JEP'14)*: 366-373.
- Ejzenberg, R. 2000. The juggling act of oral fluency: A psycho-sociolinguistic metaphor. In H. Riggenbach (Ed.), *Perspectives on Fluency*, Ann Arbor, The University of Michigan Press: 288-313.
- Eklund, R. 2004. *Disfluency in Swedish Human-human and Human-machine Travel Booking Dialogues*. PhD thesis, Linköping Studies in Science and Technology.
- Eklund, R. & Shriberg, E. 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*.
- Ellis, N. 2016. Frequency in language learning and language change. In H. Behrens & S. Pfänder (eds), *Experience Counts: Frequency Effects in Language*, Berlin, Mouton de Gruyter: 239-256.
- Erard, M. 2004. Just like, er, words, not, um, throwaways. *The New York Times*, 2 January 2004: A 13 & A 15.
- Erman, B. 1987. *Pragmatic Expressions in English: A study of You Know, You See and I Mean in Face-to-face Conversation*. Acta Universitatis Stockholmiensis, Stockholm Studies in English 69. Stockholm: Almqvist & Wiksell.
- Erman, B. 2001. Pragmatic markers revisited with a focus on *you know* in adult and adolescent talk. *Journal of Pragmatics* 33: 1337-1359.
- Esser, J. 1993. *English Linguistic Stylistics*. Tübingen: Niemeyer.
- Estellés Arguedas, M. & Pons Bordería, S. 2014. Absolute initial position. In S. Pons Bordería (Ed.), *Discourse Segmentation in Romance Languages*, Amsterdam, John Benjamins: 121-155.
- Fehrer, C. & Fry, C. 2007. Hesitation phenomena in the language production of bilingual speakers: The role of working memory. *Folia Linguistica* 41(1-2): 37-72.
- Fernández Polo, F.J. 1999. *Traducción y Retórica Contrastiva. A Propósito de la Traducción de Textos de Divulgación Científica del Inglés al Español*. Santiago de Compostela: Universidade de Santiago de Compostela. Anexo de Moenia 6.
- Fetzer, A. & Fischer, K. (eds). 2007. *Lexical Markers of Common Grounds* [Studies in Pragmatics 3]. Amsterdam: Elsevier.
- Fiksdal, S. 2000. Fluency as a function of time and rapport. In H. Riggenbach (Ed.), *Perspectives on Fluency*, Ann Arbor, The University of Michigan Press: 128-140.
- Fillmore, C. 1979. On fluency. In C. Fillmore, D. Kempler & W. Wang (eds), *Individual Differences in Language Ability and Language Behavior*, New York, Academic Press: 85-102.
- Fillmore, C. 2000. On fluency. In H. Riggenbach (Ed.), *Perspectives on Fluency*, Ann Arbor, The University of Michigan Press: 43-60.
- Fillmore, C. & Kay, P. 1993. *Construction Grammar*. Berkeley: University of California.
- Fischer, K. 1999. Repeats, reformulations, and emotional speech: Evidence for the design of human-computer speech interfaces. In H.-J. Bullinger & J. Ziegler (eds.), *Human-Computer Interaction: Ergonomics and User Interfaces, Volume 1 of the Proceedings of the 8th*

- International Conference on Human-Computer Interaction*, Lawrence Erlbaum Ass., London: 560-565.
- Fischer, K. 2000. *From Cognitive Semantics to Lexical Pragmatics: The Functional Polysemy of Discourse Particles*. Berlin: Walter de Gruyter.
- Fischer, K. (Ed.) 2006a. *Approaches to Discourse Particles*. [Studies in Pragmatics 1]. Amsterdam: Elsevier.
- Fischer, K. 2006b. Towards an understanding of the spectrum of approaches to discourse particles: Introduction to the volume. In Kerstin Fischer (Ed.), *Approaches to discourse particles* [Studies in Pragmatics 1], Amsterdam, Elsevier: 1-20.
- Fischer, K. 2010. Beyond the sentence: Constructions, frames and spoken interaction. *Constructions and Frames* 2: 1-28.
- Fischer, K. 2016. Definitions of discourse markers and their functions as discourse-relational devices. Paper presented at the *Discourse Relational Devices Conference (LPTS2016)*, *Linguistic & Psycholinguistic Approaches to Text Structuring*, January 24-26, Valencia, Spain.
- Fischer, K. & Alm, M. 2013. A radical construction grammar perspective on the modal particle-discourse particle distinction. In L. Degand, B. Cornillie & P. Pietrandrea (eds), *Discourse Markers and Modal Particles. Categorization and Description*, Amsterdam, John Benjamins: 47-88.
- Fleischman, S. & Yaguello, M. 2004. Discourse markers across languages. Evidence from English and French. In C.L. Moder & A. Martinovic-Zic (eds), *Discourse across Languages and Cultures*, Philadelphia, John Benjamins: 129-148.
- Flower, L. & Hayes, J. 1981. A cognitive process theory of writing. *College Composition and Communication* 32(4): 365-87.
- Fortescue, M. 2007. The non-linearity of speech production. In M. Hannay & G. Steen (eds), *Structural-Functional Studies in English Grammar*, Amsterdam, John Benjamins: 337-351.
- Foster, P., Tonkyn, A. & Wigglesworth, G. 2000. Measuring spoken language: A unit for all reasons. *Applied Linguistics* 21(3): 354-375.
- Fox, B., Hayashi, M. & Jasperson, R. 1996. Resources and repair: A cross-linguistic study of syntax and repair. In E. Ochs, E. Schegloff & S. Thompson (eds), *Interaction and Grammar. Studies in Interactional Sociolinguistics* 13, Cambridge, Cambridge University Press: 185-237.
- Fox Tree, J.E. 1995. The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory and Language* 34(6): 709-738.
- Fox Tree, J.E. 2001. Listeners' uses of um and uh in speech comprehension. *Memory and Cognition* 29(2): 320-326.
- Fox Tree, J.E. 2014. Discourse markers in writing. *Discourse Studies* 17(1): 64-82.
- Fox Tree, J.E. & Schrock, J.C. 2002. Basic meanings of you know and I mean. *Journal of Pragmatics* 34: 727-747.
- Fraser, B. 1988. Types of English discourse markers. *Acta Linguistica Hungarica* 38(1-4): 19-33.
- Fraser, B. 1996. Pragmatic markers. *Pragmatics* 6(2): 167-190.

- Freed, B. 2000. Is fluency, like beauty, in the eyes (and ears) of the beholder? In H. Riggenbach (Ed.), *Perspectives on Fluency*, Ann Arbor, The University of Michigan Press: 243-265.
- Fuller, J. 2003. The influence of speaker roles on discourse marker use. *Journal of Pragmatics* 35: 23-45.
- Gabarró-López, S. under review. Marqueurs du discours en langue des signes de Belgique francophone (LSFB) et langue des signes catalane (LSC): Les “balise-listes” et les “palm-ups”. In O. Loureda, G. Álvarez Sellán & M. Rudka (eds), *Marcadores del Discurso y Lingüística Contrastiva en las Lenguas Románicas*, Madrid, Iberoamericana Vervuert.
- Gaines, P. 2011. The multifunctionality of discourse operator *okay*: Evidence from a police interview. *Journal of Pragmatics* 43: 3291-3315.
- Geeraerts, D. 1998. Where does prototypicality come from? In B. Rudzka-Ostyn (Ed.), *Topics in Cognitive Linguistics*, Amsterdam, John Benjamins: 207-229.
- Geluykens, R. 1994. *The Pragmatics of Discourse Anaphora in English. Evidence from Conversational Repair*. Berlin: Mouton de Gruyter.
- Georgakopoulou, A. & Goutsos, D. 1998. Conjunctions versus discourse markers in Greek: The interaction of frequency, position, and functions in context. *Linguistics* 36(5): 887-917.
- Gerstenberg, A. & Bolly, C. 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, G. 2006. The place of prototypicality in corpus linguistics. Causation in the hot seat. In S. Gries & A. Stefanowitsch (eds), *Corpora in Cognitive Linguistics: Corpus-based Approaches to Syntax and Lexis*, Berlin, Mouton de Gruyter: 159-191.
- Gilquin, G. 2008. What you think ain't what you get: Highly polysemous verbs in mind and language. In J.-R. Lapaire, G. Desagulier & J.-B. Guignard (eds), *Du Fait Grammatical au Fait Cognitif. From Gram to Mind: Grammar as Cognition. Volume 2*, Pessac, Presses Universitaires de Bordeaux: 235-255.
- Gilquin, G. 2016. Discourse markers in L2 English. From classroom to naturalistic input. In O. Timofeeva, A.-C. Gardner, A. Honkapohja & S. Chevalier (eds), *New Approaches to English Linguistics: Building Bridges*, Amsterdam, John Benjamins: 213-249.
- Gilquin, G. & De Cock, S. 2011. Errors and disfluencies in spoken corpora. Setting the scene. *International Journal of Corpus Linguistics* 16(2): 141-172.
- Gilquin, G. & Gries, S. 2009. Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory* 5(1), 1-26.
- Ginzburg, J., Fernandez, R. & Schlangen, D. 2014. Disfluencies as intra-utterance dialogue moves. *Semantics & Pragmatics* 7: 1-64.
- Givón, T. 1975. Focus and the scope of assertion: Some Bantu evidence. *Studies in African Linguistics* 6: 185-205.
- Glynn, D. 2010. Corpus-driven cognitive semantics. Introduction to the field. In D. Glynn & K. Fischer (eds), *Quantitative Methods in Cognitive Semantics: Corpus-Driven Approaches*, Berlin, De Gruyter Mouton: 1-41.

- Goldberg, A. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: Chicago University Press.
- Goldberg, A. 2006. *Constructions at Work: The Nature of Generalization in Language*. Oxford: Oxford University Press.
- Goldman, J.-P., Prsirr, T. & Auchlin, A. 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, F. 1968. *Psycholinguistics: Experiments in Spontaneous Speech*. London: Academic Press.
- Gómez González, M. 2014. Canonical tag questions in English, Spanish and Portuguese. A discourse-functional study. *Languages in Contrast* 14(1): 93-126.
- González, M. 2004. *Pragmatic Markers in Oral Narrative: The Case of English and Catalan*. Amsterdam: John Benjamins.
- González, M. 2005. Pragmatic markers and discourse coherence relations in English and Catalan oral narrative. *Discourse Studies* 77(1), 53-86.
- Götz, S. 2013. *Fluency in Native and Nonnative English Speech*. Amsterdam: John Benjamins.
- Granger, S. & Petch-Tyson, S. 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.
- Gras, P. 2011. *Gramática de Construcciones en Interacción. Propuesta de un Modelo y Aplicación al Análisis de Estructuras Independientes con Marcas de Subordinación en Español*. PhD thesis, Universitat de Barcelona.
- Greene, J. & Cappella, J. 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, H.P. 1957. Meaning. *The Philosophical Review* 66: 377-388.
- Gries, S. 2011. Corpus data in usage-based linguistics: What's the right degree of granularity for the analysis of argument structure constructions? In M. Brdar, S. Gries & M. Žic Fuchs (eds), *Cognitive Linguistics: Convergence and Expansion*, Amsterdam, John Benjamins: 237-256.
- Gries, S. & Stefanowitsch, A. (eds). 2006. *Corpora in Cognitive Linguistics: Corpus-based Approaches to Syntax and Lexis*. Berlin: Mouton de Gruyter.
- Grosjean, F. & Deschamps, A. 1975. Analyse contrastive des variables temporelles de l'anglais et du français: Vitesse de parole et variables composantes, phénomènes d'hésitation. *Phonetica* 31: 144-184.
- Grosman, I. 2016. How do French humorists manage their persona across situations? A corpus study on their prosodic variation. In L. Ruiz-Gurillo (Ed.), *Metapragmatics of Humor: Current Research Trends*, Amsterdam, John Benjamins: 147-175.
- Guillemin-Flescher, J. 1981. *Syntaxe Comparée du Français et de l'Anglais*. Paris: Ophrys.
- Gülich, E. & Kotschi, T. 1987. Les actes de reformulation dans la consultation *La dame de Caluire*. In P. Bange (Ed.), *L'Analyse des Interactions Verbales. La Dame de Caluire: Une Consultation*, Bern, Peter Lang: 15-81.

- Gülich, E. & Kotschi, T. 1995. Discourse production in oral communication. In U.M. Quasthoff (Ed.), *Aspects of Oral Communication*, Berlin, Walter de Gruyter: 30-66.
- Halliday, M.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, M.A.K. 1987. Spoken and written modes of meaning. In R. Horowitz & S.J. Samuels (eds), *Comprehending Oral and Written Language*, New York, Academic Press: 55-82.
- Halliday, M.A.K. & Hasan, R. 1989. *Language, Context, and Text: Aspects of Language in a Social-Semiotic Perspective*. Oxford: Oxford University Press.
- Hansen, M.-B.M. 1997. “Alors” and “donc” in spoken French: A reanalysis. *Journal of Pragmatics* 28: 153-187.
- Hansen, M.-J.M. 2005. From prepositional phrase to hesitation marker. The semantic and pragmatic evolution of French *enfin*. *Journal of Historical Pragmatics* 6(1): 37-68.
- Hansen, M.-B.M. 2006. A dynamic polysemy approach to the lexical semantics of discourse markers (with an exemplary analysis of French *toujours*). In K. Fischer (Ed.), *Approaches to Discourse Particles*, Amsterdam, Elsevier: 21-41.
- Hansen, M.-B.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, A. 2011. Discourse marker and modal particle: The functions of utterance-final *then* in spoken English. *Journal of Pragmatics* 43: 3603-3623.
- Haselow, A. 2012. Subjectivity, intersubjectivity and the negotiation of common ground in spoken discourse: Final particles in English. *Language and Communication* 32:182-204.
- Hasselgren, A. 2002. Learner corpora and language testing: Small words as markers of learner fluency. In S. Granger, J. Hung & S. Petch-Tyson (eds), *Computer-Learner Corpora, Second Language Acquisition, and Foreign Language Teaching*, Philadelphia, John Benjamins: 143-173.
- Heeman, P. & Allen, J. 1999. Speech repairs, intonational phrases and discourse markers: Modeling speakers’ utterances in spoken dialog. *Computational Linguistics* 25(4): 1-45.
- Hesson, A. & Shellgren, M. 2015. Discourse marker *like* in real time: Characterizing the time-course of sociolinguistic impression formation. *American Speech* 90(2): 154-186.
- Hieke, A.E. 1985. A componential approach to oral fluency evaluation. *The Modern Language Journal* 69(2): 135-142.
- Hoffmann, S. 2004. Are low-frequency complex prepositions grammaticalized? On the limits of corpus data – and the importance of intuition. In H. Lindquist & C. Mair (eds), *Corpus Approaches to Grammaticalization in English*, Amsterdam, John Benjamins: 171-210.
- Hopper, P. 1987. Emergent grammar. In J. Aske, N. Beery, L. Michaelis & H. Filip (eds), *Proceedings of the Thirteenth Annual Meeting of the Berkeley Linguistics Society*, Berkeley, CA: Berkeley Linguistics Society.
- Hopper, P. & Traugott, E.C. 2003. *Grammaticalization (Second Edition)*. Cambridge: Cambridge University Press.

- Horowitz, R. & Samuels, S.J. 1987. Comprehending oral and written language: Critical contrasts for literacy and schooling. In R. Horowitz & S.J. Samuels (eds), *Comprehending Oral and Written Language*, San Diego, CA, Academic Press: 1-52.
- Hosman, L.A. & Siltanen, S.A. 2011. Hedges, tag questions, message processing, and persuasion. *Journal of Language and Social Psychology* 30(3): 341-349.
- Hothorn, T., Hornik, K. & Zeileis, A. 2006. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics* 15(3): 651-674.
- Hough, J., de Ruiter, L., Betz, S. & Schlangen, D. 2015. Disfluency and laughter annotation in a light-weight dialogue mark-up protocol. In *Proceedings of the 7th Workshop on Disfluency in Spontaneous Speech (DISS)*, Edinburgh, UK.
- House, J. 2013. Developing pragmatic competence in English as a lingua franca: Using discourse markers to express (inter)subjectivity and connectivity. *Journal of Pragmatics* 59: 57-67.
- Hunt, K. 1965. Grammatical Structures Written at Three Grade Levels. *NCTE Research Report 3*. Champaign Ill.: NCTE.
- Husserl, E. 1964. *The Phenomenology of Internal Time-Consciousness*. Bloomington, IN: Indiana University Press.
- Izutsu, M. & Izutsu, K. 2014. Truncation and backshift: Two pathways to sentence-final coordinating conjunctions. *Journal of Historical Pragmatics* 15(1): 62-92.
- James, C. 1980. *Contrastive Analysis*. Harlow: Longman.
- Janda, L. 2010. Cognitive linguistics in the year 2010. *International Journal of Cognitive Linguistics* 1(1): 1-30.
- Jaszczolt, K. 2003. On translating “what is said”: Tertium comparationis in contrastive semantics and pragmatics. In K. Jaszczolt & K. Turner (eds), *Meaning through Language Contrast*, Amsterdam, John Benjamins: 441-462.
- Jucker, A. 1993. The discourse marker *well*: A relevance-theoretical account. *Journal of Pragmatics* 19: 435-452.
- Jucker, A. & Ziv, Y. (eds). *Discourse Markers. Descriptions and Theory*. Amsterdam: John Benjamins.
- Kemmer, S. & Barlow, M. 2000. Introduction: A usage-based conception of language. In M. Barlow & S. Kemmer (eds), *Usage Based Models of Language*, Stanford, CSLI: vii-xxviii.
- Kennedy, S. & Trofimovich, P. 2008. Intelligibility, comprehensibility, and accentedness of L2 speech: The role of listener experience and semantic context. *Canadian Modern Language Review* 64: 459-490.
- Knott, A. & Dale, R. 1994. Using Linguistic Phenomena to Motivate Coherence Relations. *Discourse Processes* 18(1): 35-62.
- Koch, P. & Osterreicher, W. 2001. Gesprochene Sprache und geschriebene Sprache / Langage parlé et langage écrit. In G. Holtus, M. Metzeltin & C. Schmitt (eds), *Lexikon der Romanistischen Linguistik*, Bd. I / 2. Tübingen, Niemeyer: 584-627.
- Kohn, K. 2012. Pedagogic corpora for content and language integrated learning. Insights from the BACKBONE project. *The Eurocall Review* 20(2).

- Koponen, M. & Riegenbach, H. 2000. Overview: Varying perspectives on fluency. In H. Riegenbach (Ed.), *Perspectives on Fluency*, Ann Arbor, MI, The University of Michigan Press: 5-25.
- Kormos, J. 2006. *Speech Production and Second Language Acquisition*. London: Lawrence Erlbaum Associates.
- Krzeszowski, T.P. 1981. Tertium Comparationis. In J. Fisiak (Ed.), *Linguistics: Prospects and Problems*, Berlin, Mouton de Gruyter: 301-312.
- Kunz, K. & Laphinova-Koltunski, E. 2015. Cross-linguistic analysis of discourse variation across registers. *Nordic Journal of English Studies* 14(1): 258-288.
- Labov, W. 1972. *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.
- Lacheret, A., Kahane, S. & Pietrandrea, P. (eds). 2014. *Rhapsodie: A Prosodic and Syntactic Treebank for Spoken French*. Amsterdam: John Benjamins.
- Lakoff, G. 1987. *Women, Fire, and Dangerous Things. What Categories Reveal about the Mind*. Chicago: The University of Chicago Press.
- Lakoff, G. & Johnson, M. 1980. *Metaphors We Live By*. Chicago: University of Chicago Press.
- Lam, P. 2006. *Well but that's the effect of it: The use of well as a discourse particle in talk shows*. *Sprache und Datenverarbeitung (International Journal for Language Data Processing)* 30(1): 99-108.
- Langacker, R. 1987. *Foundations of Cognitive Grammar, Vol.1: Theoretical Prerequisites*. Stanford: Stanford University Press.
- Langacker, R. 1988a. An overview of Cognitive Grammar. In B. Rudzka-Ostyn (Ed.), *Topics in Cognitive Linguistics*, Amsterdam, John Benjamins: 3-48.
- Langacker, R. 1988b. A usage-based model. In B. Rudzka-Ostyn (Ed.), *Topics in Cognitive Linguistics*, Amsterdam, John Benjamins: 127-161.
- Langacker, R. 1990. *Concept, Image, and Symbol: The Cognitive Basis of Grammar* [Cognitive Linguistics Research 1]. Berlin: Mouton de Gruyter.
- Langacker, R. 2000. A dynamic usage-based model. In M. Barlow & S. Kemmer (eds), *Usage-Based Models of Language*, Stanford, CSLI: 1-63.
- Langacker, R. 2005. Construction grammars: Cognitive, Radical, and less so. In F. Ruiz de Mendoza Ibáñez & M. Sandra Peña Cervel (eds), *Cognitive Linguistics: Internal Dynamics and Interdisciplinary Interaction*, Berlin, Mouton de Gruyter: 101-159.
- Langacker, R. 2013. *Essentials of Cognitive Grammar*. Oxford: Oxford University Press.
- Lapshinova-Koltunski, E., Nedoluzhko, A. & Kunz, K. 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, H.-K. 2002. Towards a new typology of connectives with special reference to conjunction in English and Korean. *Journal of Pragmatics* 34: 851-866.
- Lehmann, C. 1985. Grammaticalization: Synchronic variation and diachronic change. *Lingua e Stile* 20: 303-318.

- Leijten, M. & Van Waes, L. 2013. Keystroke logging in writing research using inputlog to analyze and visualize writing processes. *Written Communication* 30(3): 358-392.
- Lenk, U. 1998. Discourse markers and global coherence in conversation. *Journal of Pragmatics* 30: 245-257.
- Lennon, P. 2000. The lexical element in spoken second language fluency. In H. Riggenbach (Ed.), *Perspectives on Fluency*, Ann Arbor, The University of Michigan Press: 25-42.
- Léon, P. 1993. *Précis de Phonostylistique, Parole et Expressivité*. Paris: Nathan Université.
- Levelt, W.J.M. 1981. The speaker's linearization problem. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences* 295(1077): 305-315.
- Levelt, W.J.M. 1983. Monitoring and self-repair in speech. *Cognition* 14: 41-104.
- Levelt, W.J.M. 1989. *Speaking: From Intention to Articulation*. Cambridge: MIT Press.
- Lewis, D. 2006a. Contrastive analysis of adversative relational markers, using comparable corpora. In K. Aijmer & A.-M. Simon-Vandenberghe (eds), *Pragmatic Markers in Contrast*, Amsterdam, Elsevier: 139-153.
- Lewis, D. 2006b. Discourse markers in English: A discourse-pragmatic view. In K. Fischer (Ed.), *Approaches to Discourse Particles*, Amsterdam, Elsevier: 43-60.
- Lindström, J. 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, P. 1982. *The Written Language Bias in Linguistics*. Linköping, Sweden: University of Linköping.
- Little, D.R., Oehmen, R., Dunn, J., Hird, K. & Kirsner, K. 2013. Fluency Profiling System: An automated system for analyzing the temporal properties of speech. *Behavioral Research Methods* 45(1): 191-202.
- Liu, K. & Fox Tree, J.E. 2012. Hedges enhance memory but inhibit retelling. *Psychon Bull Rev* 19, 892-898.
- Lopes, A., Martins de Matos, D., Cabarrão, V., Ribeiro, R., Moniz, H., Trancoso, I. & Mata, A.I. 2015. Towards using machine translation techniques to induce multilingual lexica of discourse markers. <http://arxiv.org/abs/1503.0914>.
- Lundholm, K. 2015. *Production and Perception of Pauses in Speech*. PhD thesis, University of Gothenburg.
- Luscher, J.-M. 1993. La marque de connexion complexe. *Cahiers de Linguistique Française* 14: 173-188.
- Luscher, J.-M. 1994. Marques de connexion: Procédures de traitement et guidage référentiel. In J. Moeschler, A. Reboul, J.-M. Luscher & J. Jayez, *Langage et pertinence. Référence Temporelle, Anaphore, Connecteurs et Métaphore*, Nancy, Presses universitaires de Nancy: 175-228.
- MacGregor, L., Corley, M. & Donaldson, D. 2009. Not all disfluencies are equal: The effects of disfluent repetitions on language comprehension. *Brain & Language* 111: 36-45.
- Maclay, H. & Osgood, C. 1959. Hesitation phenomena in spontaneous English speech. *Word* 15: 19-44

- 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.
- Mann, W. & Thompson, S. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text* 8(3): 243-281.
- Marchi, A. & Taylor, C. 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.
- Martin, R. & Slevc, I. 2014. Language production and working memory. In M.A. Goldrick, V.S. Ferreira & M. Miozzo (eds), *The Oxford Handbook of Language Production*, Oxford, Oxford University Press: 437-450.
- McCarthy, M. 2009. Rethinking spoken fluency. *Estudios de Lingüística Inglesa Aplicada* 9: 11-29.
- McHugh, M. 2012. Interrater reliability: The kappa statistic. *Biochemia Medica* 22(3): 276-282.
- Merlo, S. & Mansur, L. 2004. Descriptive discourse: Topic familiarity and disfluencies. *Journal of Communication Disorders* 37: 489-503.
- Meteer, M., Taylor, A., MacIntyre, R. & Iver, R. 1995. Disfluency annotation stylebook for the Switchboard corpus. Technical report, Linguistic Data Consortium.
- Meyer, T. & Popescu-Belis, A. 2012. Using sense-labeled discourse connectives for statistical machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, April 23-27, Avignon, France.
- Meyer, T., Popescu-Belis, A., Hajlaoui, N. & Gesmundo, A. 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, A., Wheeldon, L. & Krott, A. (eds). 2007. *Automaticity and Control in Language Processing*. Hove: Psychology.
- Meyer, D., Zeileis, A. & Hornik, K. 2014. vcd: Visualizing Categorical Data. R package version 1.3-2.
- Mieskes, M. & Strube, M. 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, M. 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.
- Moniz, H. 2013. *Processing Disfluencies in European Portuguese*. PhD thesis, Universidade de Lisboa.
- Moniz, H., Ferreira, J., Batista, F. & Trancoso, I. 2015. Disfluency detection across domains. In *Proceedings of DiSS 2015*.
- Moniz, H., Trancoso, I. & Mata, A.I. 2009. Classification of disfluent phenomena as fluent communicative devices in specific prosodic contexts. In *Proceedings of Interspeech 2009, ISCA*, Brighton, UK: 1719-1722.

- Mortier, L. & Degand, L. 2009. Adversative discourse markers in contrast: The need for a combined corpus approach. *International Journal of Corpus Linguistics* 14: 338-366.
- Mukherjee, J. 2005. The native speaker is alive and kicking: Linguistic and language-pedagogical perspectives. *Anglistik* 16(2): 7-23.
- Mulder, J. & Thompson, S. 2008. The grammaticization of *but* as a final particle in English conversation. In R. Laury (Ed.), *Crosslinguistic Studies of Clause Combining: The Multifunctionality of Conjunctions*, Amsterdam, John Benjamins: 179-204.
- Müller, S. 2005. *Discourse Markers in Native and Non-native English Discourse*. Amsterdam: John Benjamins.
- Murillo, S. 2016. Reformulation markers and polyphony. A contrastive English-Spanish analysis. *Languages in Contrast* 16(1): 1-30.
- Mustajoki, A. 2012. A speaker-oriented multidimensional approach to risks and causes of miscommunication. *Language and Dialogue* 2(2): 216-243.
- Nakatani, C. & Hirschberg, J. 1994. A corpus-based study of repair cues in spontaneous speech. *Journal of the Acoustical Society of America* 95(3): 1603-1616.
- Nelson, G., Wallis, S. & Aarts, B. 2002. *Exploring Natural Language: Working with the British Component of the International Corpus of English*. Amsterdam: John Benjamins.
- Neumann, S. 2014. Cross-linguistic register studies. Theoretical and methodological considerations. *Languages in Contrast* 14(1): 35-57.
- Nølke, H. 2006. Pour une théorie linguistique de la polyphonie: Problèmes, avantages, perspectives. In L. Perrin (Ed.), *Le Sens et ses Voix. Dialogisme et Polyphonie en Langue et en Discours*, Metz, Université Paul Verlaine: 243-269.
- Nome, A. & Haff, M. 2011. Une analyse contrastive de “donc”. In E. Khachaturyan (Ed.), *Discourse Markers in Romance Languages, Oslo Studies in Language* 3(1): 47-67.
- Norrick, N. 2009. Interjections as pragmatic markers. *Journal of Pragmatics* 41: 866-891.
- Nzoimbengene, P. 2016. *Les ‘Discourse Markers’ en Lingála. Étude Sémantique et Pragmatique sur Base d’un Corpus de Lingála de Kinshasa Oral*. PhD thesis, Université catholique de Louvain.
- O’Connell, D.C. & Kowal, S. 1972. Cross-linguistic pause and rate phenomena in adults and adolescents. *Journal of Psycholinguistic Research* 1: 155-164.
- O’Donnell, W. & Todd, L. 1980. *Variety in Contemporary English*. London: Allen and Unwin.
- Osborne, J. 2011. Fluency, complexity and informativeness in native and non-native speech. *International Journal of Corpus Linguistics* 16(2): 276-298.
- Östman, J.-O. 1995. Pragmatic particles twenty years after. In B. Wårvik, S.-K. Tanskanen & R. Hiltunen (eds), *Organization in Discourse* [Anglicana Turkuensia, Vol. 14], Department of English, University of Turku, Finland: 95-108.
- Oza, U., Prasad, R., Kolachina, S., Meena, S., Sharma, D. & Joshi, A. 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, S. 1997. “Artisans”, “Assureurs”, Conversations Téléphoniques en Entreprise. Retrieved from <http://clapi-univ.lyon2.fr> (last accessed March 2014).

- Pallaud, B., Rauzy, S. & Blâche, P. 2013a. Auto-interruptions et disfluences en français parlé dans quatre corpus du CID. *TIPA Travaux Interdisciplinaires sur la Parole et le Langage* 29.
- Pallaud, B., Rauzy, S. & Blâche, P. 2013b. Identification et annotation des auto-interruptions et des disfluences dans les corpus du CID. Technical report, Laboratoire Parole et Langage.
- Pander Maat, H.L.W. & Degand, L. 2001. Scaling causal relations and connectives in terms of speaker involvement. *Cognitive Linguistics* 12(3): 211-245.
- Pander Maat, H.L.W. & Sanders, T.J.M. 2000. Domains of use and subjectivity. On the distribution of three Dutch causal connectives. In B. Kortmann & E. Couper-Kuhlen (eds), *Cause, Condition, Concession and Contrast: Cognitive and Discourse Perspectives*, Berlin, Mouton de Gruyter: 57-82.
- Pander Maat, H.L.W. & Sanders, T.J.M. 2001. Subjectivity in causal connectives: An empirical study of language in use. *Cognitive Linguistics* 12(3): 247-273.
- Pascual, E. & Crible, L. 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, A. & Syder, F. 1975. Sentence formulation in spontaneous speech. *New Zealand Speech Therapists' Journal* 30(2): 2-11.
- Pawley, A. & Syder, F. 1983. Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. Richards & R. Schmidt (eds), *Language and Communication*, London, Longman: 191-225.
- Pawley, A. & Syder, F. 2000. The one-clause-at-a-time hypothesis. In H. Riggebbach (Ed.), *Perspectives on Fluency*, Ann Arbor, The University of Michigan Press: 163-199.
- Pichler, H. 2016. Uncovering discourse-pragmatic innovations: Innit in multicultural London English. In H. Pichler (Ed.), *Discourse-Pragmatic Variation and Change in English: New Methods and Insights*, Cambridge, Cambridge University Press: 59-85.
- Pichler, H. & Hesson, A. 2016. Discourse-pragmatic variation across situations, varieties, ages: I DON'T KNOW in sociolinguistic and medical interviews. *Language & Communication* 49: 1-18.
- Pitler, E. & Nenkova, A. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP Conference Short Papers*: 13-16.
- Poggi, I. & Magno Caldognetto, E. 1996. A score for the analysis of gestures in multimodal communication. In L. Messing (Ed.), *Proceedings of the Workshop on the Integration of Gesture and Language in Speech*, Applied Science and Engineering Laboratories, Newark and Wilmington, Delaware: 235-244.
- Pons Bordería, S. 2006. A functional approach to the study of discourse markers. In K. Fischer (Ed.), *Approaches to Discourse Particles*, Amsterdam, Elsevier: 77-100.
- Pons Bordería, S. 2008. La combinación de marcadores del discurso en la conversación coloquial: Interacciones entre posición y función. *Estudios Lingüísticos/Linguistic Studies* 2 :141-159.
- Pons Bordería, S. & Estellés Arguedas, M. 2009. Expressing digression linguistically: Do digressive markers exist? *Journal of Pragmatics* 41: 921-936.

- Popescu-Belis, A., Meyer, T., Liyanapathirana, J., Cartoni, B. & Zufferey, S. 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, A., Kolk, H. & Povel, D.-J. 1990. On the relation among speech errors, disfluencies, and self-repairs. *Language and Speech* 33(1): 19-29.
- Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A. & Webber, B. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 08)*, Marrakech, Morocco: 2961-2968.
- Rabab'ah, G. & Abuseileek, A. 2012. The pragmatic functions of repetition in TV discourse. *Research in Language* 10(4): 445-460.
- Radden, G. 1992. The cognitive approach to natural language. In M. Pütz (Ed.), *Thirty Years of Linguistic Evolution*, Amsterdam, John Benjamins: 513-542.
- Ragan, S. 1983. Alignment and Conversational Coherence. In R. Craig & K. Tracy (eds), *Conversational Coherence: Form, Structure and Strategy*, Beverly Hills, Sage Publications: 157-171.
- Razgouliaeva, A. 2002. Combinaisons des connecteurs mais enfin. *Cahiers de Linguistique Française* 24: 143-168.
- Redeker, G. 1991. Linguistic markers of discourse structure. *Linguistics* 29: 1139-1172.
- Reese, B. & Asher, N. 2007. Prosody and the interpretation of tag questions. In E. Puig-Waldmüller (Ed.), *Proceedings of Sinn und Bedeutung* 11, Barcelona, Universitat Pompeu Fabra: 448-462.
- Rendle-Short, J. 2004. Showing structure: Using um in the academic seminar. In *Pragmatics* 14(4): 479-498.
- Roberts, B. & Kirsner, K. 2000. Temporal cycles in speech production. *Language and Cognitive Processes* 15(2): 129-157.
- Roekhaut, S., Brognaux, S., Beaufort, R. & Dutoit, T. 2014. eLite-HTS: un outil TAL pour la génération de synthèse HMM en français. Paper presented at the *Journées d'Etude de la Parole (JEP)*, Le Mans, France.
- Rohde, H. & Horton, W. 2014. Anticipatory looks reveal expectations about discourse relations. *Cognition* 133(3): 667-691.
- Romaine, S. & Lange, D. 1991. The use of *like* as a marker of reported speech and thought: A case of grammaticalization in progress. *American Speech* 66(3): 227-279.
- Romano, M. & Cuenca, M.J. 2013. Discourse markers, structure and emotionality in oral narratives. *Narrative Inquiry* 23: 344-370.
- Rosch, E. 1975. Cognitive reference points. *Cognitive Psychology* 7: 532-547.
- Rossari, C. 1990. Projet pour une typologie des opérations de reformulation. *Cahiers de Linguistique Française* 11: 345-359.
- Rossari, C. 1994. *Les Opérations de Reformulation*. Bern: Peter Lang.
- Rouchota, V. 1996. Discourse connectives: What do they link? *UCL Working papers in Linguistics* 8: 1-15.

- Roulet, E., Auchlin, A., Moeschler, J. & Rubattel, C. 1985. *L'Articulation du Discours en Français Contemporain*. Bern: Peter Lang.
- Roze, C., Danlos, L. & Muller, P. 2012. LEXCONN: A French lexicon of discourse connectives. *Discours* 10.
- Rühlemann, C. & O'Donnell, M. 2012. Introducing a corpus of conversational stories. Construction and annotation of the Narrative Corpus. *Corpus Linguistics and Linguistic Theory* 8(2): 313-350.
- Sanders, T.J.M. 1997. Semantic and pragmatic sources of coherence: On the categorization of coherence relations in context. *Discourse Processes* 24(1): 119-147.
- Sanders, T.J.M, Spooren, W. & Noordman, L. 1992. Toward a taxonomy of coherence relations. *Discourse Processes* 15: 1-35.
- Sanders, T.J.M, Demberg, V., Evers-Vermeul, J., Hoek, J., Scholman, M. & Zufferey, S. forthcoming. Unifying dimensions in discourse relations: How various annotation frameworks are related.
- Santorini, B. 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, E., Jefferson, G. & Sacks, H. 1977. The preference for self-correction in the organization of repair in conversation. *Language* 53, 361-382.
- Schiffrin, D. 1987. *Discourse Markers*. Cambridge: Cambridge University Press.
- Schmid, H. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing* 12: 44-49.
- Schmid, H.-J. 2000. *English Abstract Nouns as Conceptual Shells. From Corpus to Cognition*. Berlin: Mouton de Gruyter.
- Schmid, H.-J. 2012. Generalizing the apparently ungeneralizable. Basic ingredients of a cognitive-pragmatic approach to the construal of meaning-in-context. In H.-J. Schmid (Ed.), *Cognitive Pragmatics*, Berlin, Mouton de Gruyter: 3-22.
- Schmidt, R. 1992. Psychological mechanisms underlying language fluency. *Studies in Second Language Acquisition* 14: 357-385.
- Schmidt, T. & Wörner, K. 2012. EXMARaLDA. In J. Durand, G. Ulrike & G. Kristoffersen (eds), *Handbook on Corpus Phonology*, Oxford, Oxford University Press: 402-419.
- Schönefeld, D. 1999. Corpus linguistics and cognitivism. *International Journal of Corpus Linguistics* 4(1): 137-171.
- Schourup, L. 1999. Discourse markers. *Lingua* 107: 227-265.
- Schourup, L. 2001. Rethinking well. *Journal of Pragmatics* 33: 1025-1060.
- Schourup, L. 2011. The discourse marker *now*: A relevance-theoretic approach. *Journal of Pragmatics* 43: 2110-2129.
- Segalowitz, N. 2010. *Cognitive Bases of Second Language Fluency*. New York: Routledge.
- Seyfeddinipur, M. 2006. *Disfluency: Interrupting Speech and Gesture*. MP Series in Psycholinguistics.

- Shortall, T. 2007. *Cognition, Corpus, Curriculum*. PhD thesis, University of Birmingham.
- Shriberg, E. 1994. *Preliminaries to a Theory of Speech Disfluencies*. PhD thesis, University of California at Berkeley, CA.
- Simon, A.-C., Auchlin, A., Avanzi, M. & Goldman, J.-P. 2010. Les phonostyles: Une description prosodique des styles de parole en français. In M. Abecassi & G. Ledegen (eds), *Les Voix des Français. En Parlant, en Ecrivant*, Bern, Peter Lang: 71-88.
- Šliogerienė, J., Valūnaitė Oleškevičienė, G. & Asijavičiūtė, V. 2015. Discourse relational devices of contrast in Lithuanian and English. *Santalka = Coactivity: Filologija. Edukologija* 23(2): 92-100.
- Sperber, D. & Wilson, D. 1986. *Relevance. Communication and Cognition*. Oxford: Blackwell.
- Spooren, W. & Degand, L. 2010. Coding coherence relations: Reliability and validity. *Corpus Linguistics and Linguistic Theory* 6(2): 241-266.
- Strassel, S. 2003. Simple metadata annotation specification v.5. Technical report, Linguistic Data Consortium.
- Stubbe, M. & Holmes, J. 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.
- Stukker, N. & Sanders, T.J.M. 2012. Subjectivity and prototype structure in causal connectives: A cross-linguistic perspective. *Journal of Pragmatics* 44: 169-190.
- Sweetser, E. 1990. *From Etymology to Pragmatics*. Cambridge: Cambridge University Press.
- Sweller, J. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive Science* 12: 257-285.
- Swerts, M. 1998. Filled pauses as markers of discourse structure. *Journal of Pragmatics* 30: 485-496.
- Tesnière, L. 1959. *Éléments de Syntaxe Structurale*. Paris: Klincksieck.
- Tonelli, S., Riccardi, G., Prasad, R. & Joshi, A. 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.
- Torgersen, E., Gabrielatos, C., Hoffmann, S. & Fox, S. 2011. A corpus-based study of pragmatic markers in London English. *Corpus Linguistics and Linguistic Theory* 7(1): 93-118.
- Torrey, C., Fussell, S.R. & Kiesler, S. 2013. How a robot should give advice. In *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*: 275-282.
- Tottie, G. 2011. Uh and Um as sociolinguistic markers in British English. *International Journal of Corpus Linguistics* 16(2): 173-197.
- Tottie, G. 2015a. Uh and um in British and American English: Are they words? Evidence from co-occurrence with pauses. In N. Dion, A. Lapierre & R. Torres Cacoullos (eds), *Linguistic variation: Confronting Fact and Theory*, New York, Routledge: 38-54.
- Tottie, G. 2015b. From pause to word: Uh and um in written language. Paper presented at *ICAME 36*, May 27-31, Trier, Germany.

- Traugott, E.C. 1995. The role of the development of discourse markers in a theory of grammaticalization. Paper presented at *ICHL XII*, Manchester, UK.
- Traugott, E.C. 2007. (Inter)subjectification and unidirectionality. *Journal of Historical Pragmatics* 8: 295-309.
- Traxler, M.J., Bybee, M.D. & Pickering, M.J. 1997. Influence of connectives on language comprehension: Eye-tracking evidence for incremental interpretation. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, 50A(3): 481-497.
- Turner, R. 1971. Words, utterances, activities. In J.D. Douglas (Ed.), *Understanding Everyday Life: Towards a Reconstruction of Sociological Knowledge*, London, Routledge & Kegan Paul: 169-187.
- Underhill, R. 1988. Like is, like, focus. *American Speech* 63(3): 234-246.
- Unger, C. 1996. The scope of discourse connectives: Implications for discourse organization. *Journal of Linguistics* 32: 403-438.
- Urgelles-Coll, M. 2012. *The Syntax and Semantics of Discourse Markers*. London: Bloomsbury.
- Valdmets, A. 2013. Modal particles, discourse markers, and adverbs with *It*-suffix in Estonian. In L. Degand, B. Cornillie & P. Pietrandrea (eds), *Discourse Markers and Modal Particles. Categorization and Description*, Amsterdam, John Benjamins: 107-132.
- Van Bogaert, J. 2011. I think and other complement-taking mental predicates: A case of and for constructional grammaticalization. *Linguistics* 49(2): 295-332.
- Van Dijk, T. 1979. Pragmatic connectives. *Journal of Pragmatics* 3: 447-456.
- Van Dijk, T. 1997. Cognitive context models and discourse. In M. Stamenow (Ed.), *Language Structure, Discourse and the Access to Consciousness*, Amsterdam, John Benjamins: 189-226.
- van Enschoot, R., Spooren, W., van den Bosch, A., Burgers, C., Degand, L., Evers-Vermeul, J., Kunneman, F., Liebrecht, C., Linders, Y. & Maes, A. under review. Taming our wild data: On intercoder reliability in discourse research.
- Vasilescu, I., Nemoto, R. & M. Adda-Decker. 2007. Vocalic hesitations vs vocalic systems: A cross-language comparison. In *Proceedings of the ICPHS 16th International Congress of Phonetic Science*.
- Vinay, J.-P. & Darbelnet, J. 1995. *Comparative Stylistics of French and English: A Methodology for Translation*. Translated and ed. by J. Sager & M.-J. Hamel. Amsterdam: John Benjamins.
- Vincent, D. 1993. *Les Ponctuels de la Langue et Autres Mots du Discours*. Québec: Nuit Blanche Editeur.
- Wagner, S. & Hesson, A. 2014. Individual sensitivity to the frequency of socially meaningful linguistic cues affects language attitudes. *Journal of Language and Social Psychology* 33(6): 651-666.
- Waltereit, R. 2007. À propos de la genèse diachronique des combinaisons de marqueurs. L'exemple de *bon ben* et *enfin bref*. *Langue Française* 154: 94-128.
- Waltereit, R. & Detges, U. 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, M., Hirose, K., Den, Y. & Minematsu, N. 2008. Filled pauses as cues to the complexity of up-coming phrases for native and non-native listeners. *Speech Communications* 50: 81-94.

- Wauquier-Gravelines, S. 1999. Segmentation lexicale de la parole continue: La linéarité en question. *Recherches linguistiques de Vincennes* 28: 133-156.
- Willems, D. & Demol, A. 2006. *Vraiment* and *really* in contrast: When truth and reality meet. In K. Aijmer & A.-M. Simon-Vandenberghe (eds), *Pragmatic Markers in Contrast*, Amsterdam, Elsevier: 215-235.
- Wilson, D. 2011. The conceptual-procedural distinction: Past, present and future. In V. Escandell-Vidal, M. Leonetti & A. Ahern (eds), *Procedural Meaning: Problems and Perspectives* [Current Research in the Semantics/Pragmatics Interface 25], Emerald, Bingley: 3-31.
- Wilson, D. & Sperber, D. 1993. Linguistic form and relevance. *Lingua* 90: 1-25.
- Wingate, M. 1987. Fluency, disfluency: Illusion and identification. *Journal of Fluency Disorders* 12(2): 4-87.
- Zechner, K. 2001. *Automatic Summarization of Spoken Dialogues in Unrestricted Domains*. PhD thesis, Carnegie Mellon University.
- Zeileis, A., Meyer, D. & Hornik, K. 2007. Residual-based shadings for visualizing conditional independence. *Journal of Computational and Graphical Statistics* 16(3): 507-525.
- Zeyrek, D., Demirşahin, I., Sevdik Çalli, A. & Çakici, R. 2013. Turkish Discourse Bank: Porting a discourse annotation style to a morphologically rich language. *Dialogue & Discourse* 4(2): 174-184.
- Zhao, Y. & Jurafsky, D. 2005. A preliminary study of Mandarin filled pauses. In *Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop*, September 10-12, Aix-en-Provence, France: 179-182.
- Zufferey, S. 2010. *Lexical Pragmatics and Theory of Mind. The Acquisition of Connectives*. Amsterdam: John Benjamins.
- Zufferey, S. & Cartoni, B. 2012. English and French causal connectives in contrast. *Languages in Contrast* 12(2): 232-250.
- Zufferey, S. & Degand, L. in press. Representing the meaning of discourse connectives for multilingual purposes. *Corpus Linguistics and Linguistic Theory* 10.
- Zufferey, S. & Popescu-Belis, A. 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.

Appendices

Appendix 1: DM-level annotation protocol (Crible 2014)

The first appendix consists in the annotation protocol designed for all DM-level variables. The full document is reported below in its original format, and thus contains its own references and section numbers. The original standalone document is available upon request.

Identifying and describing discourse markers in spoken corpora

Annotation protocol v.8

Ludivine Crible

15th December, 2014



Université Catholique de Louvain
Institute Language & Communication

Ludivine CRIBLE
ludivine.crible@uclouvain.be

ARC Research Grant 12/17-044

1 Introduction

1.1 Background

Discourse marker research today still faces many terminological and theoretical issues which restrain progress in the field, despite the multiplicity of theoretical frameworks and approaches taken by many valuable works over the last decades (e.g. Schiffrin 1987; Brinton 1996; Fraser 1999; Aijmer and Simon-Vandenberg 2006; Fischer 2006a; Traugott 2007; Waltereit and Detges 2007; Degand, Cornillie, and Pietrandrea 2013 to name but a few). The field suffers from lack of consensus on the category of discourse markers (henceforth DMs), its definition and what it contains. Such differences render comparisons of results inadequate, since there is usually only limited overlap between the scope of the various studies. Reasons for these discrepancies may lie in the choice of theoretical framework (coherence theory vs relevance theory, for instance), restrictions of items under consideration, type of data (medium, register), method and purpose of annotation, and possibly others.

Existing literature is usually of two kinds: either a theoretical, usually quite abstract account of variables that might affect the behavior of DMs (Schiffrin 1987; Brinton 1996) (see Bolly, Crible, et al. (2014) for an operational exception), or more in-depth case studies that specify a method but only for a certain type of elements, for instance connectives (Meyer et al. 2011; Zufferey and Degand 2014), markers of concession (Taboada and Gómez-González 2012), contrastive pairs (Bazzanella et al. 2007; Hasselgard 2006) or even, for a great majority of works, only one discourse marker (e.g. Aijmer 1997; Cuenca 2008; Denturck 2008). The first type is rarely operationalized, while the second is hardly reproducible to a larger extent or to other data. However, we acknowledge our debt to the major contribution of the Penn Discourse Tree Bank (PDTB) initiative (Prasad et al. 2007) - and especially the revised scheme in Zufferey and Degand (2014) - which serves as a model for many language-specific taxonomies, with only “a number of adjustments in the sublevels in order to account for all the specificities of their language” (ibid.: 5). The protocol proposed here is clearly situated in the line of the PDTB, although major modifications were implemented to improve its operational application and to extend its scope to all types of DMs, not only so-called connectives.⁶⁴

This research is part of the collaborative project “Fluency and disfluency markers: a multimodal contrastive perspective”⁶⁵ which investigates the ambivalent phenomenon of (dis)fluency through the lens of language, modality and various “fluencemes” (Gotz 2013) i.e. potential clues of fluent or disfluent speech such as prosodic information, reformulations, repetitions, filled pauses and discourse markers. Our contribution to this project will be to situate the role of DMs within a typology of fluencemes according to situational and linguistic variables that affect their behavior in context. In other words, our project aims at determining operational categorizing tools and methods for the evaluation and interpretation of a marker, and the text segment it belongs to, as fluent or disfluent. The present manual thus corresponds to the first annotation phase (parametric description of DMs), before the cross-examination of the obtained values with the annotation of disfluent phenomena, in line with Shriberg (1994) and following guidelines designed by Crible, Dumont, et al. (2015).

1.2 Purpose of this manual

This report documents the annotation protocol and method that was applied to *DisFrEn* (Crible 2014), a comparable database⁶⁶ of seventeen hours of speech in native French and English across various

⁶⁴ Terminological choices will be explained in section 2.1.

⁶⁵ ARC Research Grant number 12/17-044, Université Catholique de Louvain, spokesperson Liesbeth Degand.

⁶⁶ *DisFrEn* is a collection of texts sampled from existing corpora and balanced across languages and situations. Details and references will be given in section 2.2.

situations such as conversations, news broadcast or phone calls. It is designed to provide operational guidelines to the parametric description of DMs in context regarding syntactic and pragmatic variables such as position or function. It will also serve to explain all the theoretical decisions made in order to achieve the present protocol, based on existing literature in the field of discourse marker research and sometimes supplemented by specific requirements of our own project.

This manual accounts for the eight version of my annotation scheme which is itself a revision of previous versions that revealed several flaws, although the general approach had been broadly reviewed by experts in the field (many thanks to Pr. Liesbeth Degand⁶⁷, Pr. Sandrine Zufferey⁶⁸ and Pr. Gaëtanelle Gilquin⁶⁹ for their careful readings and advice on the different steps of this protocol). This protocol is currently undergoing experimentation with different coders (both naive and experts) on different data (other languages and modalities) (see Bolly and Crible 2015; Crible and Degand 2015; Crible and Zufferey 2015). After empirical testing on a subcorpus of French and English interviews from the Backbone corpus (Kohn 2012), problems encountered during this pilot annotation especially in terms of operationalization have been addressed and we hope that this new scheme will provide a robust base for further exploration of discourse markers in *DisFrEn* or any other corpus.

This document is structured as follows: definition of the object of study, corpus and conventions, summary of the annotations covered and specificities of this version (section 2); detailed account of all the annotation tiers and possible values (section 3); focus on the disambiguation of semantic and pragmatic pairs of functions (section 4); focus on the disambiguation of frequent polysemous DMs (section 5); discussion and conclusion (section 6).

2 Overview of the protocol: design and applications

2.1 Scope of this scheme: What are discourse markers ?

As mentioned before, a lot of different definitions of what is to be included in the category of discourse markers are conflicting in previous works. Therefore, before turning to the actual annotation protocol, it seems necessary to specify what are the elements the protocol applies to. First, a terminological note: the term “discourse marker” is preferred to “pragmatic markers” as used by Brinton (1996) or Aijmer and Simon-Vandenberg (2006), since the latter tends to refer not only to the items under investigation here (e.g. *you know, well, because, in other words, sort of*) but also to any pragmatic item. We follow here many authors (e.g. Hansen 2006; Waltereit and Detges 2007) who assign to “pragmatic markers” a much broader definition (Hansen 2006: 28):

Discourse marker should be considered a hyponym of *pragmatic marker*, the latter being a cover term for all those nonpropositional functions which linguistic items may perform 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.

Figure 1 represents the category of pragmatic markers with their subgroups, while there may be more.

⁶⁷ Université Catholique de Louvain

⁶⁸ Université de Fribourg, Switzerland

⁶⁹ Université Catholique de Louvain, Center for English Corpus Linguistics

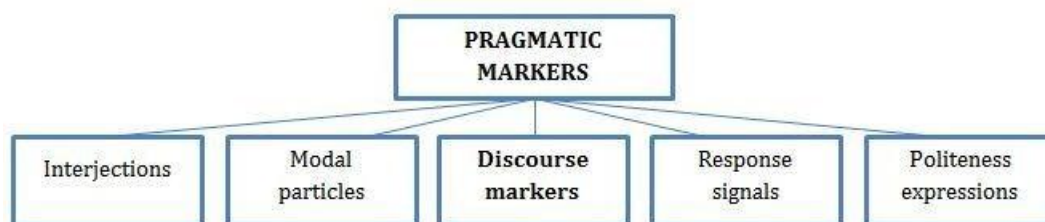


Figure 1: Taxonomy of pragmatic markers

As for the definition of what counts as a DM, we elaborated a combined version of several existing proposals (Brinton 1996; Hansen 2006; Schiffrin 1987; Schourup 1999) in order to have the most explicit, unequivocal phrasing in as few words as possible. Therefore, DMs are a grammatically heterogeneous, multifunctional type of pragmatic markers, hence syntactically optional and non-truth-conditional, constraining the inferential mechanisms of interpretation processes. Their specificity as part of the PM category is to function on a metadiscursive level as a cue to situate the host unit in a co-built representation of on-going speech. They do so by either signaling a discourse relation between the host unit and its context (see section 3.3 for delimitation of the context), explicating the structural sequencing of discourse segments, expressing the speaker's meta-comment on his phrasing, or contributing to interpersonal collaboration.

Additional characteristics of DMs are: a procedural meaning, a variable degree of syntactic integration referred to as “weak-clause association” (Schourup 1999: 232), a variable scope (over either one host unit, two related textual units of various types, one host unit and contextual assumptions). Several authors (e.g. Brinton 1996; Schiffrin 1987) mention other features of discourse marker behavior which are prototypical but mostly optional and not systematic, such as: short lexemes, characteristic of spoken-like register, high frequency, prosodic independence and stress.

Hesitations may arise about several elements of speech which are problematic to categorize:

Fillers: if some studies accept language-specific “fillers” such as *uhm* or *euh* in the category of DMs (Rendle-Short 2004; Swerts 1998; Tottie 2011), their semantic content is too abstract and functional differences very subjective to distinguish. In our project, fillers will be accounted for in a further step of the annotation, where they will be considered a specific fluenceme, along with silent pauses, reformulations etc.

Interjections: for similar reasons, standalone interjections like *ah* or *oh* are not considered DMs in their phatic or modal use linked to information state, but rather belong to another subcategory of pragmatic markers (Norrick 2009). However, interjections sometimes do express a clear discursive function, namely introduction of reported speech for English *oh*. In other cases, they are combined in a complex marker (i.e. composed of two distinct DMs) where the individual meanings cannot be perceived anymore, as in French *eh bien*⁷⁰. These exceptions can be considered DMs after disambiguation in context.

Response signals: while primary agreement particles like *yes*, *yeah*, *right* or *voilà* are most often instances of pragmatic markers (except for rare lexical uses such as “pour un oui ou pour un non”), they are only considered DMs when they display another function like topic-closing. In other words, these elements behave as DMs only when they do not correspond to the propositional content of an answer to a question, but rather when they perform a discourse function such as closing a topic unit or turn-taking.

⁷⁰ We prefer the spelling *eh bien* as opposed to *et bien* for reasons of frequency in use and semantic demotivation.

Epistemic parentheticals: it is quite complex to distinguish between full propositional and discursive uses of so-called “epistemic parentheticals” such as *I think, I suppose, je pense*. Dehé and Wichmann (2010) identify several prosodic criteria to help disambiguate the two types of uses in English which are actually on a cline of grammaticalisation from propositional to formulaic. According to them, non-propositional uses of epistemic parentheticals in English can be identified by their prosodic integration (no pause at left and right periphery of the expression) and their deaccentuation (unstressed) (Dehé and Wichmann 2010: 24). Additionally, Kaltenbock (2009) mentions that *I think* in initial position is rarely the main clause of the utterance, i.e. it is used primarily as a DM, in a “secondary status as a qualifier of the proposition” (ibid.: 67). These discriminating features will be used to identify the DM uses of epistemic parentheticals upon hearing the passage containing them, provided that sound is available to the annotator.

General extenders: these items, also called “vague category markers” (Stenstrom 2009), correspond to expressions such as *or something, and things like that* or *et tout* which show all characteristics of DMs except for the fact that their position is rather fixed at the end of a unit, and that some of them tend to be quite long. However, other well-established DMs are also fixed in final position (e.g. French *quoi*) and, as to length, only grammaticalised items will be considered DMs, thus eliminating longer and somewhat idiosyncratic expressions. Both Cheschire (2007) and Pichler and Levey (2011) discuss their variable degree of grammaticalisation, the shorter forms being the more grammaticalised, and observe that they are subject to “a great deal of individual variation” (Cheschire 2007: 187). Buysse (2014) explicitly advocates for their categorization as DMs: “Because they fulfil (interpersonal) functions in spoken language similar to those of prototypical members of the class of pragmatics markers, they have often been considered a subset within this class, with a clearly distinguishable structure” (ibid.: 2). Stenstrom (2009) provides a list of general extenders that can be used to select them during the annotation process.

Tag questions: Literature that discusses the categorization of tag questions such as *isn't it* as DMs or not is rather scarce and doesn't reach consensus. Although they may carry rather interpersonal functions (checking for attention, monitoring), they are morphologically flexible and contain propositional value, which contradicts our definition of DMs. In some contexts, tag questions also perform rhetorical, pragmatic functions such as provocation or irony which can hardly be described in terms of discourse marking. Lacking further arguments in favour of the treatment of tag questions as DMs, they will not be considered within the scope of this annotation protocol.

Editing terms: Finally, certain items, otherwise identified as “explicit editing terms” (Shriberg 1994), signal production troubles on the part of the speaker: *comment dirais-je, I don't know, si je peux dire* etc. These rather formulaic expressions refer explicitly to the locutionary act with a reference to first-person pronoun and verbs of saying or knowing. The boundary between discourse marking and text editing can be sometimes thin. Therefore, the following criteria were used to distinguish between DMs and editing terms, the latter being accounted for in a second step of the annotation as markers of (dis)fluency: explicit reference to lexical access trouble (which excludes from DMs *comment, comment dire, comment dirais-je*), low degree of grammaticalisation and traces of propositional content (which exclude *si je peux dire, if I may say so*). Borderline cases are, for instance, *si vous voulez, si tu veux, if you will, je dirais, on va dire, I don't know*, which present a high degree of fixation while explicitly referring to the act of speaking or thinking. These will be considered DMs for our purposes. It remains that the selection of DMs, unlike their functional description, can never be devoid of language-specific considerations and remains partly intuitive when it comes to items at the edges of the category.

Further details regarding the categorization of DMs can be found in Crible (forthcoming) or further in this manual (see section 3). It remains that selection of DMs may still be subject to the annotator's subjectivity and intuition, especially since some of the DM candidates are still undergoing grammaticalisation and fixation. In this, DM studies do not differ from other domains in semantics-pragmatics, which is a rather shady area of research with fuzzy categories, hence prone to subjectivity.

2.2 Source corpora and annotation method

The annotation protocol under discussion here, although potentially universal, was tested and designed for the description of DMs in *DisFrEn* (Crible 2014). This bilingual corpus collection has been compiled from the following source corpora of English and French: ICE-GB (Nelson, Wallis, and Aarts 2002), Backbone (Kurt 2012), VALIBEL (Dister et al. 2009), LOCAS-F (Degand, Martin, and Simon 2014)⁷¹, CLAPI (Palisse 1997), CPhonoGenre (Prsir, Goldman, and Auchlin 2013) and Rhapsodie Treebank (Lacheret, Kahane, and Pietrandrea 2014). All these corpora had different file format and diverging transcription conventions, which were homogenized and converted into a machinereadable form to be annotated under the EXMARaLDA suite (Schmidt and Wörner 2012).

Transcripts are sound-aligned, using the automatic aligner EasyAlign (Brogniaux et al. 2012; Roekhaut et al. 2014) when needed and loaded onto Partitur Editor, EXMARaLDA's annotation tool (see figure 2 for a visualisation of the interface).

The annotation process is entirely manual (except for an optional part-of-speech tagging performed by and imported from TreeTagger (Schmid 1995)) and horizontal: the analyst browses the transcript until she finds an item that suits the definition of DM. This item is then described in separate tiers (see section 2.3) according to criteria defined in this protocol. All variables for the same item are assigned at once, for an economy of time and effort. Although it might be advocated that “vertical” annotation (i.e. the same variable is assigned at once to all items, before turning to the next variable etc.) is better suited to avoid circularity, it would be too time-consuming in the case of pragmatic annotation which requires a certain undersanding of the item's context.

The annotated transcripts are converted into a .coma file that can be run into EXAKT, EXMARaLDA's concordancer, thus giving Excel-style concordance lines and the annotation values for each marker. However, the present annotation protocol does not depend on the software used, provided the program allows multi-layer annotation (e.g. ELAN or Praat).

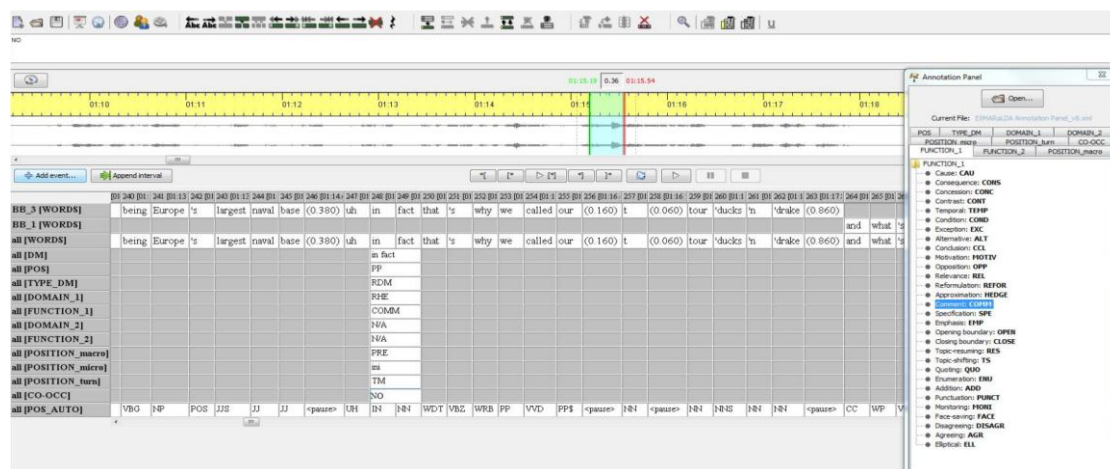


Figure 2: Interface of Partitur Editor annotation tool

⁷¹ Some of these texts were originally compiled in the C-PROM corpus (Avanzi et al. 2010).

2.3 Summary of annotations

The present protocol specifies eleven annotation tiers, each one depending on the transcription tiers (one per speaker, plus an additional tier that concatenates all transcriptions, for technical reasons regarding extraction of the data). Each tier corresponds to a different linguistic variable, and these parameters can be subsumed in three main groups: type of DM, function(s) and position. Although all tiers will be detailed at length in section 3, a brief summary will be provided here to give some overview of what this protocol covers. Table 1 presents, for each tier, its tag, definition, number of possible values, and some references when used for the design of the particular variable.

Nbr	Tag	Tier definition	Nbr of values	References
1	DM	full-word orthographic transcription of the utterance of a discourse marker	corpus-driven	Brinton 1996; Hansen 2006
2	POS	source grammatical class of the marker	9	Santorini 1990
3	TYPE DM	situation of the marker on the scale of relationality	3	Degand and Simon-Vandenberg 2011; Degand and Gilquin 2013
4	DOMAIN 1	component of language structure affected by the marker, source of coherence	4	Halliday and Hasan 1976; Sweetser 1990
5	FUNCTION 1	specifies the discourse relation or otherwise pragmatic function of the marker	30	Prasad et al. 2007; Zufferey and Degand 2014
6	DOMAIN 2	possible second component of language structure affected by the marker, source of coherence	4	Halliday and Hasan 1976; Sweetser 1990
7	FUNCTION 2	specifies the possible second discourse relation or otherwise pragmatic function of the marker	30	Prasad et al. 2007; Zufferey and Degand 2014
8	POSITION macro	position of the marker in relation to the macro-syntactic dependency structure	7	Lindström 2001
9	POSITION micro	position of the marker in relation to the micro-syntactic structure, including subordinated clauses	5	—

10	POSITION turn	position of the marker in relation to the turn of speech	4	Bolly, Crible, et al. 2014
11	CO-OCC	whether the annotated item cooccurs with another discourse marker and in what periphery	4	Bolly, Crible, et al. 2014

Table 1: Overview of the annotation tiers specified by the protocol

In the first tier, the number of possible values is not pre-defined but rather corpusdriven, which means that all markers found in a particular text will be annotated, provided they fit the categorical definition. As mentioned before, this protocol is meant to fit the purpose of any research on DMs and is hence not limited to particular subgroups of items. However, there would be no consequences for the annotation if the number of possible values were restricted to a closed list. All other tiers have a specific number of values which is the result of data exploration and empirical testing of this protocol on a pilot corpus.

As one can see, domain and function can be assigned twice, when a particular DM appears to show two functions, either from the same functional domain (in which case the domain would still be assigned twice to keep the number of functions equal to the number of domains) or from two different ones. This option allows for not only ambiguous cases (which should be resolved as much as possible) but mostly for the not so rare cases of multifunctional markers. Simultaneous functions can be equally salient or ranked by their relative salience, but for operationalisation purposes such a distinction will not be made. In fact, it is not always easy nor relevant to determine which function primes over the other, and whether there is such priming at all: “no one function is necessarily predominant in a particular context” (Brinton 1996: 35).

An additional, automatic POS-tagging tier can be added if needed to save some time, although POS-tags are not always accurate when it comes to discourse markers because of their non-prototypical syntactic behavior (conjunctions tagged as adverbs, for instance).

Finally, although this set of tiers could suit many different research purposes, it can be interesting to name a few direct applications for each of them. The ideas listed below are not specific to our research project, although inspired by it.

POS information can be used to account for creativity and diversity of discourse markers across the genres and languages in focus, while offering an objective description of the controversial and interpretative category of linguistic elements that is the category of discourse markers.

TYPE DM can be tested for language-specific preferences regarding the use of connective elements over non-relational markers, and *vice versa*.

DOMAIN contributes to DM categorization and annotation in large corpora, by offering a filter into the many functions DMs can perform, and their balance across genres and languages.

FUNCTION offers a more fine-grained inspection of discourse markers description, to check whether certain functions show patterns of preferences with other variables (e.g. position). It can also be useful when wanting to see which markers perform the same functions in different languages.

POSITION macro can test the assumption that discourse markers mostly occur outside the dependency structure. Counter-examples can be examined in more scrutiny to understand under what conditions and with what consequences can markers be syntactically integrated.

POSITION micro provides a closer definition of the syntactic behavior of the markers, by indicating whether they appear before, within or after their host unit. It can be assumed that medial position will be more prone to interrupt the flow of speech, while initial markers will tend to be more clearly functional to the structure of discourse. This tier completes the previous one by relating the marker to both macro and micro-syntax, thus giving a more reliable account of its syntactic behavior (see section 3.7 for interesting cases).

POSITION turn is rather practical and offers to automatically extract all turn-initials and turn-finals, for any hypothesis-testing (matching position within the turn with functional domain, for example).

CO-OCC primarily serves for the extraction of complex, co-occurring markers, and can also disambiguate emphatic markers, which function is only dependent on the function of the item directly contiguous to it.

Further definitions and criteria for all possible values of these parameters will be given in section 3.

2.4 Main revisions from previous schemes

The present protocol offers improvements from other existing annotation schemes in the literature, as well as from previous versions of this one. In general, the priority was to make every decision explicit, theoretically motivated and as replicable as possible. The originality of this proposal lies in the merging of functional taxonomies for connectives on the one hand, and more interpersonal discourse markers on the other (see section 3.3 for further details on this distinction).

After testing version 7 of this annotation protocol on authentic data of bilingual interviews, a number of revisions were made to cope with problems and hesitations encountered in the process. In this section, the main modifications will be only briefly summarized, since all decisions will be further explained in the remainder of this report. The following changes are ranked hierarchically by their impact on the whole annotation process:

1. scope: more detailed and documented definition of the items under consideration in the scope of this protocol;
2. terminology: choice of the label “discourse markers” as the broad category that comprises relational discourse markers (sometimes referred to as connectives) and non-relational discourse markers, while keeping the same hyperonym to highlight the similarity of their functions;
3. possibility to annotate two domains and two functions for each marker instead of having to choose, sometimes arbitrarily, between two equally marked values;
4. splitting the annotation of position in three tiers that account for position within the macro-syntactic structure, the micro-structure of the host-unit and in relation to the whole turn of speech;
5. addition of a tier that accounts for the contiguous presence of another marker and the resulting sequence of combination;
6. addition and modification of certain functions for exhaustivity and efficiency;
7. specification and detailed definition of all values, with the addition of an unambiguous paraphrase for each possible value and authentic examples from the corpus;
8. addition of a section on semantic mapping and a guide on polysemous DMs to help resolve interpretation issues of highly ambiguous occurrences.

All these changes are hoped to improve the robustness and reproducibility of the annotation.

2.5 Conventions

Before turning to the criteria of all tiers and their values, a few precisions need to be made regarding the conventions specified by this annotation protocol. They only concern some practical information that are thought to facilitate comparability and technical treatment of data.

A first recommendation advises the analyst to reproduce a standardized orthographic form of the annotated item, in the first tier “DM”, so that articulation, diatopic or diastratic variation does not corrupt technical treatment of data. In this matter, transcriptions will not be trusted, in the sense that some transcribers may wrongly assign a (non-)standard pronunciation to the speaker, when the actual realization of the item shows a different phonetic output. We advocate to always listen to the sound file instead of relying on the transcription. However, even so, the process remains subjective, hence a bias towards standardized pronunciations unless the item sounds phonologically very different (e.g. *eh ben* instead of *eh bien*).

Also, certain DMs appear to always co-occur together, until they are fully grammaticalised and irrelevant to annotate separately. The limit between co-occurrence and fixation is subtle and mostly based on frequency criteria: the more often two items appear jointly, the more fixed their respective position becomes. Therefore, in a very limited number of cases, such “complex” DMs (i.e. fixed co-occurring DMs) will be annotated as one item. This convention concerns the following expressions: *mais bon*, *et puis*, *bon ben*, *eh ben*, *ou sinon*, and *then*. In this closed list, it is impossible to assign a function or meaning to the elements taken separately, which motivates their fusion.

Another convention is the writing of all tags in uppercase letters, for visibility reasons and because not all softwares recognize the difference with lowercase.

In case of overlapping speech and depending on how sound-text alignment was processed, it may happen that only one of two simultaneously occurring DMs is transcribed.

In this case, only the marker that belongs to the initiating turn (i.e. the new turn, not the one being interrupted or overlapped) will be annotated. This is motivated by technical reasons, in order to preserve the integrity of the text-sound correspondence. However, specific research interests might need to maintain the systematic integrity of one transcription over the other (for instance the interviewee vs. the interviewer).

Since this protocol only deals with the annotation of discourse markers, transcription conventions will not be commented. However, it is always recommended to have homogeneous transcription styles across all texts, even from different source corpora. Pauses can be helpful in the disambiguation of discourse marker functions and uses, so their notation in the transcript is desirable. Finally, sound-alignment, although not always necessary in discourse marker description, can be required for certain research questions, and is in some cases needed for disambiguation purposes. The present protocol was tested on a corpus that was not sound-aligned at the time of the annotation.

3 Annotation tiers and values

In this section, all tier values will be defined and explained in relation to previous works where they might have been used differently, and in relation to one another to explain their differences. When no reference is mentioned, the value has never been used or documented before, to our knowledge.

3.1 Tier 1: DM

This first tier stands for “discourse marker” and will be filled by the standardized orthographic transcription of the token under consideration. In this case, the term “discourse marker” is used as the umbrella term that covers all types of DMs (see section 3.3 for further details on the two types of DMs). Tier 1 will thus contain all occurrences of DMs in one text.

Starting from the definition proposed in section 2.1, we can list a number of prescriptive criteria which will hopefully improve the replicability and consistency of the identification phase. Items contained in this first tier will thus:

- function as procedural instructions for the interpretation of discourse, within one of the following four domains: ideational, rhetorical, sequential, interpersonal (see below section 3.4) ;
- be syntactically optional: their removal does not alter the grammaticality of the utterance;
- apply to an autonomous unit, both syntactically and semantically, i.e. there must be a finite or implicit predicate, which includes subclauses but excludes a number of components such as relative clauses, infinitive phrases, nominal phrases (except when these are acting as a-verbal predicates). This excludes in effect from our selection all intra-sentential conjunctions such as *cats and dogs* and prepositional phrases such as *because of*, *in order to*, *instead of* etc.;
- show a high degree of grammaticalization, hence fixed (for multi-word units), frequent in a given linguistic community (i.e. not idiosyncratic) and semantically bleached (non-compositional);
- be incompatible with membership of one of the categories mentioned in section 2.1, viz. fillers, interjections, response signals, epistemic parentheticals, general extenders, tag questions and editing terms, following the criteria and motivations explained above.

This definition (and its improvements from previous versions of the scheme) were tested by two coders in an independent annotation experiment (Crible and Zufferey 2015). Although some disagreements remain (due to individual biases and the inherent ambiguity of speech phenomena), this list of criteria should help coders strive towards an operational identification process. It is important to stress that the primary criterion is functional, and in this sense the selection is somewhat circular with the annotation of functions: whatever fits in the taxonomy (detailed below) will be considered a DM, provided they match the syntactic restrictions. In other words, function primes but syntax filters.

I would also like to note that several of these restrictions - while all consistent with the definition presented here - are rather specific to the author’s research questions, and may thus vary depending on the purpose of the annotation. In our view, this would only be a problem insofar as such decisions were not documented nor motivated in the annotation protocol, which is not the case with the present document.

3.2 Tier 2: POS

The set of POS tags used in this protocol is borrowed from the PDTB annotation guidelines in Santorini (1990), with the exception of interjections and prepositional phrases that will be developed more explicitly. The PDTB manual did not provide generic definitions. Examples are fictional and prototypical.

Coordinating conjunction - CC is a short, invariable unit that relates paratactically two units of usually the same class and function. This category is sufficiently defined and small to be operationalized by a closed class, for instance in French: *mais*, *ou*, *et*, *donc*, *or*, *ni*, *car* and their English equivalents (*but*, *or* etc.).

Adverb - RB is an invariable sentence specifier indicating manner or modality, for instance. *well, actually*

Verbal phrase - VP is a unit with a verbal node. *you know, I mean*

Subordinating conjunction - SC is an expression that introduces a subordinated clause. Original tag in Santorini (ibid.) was “IN” and grouped subordinating conjunctions with prepositional phrases, which are now two distinct categories in this protocol. This dichotomy was made on the basis that the two classes do not always correspond to syntactically and functionally similar items. *because, although*

Pronoun - WP can be an interrogative or indefinite pronoun. *FR quoi*

Adjective - JJ is a variable noun or pronoun specifier. *FR bon*

Noun phrase - NN is a short unit with a nominal node. *sort of, FR genre*

Prepositional phrase - PP is a short unit that begins with a preposition. Like “SC”, this tag did not exist in the original PDTB manual and was added here for clarification. *in fact, FR en fait*

Interjection - UH: Following Ameka (1992) and Norrick (2009), the label “interjections” (tagged UH) is reserved for “primary interjections”, i.e. desemanticized, not lexicalized vocal gestures such as *oh* or *euh*. “Secondary” interjections like *well* and *bon* will be coded as their original grammatical class (in this case, adverb and adjective respectively). Any item that does not fit in any of the above categories will be tagged “UH”. *oh, yeah*

This system hence assigns a POS tag to the whole DM unit, and not to each component, in the case of a multi-word unit. A similar approach is taken by Pitler and Nenkova (2009) who refer to this syntactic feature as “self category”:

Self Category: The highest node in the three which dominates the words in the connective but nothing else. For single word connectives, this might correspond to the POS tag of the word, however for multi-word connectives it will not. For example, the cue phrase *in addition* is parsed as (PP (IN In) (NP (NN addition))). While the POS tags of “in” and “addition” are preposition and noun, respectively, together the Self Category of the phrase is prepositional phrase (ibid.: 14).

In the case of complex DMs (see section 2.5 for their definition), only one POS-tag will be given following the major element composing the DM, for instance *et puis* will be coded as CC since the “head” of this complex DM is *et*.

For coordinating conjunctions, it is important to mention that only items connecting units above the clause will be considered DMs, thus excluding intra-sentential CC as in *cats and dogs*. Cleft constructions such as *c’est parce que tu viens que je serai là* will also be excluded since the DM is upgraded to a syntactic and propositional role in the utterance, hence no longer optional. However, it includes cases where the subject of the utterance is omitted (for instance because it is the same as in the previous utterance) and averbal or nominal sentences.

This system allows a POS-tagging method that does not depend on the function of the marker, but that focuses on its grammatical origin, thus avoiding the over-representation of adverbs (or adverbial uses) and accounting for linguistic creativity.

3.3 Tier 3: TYPE_DM

As mentioned before, our understanding of what falls within the DM category includes connecting devices that signal a discourse relation such as cause or contrast, as well as items functioning on other

semantic levels such as text-structuring, metadiscursive or interpersonal. This distinction is rarely tackled explicitly, and most authors recourse to different (and sometimes confusing) labels, but without any clear statement on what they imply. In fact, apart from Degand and Simon-Vandenberg (2011), it is fairly uncommon to see any mention of this distinction at all. They address this issue in terms of a scale between two extremes, “non-relational” and “strictly relational”, the former showing no linking function but rather (inter)subjective purposes such as *I think*, while the latter are “grammatical items in the traditional sense of the term”, i.e. conjunctive, connecting two elements (ibid.: 289).

Following these authors, DMs will be coded “relational” (RDMs) or “non-relational” (NRDMs) depending on the function they perform in context. NRDMs cover very different forms and functions, from interactive verbal expressions (*you know*, *tu vois*) to interjectional “punctuants” or punctuators (Vincent 1993) such as *well* or *bon ben* and other metadiscursive elements like *sort of* and *actually*⁷². RDMs on the other hand are restricted to signalling a two-place relation, either on the content level (e.g. a cause between two events), on the metadiscursive level (e.g. a reformulation of a previous statement) or on the textual level (e.g. a thematic shift from previous context).

For operationalization issues, we decided not to discriminate RDMs on the type of the related segments in the first stage of the annotation, since distinctions between single or complex units, textual or contextual are not always unequivocal. Moreover, although the presence or absence of a textual segment might have a cognitive impact, “connectivity is not limited to relations between neighbouring utterances” (Hansen 2006: 25). Therefore, items will be considered RDMs in the following three cases, provided they apply one of the coherence relations mentioned above:

- S1 - RDM - S2: S1 is a single textual unit, such as a clause;
- S1* - RDM - S2: S1* is a complex portion of several units, such as a thematic or informational unit;
- (S1) - RDM - S2: S1 is a contextual assumption, not textually expressed or unclear in the prior context.

This two-fold classification of DMs depending on their functioning can be found in previous works which address one end or the other, under different labels. Diewald (2013) reports two opposing views of DMs, “School 1” and “School 2”, where the former only includes markers which take scope over two explicit, textual units (as exemplified by (Fraser 1999, 2006)), while the latter includes relations between assumptions, a position which is supported by many authors (e.g. Hansen 2006; Rouchota 1996). RDMs are sometimes called “connectives” (e.g. Pons Bordería 2006; Rouchota 1996), “text-connecting marker” (Diewald 2013) or “text-relation marker” (Roulet 2006). As for NRDMs, it would seem that terminology is even more chaotic, since most uses of the label “discourse marker” (or sometimes “pragmatic marker” (Brinton 1996, 2008)) are unclear on whether they include RDMs or not. Fischer (2006b) refers to them as “discourse particles”, while this term also seems to include modal particles. Our choice to divide the category of DMs into “relational” and “non-relational” allows to avoid the problem of coining a new term to the two subclasses, while maintaining “DM” as the hyperonym, thus highlighting the similarity of their functions.

Additionally, Degand and Simon-Vandenberg (2011) also acknowledge that many DMs belong somewhere between the RDM-NRDM extremes and are thus difficult to situate, “though some have more salient connective functions than others” (ibid.: 289). In our experience of annotating authentic data, this is very often the case. Therefore, in this protocol, we will make a three-fold distinction, rather

⁷² In certain contexts of DM use, *actually* can be emphatic of previous elements in the utterance, but can also be a marker of counter-expectation or signal subjectivity (Aijmer 2002).

than a mere dichotomy, to account for these “in-between” cases. The resulting third possible value for this tier covers DMs that, in context, show simultaneously both type of functions. In practical terms, this applies, for instance, to a conjunction like English *so* which might function primarily as punctuating the flow of speech for planning purposes while maintaining a hint of conclusive or consequential meaning. In case of doubt, we advocate for a relational bias in the annotation of this parameter, when a basic relational meaning can still be perceived from the core semantism of the DM.

Here are the three possible values, their tag and definition for the tier TYPE_PM:

Relational - REL markers apply a discourse relation between a host unit and its context (either textually expressed or not).

- (1) I think we've not been protected here in Guildford, but it's certainly not been as extreme as perhaps in the North (bb_en011: 2297)

Non-relational - NREL markers do not explicitly signal a relationship but rather signal various metadiscursive functions related to word processing, interpersonal management, structuring and punctuating speech.

- (2) we're only, you know, less than an hour from the actual venues (bb_en 011: 1105)

Both - B applies to markers which function somewhere between the relational and nonrelational extremes, i.e. serving both as signals for discourse relations and other non-connecting functions.

- (3) c'était euh c'était la fête, une fête d'ailleurs qui s'est euh pérennisée (bb_fr004: 2255)

To make the annotation of this parameter as operational and reproducible as possible, one may use closed lists of functions that are either relational or not as a guideline.

Although this process would render the two parameters (TYPE_DM and FUNCTION) inter-dependent, it will certainly help the annotation process, especially if the closed lists are well designed and semantically motivated. For each possible function detailed in this protocol in section 3.5, we will indicate the prototypical TYPE value, based on the semantism of the function at stake. For instance, a causal DM is mostly relational, while a hedging DM is rather non-relational.

This parameter is potentially subject to coders' subjectivity and may lead to disagreeing annotations, but we believe in its valuable insight into the functions of a DM, for which it is a sort of filter. More details regarding the distinction of DMs (and their subtypes) with modal particles and other pragmatic markers can be found in Crible (forthcoming) or in Figure 3.

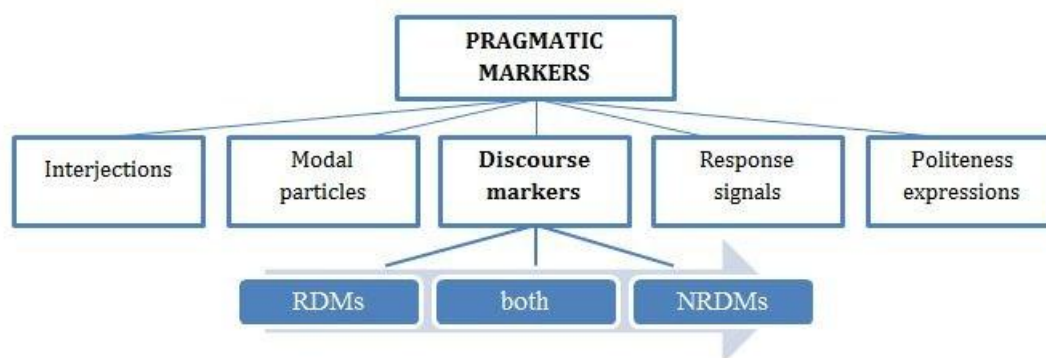


Figure 3: Taxonomy of pragmatic markers with types of discourse markers

3.4 Tiers 4 and 6: DOMAIN

Several authors have tried to come up with a categorizing system that would account for all possible functions conveyed by discourse markers (see Cuenca 2013 for a recent example). Some were actually designed for more global purposes, like Halliday and Hasan (1976), Sweetser (1990) or Redeker (1990). However, even more contemporary, DM-specific taxonomies all fail to a certain extent to meet the ideal balance between precision and application, in other words necessary and sufficient functional domains⁷³ that are distinct enough to avoid hesitations but still cover the whole range of values possibly conveyed by discourse markers.

This protocol suggests a fourfold taxonomy that borrows from existing threefold proposals. A similar system can be found, although without any operational criteria, in Haselow (2011), who identifies four equivalent domains. Each domain will be defined and compared to its counterparts from other taxonomies when necessary.

When a DM expresses two functions from the same domain, the domain will be assigned twice (DOMAIN_1 and DOMAIN_2), so that the number of given functions is equal to the number of domains, for post-treatment. These doubles can always be filtered automatically if need be.

Ideational - IDE domain is linked to states of affairs in the world, semantic relations between real events.

In other words, the relation between the two discourse objects exists independently in the real world. It includes for instance contrast or temporal ordering of events (see more examples in section 3.5.1). When a conflict with a rhetorical interpretation (see below) is possible, we suggest an ideational bias with the following criteria: the context which motivates the use of the marker must be textually expressed and concern real events, external and independent from discourse. This domain and its definition are inspired by Degand (1998).

Rhetorical - RHE domain is linked to the speaker's metadiscursive work on the ongoing speech. It consists of purely pragmatic, metadiscursive functions such as comment or emphasis. This domain also includes pragmatic equivalents of certain ideational relations, when the relation is applied between two discursive events rather than world-events (see section 3.5.2 for examples). These pragmatic relations apply to subjective claims, implicit assumptions or speech-acts. Unlike ideational relations, rhetorical functions cannot be reformulated without assigning mental states to one or both units: in this sense they can be distinguished from their ideational equivalent (see section 4), which do not necessarily require metarepresentations but only refer to events. The ideational bias mentioned above is meant to prevent the over-representation of rhetorical relations such as opposition or conclusion.

The rhetorical domain is borrowed from Gonzalez (2005), but also present in other works such as Haselow (2011: 3609) where it is called "metadiscursive". Otherwise it is either absent or grouped with interpersonal functions, as in Cuenca (2013) and the classical Halliday and Hasan (1976).

Sequential - SEQ domain is linked to the structuring of discourse segments, both at macro- and micro-level. This means that local management of smaller units (hesitation breaks, other types of filled pauses) will be included in this domain, along with more structural functions such as turn-taking or topic-shifting. Sequential functions explicitly signal the progressing steps of speech and thought. The sequential domain is close to its counterpart in Gonzalez (2005) and is also called "textual" in Halliday and Hasan (1976) and Cuenca (2013), although the latter includes

⁷³ The term "domain" thus refers to a category of functions and is sometimes referred to as "structure" (as in Gonzalez 2005).

reformulation, which belongs to the rhetorical domain according to this protocol (given its strong metadiscursive role).

Interpersonal - INT domain is linked to the interactive management of the exchange, in other words to the speaker-hearer relationship. Interpersonal functions have a phatic function to call for attention or to manifest understanding. This domain, or one similar, is mentioned in every functional taxonomy in the literature, for instance Brinton (2008).

3.5 Tiers 5 and 7: FUNCTION

This protocol provides a closed list of thirty functions that may be performed by discourse markers in speech. Many definitions are based on existing literature, but were often modified and explicitated here, with examples taken from the *DisFrEn* corpus. Most of the ideational and rhetorical functions were borrowed from the PDTB annotation manual, although they were sometimes specified and given a different tag. In the following, functions will be listed and defined, grouped by their domain, with proposals of paraphrasing, and their prototypical type.

Note 1: when quoting the PDTB, their use of “Arg1” and “Arg2” corresponds to our S1 - S2, i.e. the two (non)verbal units connected by the marker.

Note 2: In the examples taken from *DisFrEn*, the reference includes the text code and the position of the item in the transcript (a cell or “event” in Partitur Editor).

3.5.1 Ideational functions

Cause - CAU “the situations described in Arg1 and Arg2 are causally influenced and the two are not in a conditional relation” (Prasad et al. 2007: 28), when S1 and S2 are two real-world events, including future or hypothetical facts. Our tag corresponds to the PDTB’s subcategory of cause “reason”. Paraphrase: “This happened because...”. Type: RDM.

(4) they do struggle because sometimes it’s their first experience (bb_en 023: 803)

Consequence - CONS “the situation in Arg2 is the [logical] effect brought about by the situation described in Arg1” (ibid.: 29). Our tag corresponds to the PDTB’s subcategory of cause “result”. CONS also includes markers of purpose like *so that*, following the PDTB (ibid.: 39). However, it excludes under-specified additions and temporal sequences. Paraphrase: “As a consequence of that, this happened”. Type: RDM.

(5) I initially wanted to go into marketing or advertising but decided it wasn’t for me. So I did a three year course in marketing (bb_en023: 1157)

(6) Northeast said it would refile its request and still hopes for an expedited review by the FERC so that it could complete the purchase by next summer (PDTB corpus, 0013)

Temporal - TEMP “the situations described in the arguments are related temporally” (ibid.: 27). This tag includes both ordered and overlapping temporal events, i.e. synchronous and asynchronous in the terms of the PDTB. We suggest a temporal bias in case of conflict with under-specified consequence relations. Paraphrase: “After/before/during this, then ...”. Type: RDM.

(7) and after Theo was born then, did your husband have... (bb_en014: 296)

Contrast - CONT “Arg1 and Arg2 share a predicate or property and a difference is highlighted with respect to the values assigned to the shared property” (Prasad et al. 2007: 32), either as an opposite (PDTB’s subtype “juxtaposition”) or as a scalar difference (PDTB’s “opposition”). Contrast

differs from concession by explicitly referring to a verbally expressed property that is contrasted. Paraphrase: “X is this, whereas Y is that”. Type: RDM.

- (8) you can do this in a concrete sense and you can do it in a slightly more implicit sensed (bb_en025: 839)

Concession - CONC markers deny one or several clearly identified expectations explicitly related to the concessive segment. Concession can apply to both events and assumptions as long as the expectation derived from S2 is logical and verbally expressed. Paraphrase: “Although..., yet something else happened”. Type: RDM.

- (9) a place called Sutton which is actually a borough of London but it’s classed as Surry (bb_en023: 178)

Condition - COND “the situation in Arg2 is taken to be the condition and the situation in Arg1 is taken to be the consequence, i.e., the situation that holds when the condition is true” (ibid.: 29). It includes all possible subtypes identified by the PDTB group (present, past, unreal etc.). Paraphrase: “On this condition only...”. Type: RDM.

- (10) In addition, Black & Decker had said it would sell two other undisclosed Emhart operations if it received the right price. (PDTB corpus, 0807)

Exception - EXC “Arg2 specifies an exception to the generalization specified by Arg1” (ibid.: 36). In our view, the exception can be either in S1 or S2. The exception is a real content-object which is extracted from a category of content-objects, with no change in topic. Paraphrase: “with the exception of”, FR *à part ça*. Type: RDM.

- (11) juste l’accent peut-être, mais à part ça euh on utilise presque la même langue (bb_fr016: 204)

Alternative - ALT “The arguments are alternative situations, exclusive or not” (ibid.: 36). It includes PDTB’s subtype “chosen alternative” (typically marked by *instead*) and negative hypothesis (as in *otherwise*, *sinon*). The choices or the preference given by an alternative relation do not imply the speaker’s subjective appreciation of an expression that fits their intention better, unlike reformulative relations (see below 3.5.2), but merely reports competing facts. Paraphrase: “on the one hand... on the other” ; “instead”. Type: RDM.

- (12) it isn’t allowed to share in the continuing proceeds when the reruns are sold to local stations. Instead, ABC will have to sell off the rights for a one-time fee (PDTB corpus, 2451)

- (13) Obviously a lot of people will publish their poetry on the web either in their own spaces, whether it be MySpace or in Facebook, or whether they are members of groups and organisations which will actually publish online (bb_en025: 1612)

3.5.2 Rhetorical functions

Motivation - MOTIV corresponds to a pragmatic (epistemic or speech-act) cause. It applies to the subjective content of a claim or a speech-act. We suggest an ideational bias when a factual cause can be reconstructed. Paraphrase: “I say this because...”. Type: RDM.

- (14) and you were actually at Birmingham university, because I understand there are different universities in Birmingham? (bb_en023: 1842)

Conclusion - CCL corresponds to a pragmatic result, an epistemic or speech-act consequence. It includes summaries with conclusive value (PDTB's "generalization"), but excludes simple paraphrasing. Conclusion usually takes scope over a complex left context, while reformulation modifies a simple unit. Conclusion usually corresponds to an evaluation or a generalization, when the causal link between the two segments is under-specified other than by the speaker's appreciation. We suggest an ideational bias when a factual consequence can be reconstructed. Paraphrase: "We can now say that ...". Type: RDM.

(15) and it talks about different sorts of, well, settings in nature really, so it's lovely (bb_en023: 704)

Opposition - OPP corresponds to a pragmatic (epistemic or speech-act) contrast or concession, and includes counter-expectation as well. Both contrast and concession are grouped in the rhetorical domain since the former does not seem to be very frequent, based on a pilot corpus study, nor easily distinguishable from concession. Opposition differs from concession when the link between the expectation and S1 is not clear and not textually expressed. Paraphrase: "Although I said that, actually...". Type: RDM.

(16) it's quite small. It's very nice actually, it's a nice place to live (bb_en 023: 86)

Relevance - REL corresponds to a pragmatic condition, specifically when S1 and S2 are not causally related: "Arg1 holds true independently of Arg2", contrary to ideational condition (Prasad et al. 2007: 31). The condition is what makes the speech-act relevant to the particular context. Paraphrase: "I can say this only in the context of...". Type: RDM.

(17) If you are thirsty, there's beer in the fridge

Reformulation - REFOR is an equivalence between two simple units with a change in phrasing. It includes simple paraphrase ("equivalence" in the PDTB) and actual reformulation ("alternative"), sometimes too hard to distinguish. In case of a reformulation between two different contents, S2 is marked by the speaker as more appropriate (correct, relevant) than S1, which is cancelled. Paraphrase: "in other words" ; "I should rather say...". Type: RDM.

(18) you're getting more work ? I mean, is there an increasing need for translation? (bb_en016: 2331)

Approximation - HEDGE is a deliberate lack of precision, mitigating the speaker's assertion. It excludes hedging due to face-threatening contexts (see below). Paraphrase: "about", "not literally". Type: NRDM.

(19) I don't teach that sort of separately (bb_en023: 570)

Comment - COMM is a remark that is not directly related to the speech but is considered relevant for full understanding, in other words a digression or parenthesis. It may be considered a specific type of specification which is marked as an afterthought, less relevant or less important. Paraphrase: "by the way". Type: RDM.

(20) one of the things that I think is changing across all parks and we're certainly driving here is that we want to go back to those very early stage companies (bb_en024: 551)

Specification - SPE "applies when Arg2 describes the situation described in Arg1 in more detail" (Prasad et al. 2007: 34) and instantiates Arg1 with an example (PDTB's "instantiation"). The content of S2 must fall within the informational scope of S1. We restrict specification to the strict application of one of the three following paraphrases. Paraphrase: "Which is/are/does etc."; "for example"; "in particular". Type: RDM.

(21) I teach in a school called Devonshire primary school so it is a primary school which ranges the children range from... (bb_en023: 222)

(22) And that's at all levels. So, for example, I think, while it may be controversial, I think it's actually quite important for students... (bb_en 025: 1181)

Emphasis - EMP reinforces the propositional value of the utterance or of a neighbouring pragmatic function. This function mostly corresponds to cases of co-occurrences with another, semantically richer DM, or when it stresses another element. Emphasis must depend on another co-textual expression which it reinforces. Paraphrase: corresponds to prosodic stress or graphical underlining of usually the next item. Type: NRDM.

(23) but actually we also will buy expertise in from outside, emphatic of "but" (bb_en024: 959)

3.5.3 Sequential functions

Opening boundary - OPEN the item opens a new turn, in which case it indicates floor-taking, or a new sequence, within the same topic, namely an introduction to an enumeration or a narrative sequence, or possible others. Apart from turn-taking, it corresponds to any form of opening or engaging which is not covered by topicshift or any other sequential function. OPEN cannot be assigned as a double tag with another sequential function. Cuenca (2013) refers to this function as "start". Type: NRDM.

(24)... having the traditional weddind breakfast ? // So a variety of things. So there'll be things like ... (bb_en012: 437]

Closing boundary - CLOSE the item indicates the intention to close a list, a thematic unit or a turn. It must be in final or autonomous position. This function is borrowed from Cuenca (ibid.). Paraphrase: "This topic/ this turn is now closed". Type: NRDM.

(25) and the children and myself are both noticing that, so. (bb_en023: 367)

Resuming - RES The item signals the intention to link the upcoming segment to previous topic, to come back to the topic after a digression, a hesitation or a nonrelevant passage. Formal criteria include anaphora or reference to a previous topic which is taken up. This function is also borrowed from Cuenca (ibid.) who labels it "continuity". Paraphrase: "Back to our topic". Type: RDM.

(26) particular types of reactions and a particular sensory experience. So in relation to poetry of course, you can do this in a concrete sense (bb_en 025: 823)

Topic-shifting - TS The item signals a change of topic within or between turns. A distant connection to previous context can still remain, with a shift in focus. There must be no formal reference to previous context, as opposed to the continuity function. The new topic can be a subtopic of the previous one, but the latter should be definitely closed and not taken up in upcoming speech. This is mentioned in Cuenca (ibid.) as "topic change". Paraphrase: "Let's move on to...". Type: RDM.

(27) that's how practices work, is on a partnership of people of equals. And you want you asked me about the support staff. Traditionally... (bb_en 009: 184)

Quoting - QUO The item indicates the start of a reported speech segment. It must immediately precede other-attributed speech or a change in footing. This function was mentioned in Gonzalez (2005) as "opening quoted material". Paraphrase: corresponds to quotation marks. Type: NRDM.

(28) you would understand oh, the horse comes on (bb_en021: 1213)

Enumeration - ENU The item indicates a sequential ordering of discourse events, typically a list, structured by conventional items such as *firstly*. It also corresponds to textual deixis, i.e. referring to a specific discourse event. A similar function “list” was mentioned in the PDTB but with only one implicit example and a rather confusing definition: “when Arg1 and Arg2 are members of a list, defined in the prior discourse. ‘List’ does not require the situations specified in Arg1 and Arg2 to be directly related” (Prasad et al. 2007: 37). We may have restricted the scope of this function, but we believe our definition calls for less misinterpretation and hesitation. Paraphrase: “firstly..., secondly...”. Type: RDM.

(29) the site that we’re using here for surrey sports park. Firstly, it’s slightly off the main campus so that helps. Secondly... (bb_en 011)

Addition - ADD is the basic function of additive connectives when they “provide additional, discourse new information that is related to the situation described in Arg1” (ibid.: 37) when no other function applies. Since it refers to the continuation of on-going speech by adding new elements to the same topic, this function is classified in the sequential domain. It corresponds to a basic operation of addition within a topic or sequence, without any rhetorical or other value of specification or conclusion, for instance. Paraphrase: corresponds to the logical operation « + ». Type: RDM.

(30) it’s a very play based curriculum, and they do pick up the English language very quickly (bb_en023: 861)

Punctuation - PUNCT The item signals the intention to hold the floor while planning the upcoming speech, or for any other reason not mentioned by the other sequential functions. A marker will be coded as punctuating after elimination of all other sequential possibilities. Paraphrase: corresponds to typographical commas. Type: NRDM.

(31) ... a little bit more off the beaten track are, I don’t know, are quite special. (bb_en019: 2025)

3.5.4 Interpersonal functions

Monitoring - MONI checks for understanding and attention, in the form of an explicit address to the interlocutor. This is the more frequent and default function of the interpersonal domain. It includes markers of common ground. Paraphrase: “You know what I mean”. Type: NRDM.

(32) and then it’ll turn into, you know, more practical English really (bb_en 023: 921)

Face-saving - FACE expresses deference, politeness and prevents face-threats, and is therefore often found in face-threatening contexts. Paraphrase: “This is a delicate topic”. Type: NRDM.

(33) I come from a background where, you know, now I guess my family you would say is middle class (bb_en025: 2603)

Disagreeing - DISAGR expresses a disagreeing response. It differs from opposition because disagreement needs to be in response to another speaker’s turn. This function will not be coded when it is expressed by a response signal like “no”, since those are excluded from the definition of DMs (see section 2.1). Paraphrase: “No”. Type: NRDM.

(34) so you are the Cantona equivalent ? // Well, I’m not quite as great as him (bb_en020: 2665)

Agreeing - AGR expresses understanding. This function will not be coded when it is expressed by a response signal like “yes” with no additional function, since those are excluded from the definition

of DMs (see section 2.1). It includes both meanings of French *d'accord*: “I agree” and “I understand”, and is thus not necessarily in response to a question. Paraphrase: “I understand”. Type: NRDM.

- (35) So the regeneration isn't just about building places and buildings, it's also about building green parks and looking towards a more environmentally friendly future as well? // absolutely, yes. [...] by 2015 so yeah, it has very much a view of the environment (bb_en022: 1709)

Elliptical - ELL vague-category markers, indicate the inclusion of other members of a previous category without naming them. Corresponds to the category of general extenders. Usually takes the form of a conjunction followed by a routinized expression including a pronoun. Paraphrase: “and things like that”. Type: NRDM.

- (36) There was there was a lot of trade, I think a lot of spices and tobacco and things like that so obviously... (bb_en019: 1184)

3.6 POSITION_macro

Syntactic position description is challenging in many ways, two among them being the multiplicity of available theories and the non-canonical structure of spoken language. In my endeavour to assign a position to DMs in natural occurrences of speech, I will rely on the well-known framework of Dependency Grammar as originated by Tesnière (1959) and still fairly used today, although with minor terminological adjustments. I have thus chosen a model that provides the possible values for the macro-syntactic position of DMs in relation to a representation of the host unit as an utterance, i.e. a semantically complete proposition which comprises one or several clauses and their related peripheral elements. A clause is understood as a predicate, its arguments and adjuncts (respectively inside and outside the valency frame of the predicate). A clause can be independent (main clauses) or not (subordinate clauses), and all the syntactically-dependent clauses form an utterance.

A third challenge to this syntactic feature is that most DMs do not occur within these well-defined and consensual slots (predicate, arguments, adjuncts), but rather mostly outside of them. In this respect, they can be associated to “headers” and “tails” (Carter and McCarthy 2006), respectively left- and right-detached elements such as dislocations. This concerns many cases of coordinating conjunctions, adverbs and particles, while subordinating conjunctions will prototypically be described as an integrated, yet peripheral first element in a subclause.

In general, position of the DMs is assigned according to their scope: S2 for RDMs, closest and smallest clause available for NRDMs. At the macro-syntactic level, the annotation process will: (1) identify the scope of the DM, i.e. what unit it is attached to and (2) analyze the status of this unit in relation to the predicate of the main clause. Then, the item will be assigned one of five values, following the terminology in Lindström (2001) and the recommendations of the MDMA Research Group (Bolly, Crible, et al. 2014)⁷⁴. Lindström's system has been implemented by additional values (independent and interrupted) to better cope with the complexity of spoken data. The following lines will describe each possible value for the macro-syntactic position of a DM.

Pre-front field - PRE The marker is at the left periphery of the main verb and does not belong to the macro-syntactic structure of the utterance, i.e. outside the dependency structure (the clefting test cannot be successful). There must be no basic function words before the marker. Auer (1996)

⁷⁴ MDMA (Model for Discourse Marker Annotation) is a research group dedicated to the description of DMs as clusters of features, in the perspective of semi-automatic annotation.

describes the pre-front field as a position that “projects something else to come, but does not oblige the speaker to subscribe to one particular syntactic project at a time where sh/he may still be in the phase of planning” (ibid.: 313).

(37) And in fact even though the visual verbal intersection... (bb_en025: 558)

Initial field - LEFT The marker is at the left periphery of the main verb and is either syntactically integrated in the dependency structure as an adjunct itself, or between adjuncts that are included in governed units. Subordinating conjunctions are considered integrated.

(38) The new one is actually now out Canary Wharf way and although I haven’t done it it’s been on my list of things to do (bb_en019: 2673)

Middle field - MID The marker is within the core clausal structure, between two predicative elements, i.e. usually within the finite verb construction.

(39) it’s actually set by the government (bb_en023: 408)

End field - RIGHT The marker is at the right periphery of the nodal verb and is syntactically integrated within the dependency structure. This position is the right equivalent of INI.

(40) they call in our services because they need some professional help (bb_en 012: 239)

Post-field - POST The marker is at the right periphery but does not belong to the syntactic structure of the utterance. There must be no basic function words after the marker. Cases of interruptions where the marker happens to be the last uttered word will not be coded as POST but with a special tag (see below), to allow easy extraction and thus preventing mixing different phenomena in the same category.

(41) I think the negotiation takes place with the couple actually (bb_en 012: 940)

Independent - IND The marker is the only item of the unit, either as a whole turn or as a syntactically and prosodically detached element. This value was suggested by the MDMA annotation project (Bolly, Crible, et al. 2014).

(42) Oui. Bon. (bb_fr013: 26)

Interrupted - INT The position of the marker is unclear due to incompleteness and interruption. This tag will be given to a DM whenever the host unit is interrupted.

3.7 POSITION_micro

This second annotation of DMs position is much more intuition-based and simply relates whether the item is initial (which includes interruptions), medial (within function-words) or final (not followed by anything from the same speaker). This basic system takes into consideration the position of the marker within its minimal syntactic unit, starting from subordinated clauses and larger. We believe that this annotation layer provides a useful information that completes the previous positional information. It will allow for interesting cases where, for instance, a subordinating conjunction is macro-syntactically at the right periphery of the governing verb, but initial within its own subclause (see example 43). The structure of this variable is very similar to the macro-position above. Independent and interrupted (IND, INT) values match exactly the definition of their correspondents in the macro-position parameter.

Initial - ini initial position, the most leftward positional slot of the micro-syntactic unit that contains the marker (from the subclause up). The marker can be strictly or quasi initial. All RDMs generally apply to the segment they introduce, since they mostly introduce further speech (Schiffrin 1987; Schourup 1999), and will hence be coded initial.

(43) it's good for us because it puts us into a marketplace ... (bb_en011: 1196)

Medial - med medial position, integrated in the micro-syntax and preceded and/or followed by an element of the dependency structure. There must be non-optional words before and/or after the marker.

(44) people wanting to sort of be careful with budget (bb_en012: 597)

Final - fin final position, outside the dependency structure of the verb (micro-syntactic node, includes verbs in subclauses), at the right from the non-finite verb (if present). This position includes strictly final and quasi final.

(45) and are learning their trade like apprentices, if you will (bb_en009: 112)

Independent - ind the marker is the only item of the unit, either as a whole turn or as a syntactically and prosodically detached element. Here and for the next value, the same definition and example as in the macro-syntactic position apply.

Interrupted - int the position of the marker is unclear due to incompleteness and interruption.

3.8 POSITION_turn

The last variable of position concerns the turn of speech, and the values are fairly straightforward since it only looks at the exchange structure and whether the speaker takes the floor, holds it and ceases it. Turn breaks (represented either by a change in the transcription tier or by a symbol) will be the only decisive criterion. This parameter was suggested by the MDMA project (Bolly, Crible, et al. 2014).

Turn-initial - TI the marker is the first element of the speaker's turn. This tag is restricted to the very first position in the turn, with nothing being said by the same speaker in the same turn.

Turn-medial - TM the marker is in any other position within the speaker's turn, when it is neither of the other three values.

Turn-final - TF the marker is the very last element of the speaker's turn, either by choice or by interruption.

Turn - TT the marker constitutes the whole turn. This includes cases of co-occurrences or repetitions of markers.

3.9 CO_OCC

Finally, the values for this tier account for the immediately contiguous presence of another DM. The only criterion required here is the definition of what counts as a DM, which is, for coherence, the same definition as was previously presented in this protocol (cf. section 2.1). Therefore, a DM co-occurring with a filled pause such as *uhm* will not be considered as two co-occurring DMs, since filled pauses are excluded from our definition of DMs. In case of a co-occurrence, the periphery where the other DM appears in will be noted (left, right or both), following the MDMA model (Bolly, Crible, et al. 2014). Additionally, the actual combination of items, for instance *so actually*, will also appear in the annotation, thus giving in the output all possible combinations of DMs for further analysis. However, the sequence of combined items will only appear once, under the first DM of the sequence (so only for the "Yright" value), and not for all elements of the annotation, so that frequencies remain correct.

Yleft the annotated item is preceded by at least one other discourse marker at the left.

(46) but you know, you cut me half and it's red (bb_en011: 2482)

Yright: sequence the annotated item is followed by at least one other discourse marker at the right. In the following example, the annotated value would be “Yright: but you know”.

(47) but you know, you cut me half and it's red (bb_en011: 2482)

Ylr the annotated item is followed by at least one other discourse marker at either side, both left and right.

(48) but yeah so backtracking, going back I, walking along the north... (bb_en 019: 2727)

No - NO the annotated item does not directly co-occur with any contiguous discourse marker.

(49) divide between the north and the south. And this is... (bb_en 011: 2069)

4 Mapping of pragmatic and non-pragmatic functions

An interesting feature of the functional taxonomy proposed in this protocol is that it integrates the domain (IDE, RHE, SEQ or INT) within the function itself, by making these two parameters interdependent in the annotation. This is particularly relevant for discourse relations which can have both a semantic (ideational) and a pragmatic (rhetorical) use. Sweetser (1990) refers to this phenomenon as “pragmatic ambiguity” and describes it at length:

We use the same vocabulary in many cases to express relationships in the speech act and epistemic (reasoning) worlds that we use to express parallel relationships in the content domain (the “real-world” events and entities, sometimes including speech and thought, which form the content of speech and thought (ibid.: 10).

In fact, each ideational relation has a pragmatic equivalent, although the mapping only partially covers the semantics of exception (EXC) and alternative (ALT). For instance, an ideational cause will receive a different tag (CAU) than a rhetorical cause (MOTIV). One possible drawback of this design is that coding the domain was found to lead to numerous hesitations and disagreement between coders, especially when the choice is between ideational and rhetorical relations (see Zufferey and Degand in press). However, in my view, this distinction is too crucial in the interpretation of DMs to be kept as an optional sublevel or even removed.

Therefore, in this section, I will point to the semantic mapping between pragmatic and non-pragmatic functions (or pragmatic and “more” pragmatic functions, as I will explain below). For each pair of functions, I will try to provide useful distinctions and authentic examples. It is important to note that the limits between semantic and pragmatic interpretation can sometimes be rather thin, as well as flexible depending on the theoretical or other biases of the coder. In a recent annotation experiment (Crible and Zufferey 2015), we found that the issue of determining the moment when a discourse relation switches from a factual account to a discursive event is one the most problematic feature to annotate, especially in languages such as English and French where this is not grammaticalized as a different marker, and that it stems from individual biases and experience: for instance in speech, the boundary between semantics and pragmatics might be a little “higher”, that is restricted to purely metadiscursive uses, and not include all subjective facts or assumptions, as we would tend to in writing, where speech-act and intersubjective functions are quite rare.

4.1 Cause

CAUSE (ideational) and MOTIVATION (rhetorical) differ in both the source and target of the causality:

- logical cause between two facts:

“they call in our services because they need some professional help” (bb_en 012)

“il a fallu arrêter nos études parce que nos parents sont partis à la guerre” (bb_fr014, *we had to quit school because our parents left to war*) ;

- pragmatic cause between at least one mental state or speech-act:

“in Wapping actually it’s unfortunate, [I say/think this] because there is a tube in Wapping which at the moment is shut” (bb_en 019)

“alors la commune est assez importante puisque le nombre d’habitants est de cinq cent soixante quatre” (bb_fr021, *[we can say that] the town is rather big since the number of inhabitants is 564*)

Hence the difference is not whether the cause is supposedly discourse-given or known by the hearer, but how the two units are related logically and semantically. With motivation, the referential content of S1 is not logically inferred from S2, not until we add a mental state or a speech-act in the resulting metarepresentation.

4.2 Consequence

CONSEQUENCE (ideational) and CONCLUSION (rhetorical) draw on the same pattern, the latter being more of an evaluation, usually applying to a complex S1, while the former is restricted to clearly explicit consequence between events.

- logical consequence between two facts:

“we also wanted to be a facility which had maximum access, so we’ve built it with maximum disability access as well” (bb_en 011)

“par rapport à la France, un pays qui est beaucoup plus pauvre, donc tout ce qu’ on achète là-bas est beaucoup moins cher” (bb_fr018, *compared to France, a country which is much poorer, so everything you buy there is a lot less expensive*)

- evaluative, subjective conclusion or speech-act consequence:

“we’re specifically talking about wedding planning today so [this is why I say that] I also organise a lot of weddings” (bb_en 012)

“mais c’est beaucoup moins cher que euh en France. donc c’est sûr que ça fait un grand décalage, hein” (bb_fr018, *but it’s a lot cheaper than uh in France. so obviously [we can say that] there are a lot of differences, right*)

4.3 Temporal

The difference between TEMPORAL (ideational) and ENUMERATION (sequential) is slightly different, first because it involves the sequential domain as pragmatic equivalent, instead of the rhetorical one, and also because the related units can be two facts. It is the nature of the temporal relation that is different. The chronological ordering concerns in one case external events, with usually a before-after situation, and in the other members of a discursive list, an enumeration. The latter does not necessarily involves a comparison between two moments but can also be a mere pointer to a specific moment in discursive time.

- real-time chronology between events:

“has that changed since you started working in Birmingham?” (bb_en 022)

“dès qu’on arrive on a tendance à tout dépenser” (bb_fr018, *as soon as you get there you tend to spend all your money*)

- discursive chronology, ordering of textual segments, regardless of the content of the units:

“What kinds of things do you write about? Is it, first of all in terms of genre, is it poetry or ...” (bb_en 025)

“Je suis venue en France, premièrement euh parce que j’aime bien la langue française” (bb_fr016, *I came to France, firstly uh because I like French language*)

4.3 Adversative

As mentioned earlier, CONTRAST and CONCESSION are both subsumed into their unique pragmatic equivalent OPPOSITION. This might be the most difficult distinction of all possible values, since even ideational concession involves a counter-expectation or a mental state. For this reason, the boundary will recourse to formal criteria, such as textually expressed expectation (in S2) on which to draw the logical inference, for the ideational domain.

- logical counter-expectation inferred from verbally expressed cues in S2:

“as someone who comes from Northern Ireland, even though I left at a relatively young age, I have carried with me...” (bb_en 025)

“nos clubs sont en bonne santé mais nous ne pouvons pas faire venir chez nous des stars parce que...” (bb_fr003, *our clubs are healthy but we can't afford to hire stars because...*)

- unclear or distant adversative relation between assumptions or speech-acts:

“I’m not quite sure at the moment what it is but there is talk of something coming in every three years” (bb_en 023)

“puisque nous avons, je l’ai pas dit dans la présentation mais euh un euh un sous-sol tout à fait intéressant” (bb_fr019, *since we have, I didn't mention it in the presentation but uh a a basement absolutely interesting*)

4.4 Condition

While ideational CONDITION impacts the truth conditions of both units, rhetorical RELEVANCE merely states a context where S1 is pragmatically relevant, i.e. a context that motivates the speech-act in S1.

- truth-conditional, logical condition between events:

“we just put them in a room, and if they like each other, they’ll work together” (bb_en 024)

“si nous laissons faire, eh bien, dans dix quinze ans il n’y aura plus rien du tout” (bb_fr009, *if we let things go, well, in ten to fifteen years there won't be anything left at all*)

- contextualization that motivates the existence or relevance of S1:

“it wasn’t something I particularly wanted to get into, [I can say this] if I’m really honest” (bb_en016)

“et quelles sont les autres activités euh industrielles, si on peut revenir euh là-dessus” (bb_en017, *and what are the other uh industrial activities, [I can ask this] if we may come back uh to this*)

4.5 Exception

In the case of EXCEPTION, the mapping of domains only concerns the use of certain DMs, *viz.* French *sinon*, *à part ça*, *autrement*, and less so for *apart from that* and some contexts of *otherwise*, which can be interpreted as sequential TOPIC-SHIFT. In these cases, a sequential reading would imply cancelling the previous topic to bring forward a new one, much like an ideational context where a content is extracted from a category of objects. This mapping is thus partial and only motivated by the polysemy of the specific DMs mentioned here, which usually lead to ambiguity. I propose to use a rather strict criteria for the ideational interpretation, which would require the explicit mention of an overarching category from which the exception is extracted, whereas topic-shift may also relate two distantly connected subject-matters, but without any meaning of extraction from a category.

- exception to a category

“there are probably only three counties which are proud as their county, in my opinion, which would be Lancashire, Yorkshire and Cornwall, and apart from that perhaps not as proud as we are” (bb_en 011)

“il suffit d’avoir tous les papiers qu’ils demandent et il faut passer aussi l’entretien et il y a les euh les gendarmes qui viennent chez vous pour voir comment vous vivez, et ils vous posent des questions, mais à part ça euh le plus difficile c’est d’attendre” (bb_fr016, *you just need to have all the papers they ask and you also need to pass an interview and policemen come to your house to see how you live, and they ask you questions, but apart from that the hardest is the waiting*)

- new topic from cancelling of the previous one

“les appréhender, les solutionner moi aussi, hein ? euh et puis autrement, je fais des petits sujets euh un petit peu d’après la réalité” (bb_fr020, *apprehend, solve them myself, you know ? uh and then apart from that, I do little subjects euh sort of based on reality*)

4.6 Alternative

Between ALTERNATIVE and REFORMULATION, the mapping is also partial and concerns only the “inclusive” and “chosen alternative” meanings of the ideational domain, which can be marked by the same items as a rhetorical reformulation. The difficulty arises when a reformulation is motivated by the inaccuracy of the content, and not just the phrasing of S1, in which case a S2 is perceived as necessary by the speaker. For inclusive alternative (several choices presented as equally possible), the difference with reformulation lies in the fact that the latter implies a preference (be it on form or content). For chosen alternative (S2 being presented as the preferred choice), the difference with reformulation concerns the agent of the choice: an ideational interpretation will assign the preference to a real-world (acting or thinking) agent, who need not be the speaker himself, while a rhetorical reading would give this choice to the abstract speaker, as a speech-act preference. In any case, this distinction might be unclear and highly contextual.

- inclusive alternative between two equally possible facts

“something that can actually be taught and assessed, or can be learnt by a student ?” (bb_en 025)

- exclusive alternative opposing two contents that are incompatible with each other

“I’ll either do those myself or I’ll involve a barrister” (bb_en 009) “vous préférez maintenant habiter en France ou vous avez un petit peu la nostalgie de Madagascar ?” (bb_fr016, *do you prefer now to live in France or do you miss Madagascar a bit?*)

- chosen alternative expressing an agent's preference

“it isn't allowed to share in the continuing proceeds when the reruns are sold to local stations. Instead, ABC will have to sell off the rights ...” (PDTB: 2451)

- reformulation affecting content and/or form of a speech-act

“that you've got all this or Birgmingham's got all this regeneration going on” (bb_en 022)

“qui se sont expatriés pour apprendre les cultures, enfin apprendre la culture de la vigne” (bb_fr005, *who went abroad to learn the cultures, well learn the wineculture*)

4.7 Speaker- vs. hearer-oriented punctuation

This final mapping concerns PUNCTUATION (sequential) and MONITORING (interpersonal), and again only applies to certain uses of the DMs at stake. The equivalence between the two functions is less obvious than in the previous cases, especially since both functions are highly pragmatic, one being “more” subjective or intersubjective than the other. The idea is that punctuation is used by the speaker for various reasons including speech planning, and is thus highly speaker-oriented, while monitoring may also be used as speech-planning but in a hearer-oriented way, usually by an expression with a reference to the interlocutor (*you know*), an interjection (French *hein*) or in a final position (French *quoi*), which has been described elsewhere as prototypically intersubjective (Degand 2014). Therefore, these two criteria (formal reference to the hearer and/or final position) will be used to distinguish the two functions.

- speaker-oriented punctuation

“and then the science park, well, it can be for life of the company” (bb_en 024)

“alors le service, ben, c'est quand un opticien euh souhaite changer une paire de branches” (bb_fr017, *So service, well, it's when an optician wants to change a pair of spectacle arms*)

- hearer-oriented monitoring

“walls and ceilings and, you know, trying to finish before you go outside the lines” (bb_en 016)

“le marché de Louhans, hein, qui est un marché de volailles” (bb_fr004, *the market of Louhans, right, which is a poultry market*)

5 Guide to frequent polysemous discourse markers

Apart from the distinction of closely related functions in the form of (non-)pragmatic pairs, ambiguity can also arise when dealing with high-frequency, under-specified DMs, especially in speech. In an annotation experiment (Crible and Zufferey 2015), we found that the most frequent conjunctions lead to the highest number of disagreements in interrater agreement analysis. This may be the result of their frequency of use which makes them more accessible, almost automatic to retrieve from the part of the speaker. However in return, these uses seem to lack the semantic richness that allows for unambiguous interpretation in context, since their meaning is quite bleached and flexible.

This difficulty is hard to overcome, especially in speech, given the inherent ambiguous nature of natural communication and the limitations faced by the analyst in the process of off-line annotation (for instance lack of contextual cues such as physical setting or gestures). Although prosodic information can help orient our interpretation, it remains perceptive, hence subjective and not systematic. Therefore, I propose in this section to illustrate the possible values of highly polysemous DMs in English, to guide the analysts in their endeavour by providing criteria and examples, as well as the extensive list of

possible values. The following lists are corpus-driven, since they are based on observations from a pilot study on a corpus of interviews which was annotated according to a similar version of this protocol. For practical reasons, I will only focus on four DMs in English: *and*, *so*, *but* and *well*.

Function	Criteria	Example
Consequence	logical effect brought about by the situation in S1	“he left school at the age of fourteen and was an apprentice baker so he didn’t have much formal education” (bb_en 025)
Conclusion	evaluation, summary, generalization	“how that can be manipulated in order to try to achieve those effects. So I think that’s absolutely crucial” (bb_en 025)
Specification	more detail, example, or particularization	“mainly patents and kind of related legal work, so oppositions and all that kind of stuff” (bb_en 016)
Topic-resuming	come back to previous topic after a digression	(question about social impact of wedding) (several utterances about modern weddings) “...a cultural experience at a wedding. So yes, I think it’s social impact” (bb_en 012)
Reformulation	paraphrase or actual reformulation	“our deputy head every term so three times a year” (bb_en 023)
Closing boundary	intention to close a list, a topic or a turn	“the children and myself are both noticing that, so.” (bb_en 023)

Table 3: Summary of the different possible values for *so*

Function	Criteria	Example
Addition	simple addition of information within the same topic	"I'm Fiona Doloughan and I'm a lecturer" (bb_en 025)
Specification	more detail, example, or particularization	"language as being multimodal and particularly with poetry, you've got rhythms" (bb_en 025)
Consequence	logical effect brought about by the situation in S1	"that course was actually on short fiction and I spent quite a lot of time working on short fiction" (bb_en 025)
Temporal	chronological ordering of events	"And then I put that away in a drawer and I have left that since that time" (bb_en 025)
Contrast	the segments share a property which is contrasted	"you can do this in a concrete sense you can do it in a slightly more implicit sense" (bb_en 025)
Topic-shift	change of topic, possible distant connection with previous topic	"I would teach my class literacy, numeracy, history, and so on. / Ok. And what are you doing with your class at the moment?" (bb_en 025)
Opening boundary	engage a new turn or sequence, within the same topic	"And we'd like to talk to you specifically about..." (bb_en 025)

Table 2: Summary of the different possible values for *and*

Function	Criteria	Example
Punctuation	holds the floor, speaker-oriented	"different sorts of, well, settings in nature really" (bb_en023)
Reformulation	paraphrase or actual reformulation	"how language itself can actually trigger a different, well, a variety of sensory experiences"
Opening boundary	engage a new turn or sequence, within the same topic	"has that changed since you started working in Birmingham ? / Well, the perception of Birmingham is of engineering" (bb_en 022)
Disagreeing	expresses a disagreeing response (prosodic interpretation)	"there's always going to be work for a patent translator? / Well, theoretically yeah" (bb_en 016)

Table 5: Summary of the different possible values for *well*

Function	Criteria	Example
Contrast	the segments share a property which is contrasted	“it’s not a legal obligation but it is a moral obligation” (bb_en021)
Concession	denial of one or several clearly identified expectations explicitly related to the concessive segment	“people don’t do it so often but is equally quite interesting” (bb_en 019)
Opposition	unclear contrast or opposition, not verbally expressed	“we have a shooting script. But also you have time to actually blow up the image and try and work out what’s going on” (bb_en 021)
Topic-resuming	come back to previous topic after a digression	“it actually deteriorated, I suppose, in terms of a pattern. But yeah, life is turned upside down” (bb_en 014)
Topic-shifting	change of topic, possible distant connection with previous topic	“it’s a very brutal landscape a lot of it as well so. But my accent is very specific to two villages because it’s on the Yorkshire Lancashire borders” (bb_en 021)
Closing boundary	intention to close a list, a topic or a turn	“I don’t know much about him, but. / I suppose it makes sense if...” (bb_en 019)

Table 4: Summary of the different possible values for *but*

These tables are not necessarily exhaustive since genre-specific functions (or idiosyncratic uses) could emerge from other corpora. The examples are prototypical, but do not exclude the possibility of slight semantic variation. Again, I would like to stress the flexible nature of pragmatics and the necessary subjectivity that needs to be involved in the annotation process. This section merely offers a guide, with no further prescriptive agenda.

References in Appendix 1

- Aijmer, K. 1997. "I think - an English modal particle". In: *Modality in Germanic languages. Historical and comparative perspectives*. Ed. by T. Swan and O. Westvik. Berlin: Mouton de Gruyter, 1–47.
- Aijmer, K. 2002. *English discourse particles: evidence from a corpus*. Philadelphia: John Benjamins.
- Aijmer, K. and A.-M. Simon-Vandenberghe, eds. 2006. *Pragmatic markers in contrast*. Amsterdam: Elsevier.
- Ameka, F. 1992. "Interjections: the universal yet neglected part of speech". In: *Journal of Pragmatics* 18, 101–118.
- Auer, P. 1996. "The pre-front field in spoken German and its relevance as a grammatical position". In: *Pragmatics* 6.3, 223–259.
- Avanzi, M. et al. 2010. "C-PROM. Un corpus de français parlé annoté pour l'étude des proéminences". In: *Actes des 23èmes journées d'étude sur la parole* (May 25-28, Mons, Belgium).
- Bazzanella, C. et al. 2007. "Italian allora, French alors: functions, convergences and divergences". In: *Catalan Journal of Linguistics* 6, 9–30.
- Bolly, C. and L. Crible. 2015. "From context to functions and back again: Disambiguating pragmatic uses of discourse markers". In: *International Pragmatics Association (IPrA) Conference*, July 26–31, Antwerp, Belgium.
- Bolly, C., L. Crible, et al. 2014. "Towards a Model for Discourse Marker Annotation in spoken French: From potential to feature-based discourse markers." In: *International workshop Pragmatic Markers, Discourse Markers and Modal Particles: What do we know and where do we go from here?* October 1–17, Como, Italy.
- Brinton, L. 1996. *Pragmatic markers in English. Grammaticalization and discourse functions*. New York: Mouton de Gruyter.
- Brinton, L. 2008. *The comment clause in English: syntactic origins and pragmatic development*. Cambridge: CUP.
- Brogniaux, S. et al. 2012. "Train&Align: A new online tool for automatic phonetic alignment." In: *IEE Spoken Language Technology Workshop (SLT)*, 416–421.
- Buysse, L. 2014. "'We went to the restroom or something'. General extenders and stuff in the speech of Dutch learners of English". In: *The Yearbook of Corpus Linguistics and Pragmatics 2014: New Empirical and Theoretical Paradigms*. Ed. by J. Romero-Trillo. Springer, 213–237.
- Carter, R. and M. McCarthy. 2006. *Cambridge Grammar of English*. Cambridge: CUP.
- Cheshire, J. 2007. "Discourse variation, grammaticalisation and stuff like that". In: *Journal of sociolinguistics* 11.2, 155–193.
- Crible, L. and L. Degand. 2015. "Functions and syntax of discourse connectives across languages and genres: Towards a multilingual annotation scheme". In: *International Pragmatics Association (IPrA) Conference*, July 26–31, Antwerp, Belgium.
- Crible, L. 2014. "DisFrEn: Reaching bilingual comparability across multiple situational features. A challenge for corpus design". In: *IVACS*, June 18–21, Newcastle, UK.
- Crible, L. Forthcoming. Towards an operational category of discourse markers: A definition and its model.

- Crible, L., A. Dumont, et al. 2015. Annotation des marqueurs de fluence et disfluence dans des corpus multilingues et multimodaux, natifs et non natifs v.1. Tech. rep. Université Catholique de Louvain, Université de Namur.
- Crible, L. and S. Zufferey. 2015. "Using a unified taxonomy to annotate discourse markers in speech and writing". In: *Proceedings of the Joint ACL-ISO Workshop on Semantic Annotation*.
- Cuenca, M. J. 2008. "Pragmatic markers in contrast: the case of well". In: *Journal of Pragmatics* 40, 1373–1391.
- Cuenca, M. J. 2013. "The fuzzy boundaries between discourse marking and modal marking". In: *Discourse markers and modal particles. Categorization and description*. Ed. by L. Degand, B. Cornillie, and P. Pietrandrea. Amsterdam: John Benjamins, 191–216.
- Degand, L. and A.-M. Simon-Vandenberghe. 2011. "Grammaticalization and (inter-subjectification of discourse markers)". In: *Linguistics* 49, 287–294.
- Degand, L. 1998. "On classifying connectives and coherence relations". In: *Discourse relations and discourse markers. Proceedings of the workshop COLING/ACL 98*, Montréal. Ed. by M. Stede, L. Wanner, and E. Hovy, 29–35.
- Degand, L. 2014. "'So very fast, very fast then' Discourse markers at left and right periphery in spoken French". In: *The role of the left and right periphery in semantic change*. Ed. by K. Beeching and U. Detges. Amsterdam: John Benjamins.
- Degand, L., B. Cornillie, and P. Pietrandrea, eds. 2013. *Discourse markers and modal particles. Categorization and description*. Amsterdam: John Benjamins.
- Degand, L. and G. Gilquin. 2013. "The clustering of 'fluencemes' in French and English". In: *ICCL 2013*, Gent.
- Degand, L., L. Martin, and A.-C. Simon. 2014. "LOCAS-F: un corpus oral multigenres annoté". In: *CMLF 2014 - 4e Congrès Mondial de Linguistique Française*, July 19-23, Berlin, Germany. EDP Sciences.
- Dehé, N. and A. Wichmann. 2010. "The multifunctionality of epistemic parentheticals in discourse. Prosodic cues to the semantic-pragmatic boundary". In: *Functions of Language* 17.1, 1–28.
- Denturck, E. 2008. "Étude des marqueurs discursifs. L'exemple de quoi". MA thesis. Universiteit Gent.
- Diewald, G. 2013. "'Same same but different' - Modal particles, discourse markers and the art (and purpose) of categorization". In: *Discourse markers and modal particles. Categorization and description*. Ed. by L. Degand, B. Cornillie, and P. Pietrandrea. Amsterdam: John Benjamins, 19–46.
- Dister, A. et al. 2009. "Du corpus à la banque de données. Du son, des textes et des métadonnées. L'évolution de banque de données textuelles oral VALIBEL (1989-2009)". In: *Cahiers de Linguistique* 33.2, 113–129.
- Fischer, K., ed. 2006a. *Approaches to discourse particles*. Studies in pragmatics 1. Amsterdam: Elsevier.
- Fischer, K. 2006b. "Towards an understanding of the spectrum of approaches to discourse particles: introduction to the volume". In: *Approaches to discourse particles*. Ed. by K. Fischer. Amsterdam: Elsevier, 1–20.
- Fraser, B. 1999. "What are discourse markers?" In: *Journal of Pragmatics* 31, 931–952.

- Fraser, B. 2006. "Towards a theory of discourse markers". In: *Approaches to discourse particles*. Ed. by K. Fischer. Amsterdam: Elsevier, 189–204.
- Gonzalez, M. 2005. "Pragmatic markers and discourse coherence relations in English and Catalan oral narrative". In: *Discourse studies* 7.1, 53–86.
- Gotz, S. 2013. *Fluency in native and nonnative English speech*. Amsterdam: John Benjamins.
- Halliday, M. A. K. and R. Hasan. 1976. *Cohesion in English*. London: Longman.
- Hansen, M.-B. 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 K. Fischer. Amsterdam: Elsevier, 21–41.
- Haselow, A. 2011. "Discourse marker and modal particle: the functions of utterance-final *then* in spoken English". In: *Journal of Pragmatics* 43.14, 3603–3623.
- Hasselgard, H. 2006. "'Not now' - on non-correspondence between the cognate adverbs *now* and *nå*". In: *Pragmatic markers in contrast*. Ed. by K. Aijmer and A.-M. Simon- Vandenberg. Amsterdam: Elsevier, 93–113.
- Kaltenbock, G. 2009. "Initial *I think*: main or comment clause ?" In: *Discourse and Interaction* 2.1, 49–70.
- Kurt, J. 2012. "Pedagogic corpora for content and language integrated learning. Insights from the BACKBONE Project". In: *The Eurocall Review* 20.2.
- Lacheret, A., S. Kahane, and P. Pietrandrea, eds. 2014. *Rhapsodie: a prosodic and syntactic treebank for spoken French*. Studies in Corpus Linguistics. Amsterdam: John Benjamins.
- Lindström, J. 2001. "Inner and outer syntax of constructions: the case of the *x och x* construction in Swedish". In: *Pragmatic aspects of frame semantics and construction grammar*, 7th International Pragmatics Conference, Budapest July 9-14.
- Meyer, T. et al. 2011. "Multilayer annotation and disambiguation of discourse connectives for machine translation". In: *Proceedings of the SIGDIAL 2011 Conference*, June 17-18, 2011, Portland, Oregon, USA, 194–203.
- Nelson, G., S. Wallis, and B. Aarts. 2002. *Exploring natural language: Working with the British component of the International Corpus of English*. Amsterdam: John Benjamins.
- Norrick, N. R. 2009. "Interjections as pragmatic markers". In: *Journal of Pragmatics* 41, 866–891.
- Palisse, S. 1997. *Artisans, Assureurs. Conversations téléphoniques en entreprise*. CLAPI.
- Pichler, H. and S. Levey. 2011. "In search of grammaticalization in synchronic dialect data: General extenders in North-East England". In: *English Language and Linguistics* 15.3, 441–471.
- Pitler, E. and A. Nenkova. 2009. "Using syntax to disambiguate explicit discourse connectives in text". In: *Proceedings of the ACL-IJCNLP Conference Short Papers*, 13–16.
- Pons Bordería, S. 2006. "A functional approach to the study of discourse markers". In: *Approaches to discourse particles*. Ed. by K. Fischer. Amsterdam: Elsevier, 77–100.
- Prasad, R. et al. 2007. *The Penn Discourse Treebank 2.0 annotation manual*. Tech. rep. Institute for Research in Cognitive Science.

- Prsirr, T., J.-P. Goldman, and A. Auchlin. 2013. "Variation prosodique situationnelle: étude sur corpus de huit phonogenres en français". In: *Proceedings Prosody-Discourse Interface*, Leuven, September 11-13.
- Redeker, G. 1990. "Ideational and pragmatic markers of discourse structure". In: *Journal of Pragmatics* 14.3, 367–381.
- Rendle-Short, J. 2004. "Showing structure: using um in the academic seminar". In: *Pragmatics* 14.4, 479–498.
- Roekhaut, S. et al. 2014. "eLite-HTS: a NLP tool for French HMM-based speech synthesis". In: *Fifteenth Annual Conference of the International Speech Communication Association*.
- Rouchota, V. 1996. "Discourse connectives: what do they link ?" In: *UCL Working papers in Linguistics* 8, 1–15.
- Roulet, E. 2006. "The description of text relation markers in the Geneva model of discourse organization". In: *Approaches to discourse particles*. Ed. by K. Fischer. Amsterdam: Elsevier, 115–132.
- Santorini, B. 1990. Part-of-speech tagging guidelines for the Penn Treebank Project (3rd revision, 2nd printing). Technical report MS-CIS-90-47. Department of Computer and Information Science, University of Pennsylvania.
- Schiffrin, D. 1987. *Discourse markers*. Cambridge: CUP.
- Schmid, H. 1995. "Improvements in Part-Of-Speech tagging with an application to German". In: *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland.
- Schmidt, T. and K. Wörner. 2012. "EXMARaLDA". In: *Handbook on Corpus Phonology*. Oxford University Press, 402–419.
- Schourup, L. 1999. "Discourse markers". In: *Lingua* 107, 227–265.
- Shriberg, E. 1994. "Preliminaries to a theory of speech disfluencies". PhD thesis. University of California at Berkeley.
- Stenstrom, A.-B. 2009. "Vague category markers in London and Madrid Youthspeak". In: *Corpora and discourse - and stuff. Papers in honour of Karin Aijmer*. Ed. by S. O. R. Bowen M. Mobarg. Goteborg: Acta Universitatis Gothoburgensis, 287–296.
- Sweetser, E. 1990. *From etymology to pragmatics*. Cambridge: CUP.
- Swerts, M. 1998. "Filled pauses as markers of discourse structure". In: *Journal of Pragmatics* 30, 485–496.
- Szmrecsányi, B. M. 2004. "On operationalizing syntactic complexity". In: *Le poids des mots. Proceedings of the 7th international conference on textual data statistical analysis*. Ed. by C. Purnelle, C. Fairon, and A. Dister. Louvain-la-Neuve: Presses universitaires de Louvain, 1032–1039.
- Taboada, M. and M. d. I. Á. Gómez-González. 2012. "Discourse markers and coherence relations: Comparison across markers, languages and modalities". In: *Linguistic and the Human Sciences* 6, 17–41.
- Tesnière, L. 1959. *Éléments de syntaxe structurale*. Klincksieck, Paris.

- Tottie, G. 2011. "Uh and Um as sociolinguistic markers in British English". In: *International Journal of Corpus Linguistics* 16.2, 173–197.
- Traugott, E. 2007. "Discourse markers, modal particles, and contrastive analysis, synchronic and diachronic". In: *Catalan journal of linguistics* 6, 139–157.
- Vincent, D. 1993. *Les ponctuations de la langue et autres mots du discours*. Québec Nuit Blanche Editeur.
- Waltereit, R. and U. 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.
- Zufferey, S. and L. Degand. 2014. "Representing the meaning of discourse connectives for multilingual purposes". In: *Corpus Linguistics and Linguistic Theory* 10.

Appendix 2: Fluenceme-level annotation protocol (Crible et al. 2016)

The second appendix consists in the annotation protocol designed for all fluenceme-level variables. The full document is reported below in its original format, and thus contains its own references and section numbers. The original standalone document is available upon request.

**Annotation manual of fluency and disfluency markers in multilingual,
multimodal, native and learner corpora**

Version 2.0

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

9th Feb. 2016

To cite this paper: Crible L., Dumont A., Grosman I., Notarrigo I. 2016. Annotation manual of fluency and disfluency markers in multilingual, multimodal, native and learner corpora. Version 2.0. *Technical report*. Université catholique de Louvain and Université de Namur.

This research is the outcome of the collaboration between the
Université catholique de Louvain and the Université de Namur



Fluency and disfluency markers: A multimodal contrastive perspective

“Action de Recherche Concertée” funded by the Fédération Wallonie-Bruxelles

(Grant number 12/17-044)



1 Introduction

1.1 Research context

Research on fluency and disfluency phenomena is flourishing, multi-disciplinary and adopts different approaches depending on the scientific goals. This situation has led to a panel of annotation protocols available in the literature which are nevertheless rarely comparable or generalisable to different data types. Since Shriberg (1994) and her seminal typology of markers, several authors have tried to adapt her coding scheme and terminology, with a view to allowing automatic annotation or to modifying her categories. Major inspiring studies in our research are: Shriberg (1994), Meteer (1995), Candéa (2000), Eklund (2004), Dister (2007) and Götz (2013). Other ongoing projects following similar approaches include Christodoulides, Avanzi, and Goldman (2014), Moniz et al. (2014) and Pallaud, Rauzy, and Blache (2013). We will refer to more specific studies focusing on particular phenomena in the following sections when relevant.

The interest for componential approaches to fluency as proposed by Götz (2013) tends to spread. However, the specific elements participating in this vision of fluency such as repetition, substitution, pause etc., as well as annotation formats often differ on the scope of covered phenomena. Concretely, differences between typologies pertain to the number and types of phenomena, language- and modality-related specificities, technical aspects of the annotation and quantitative treatment (labelling, data extraction), etc. Generally, most protocols show a number of pitfalls, either on practical aspects such as the replicability of the annotation system and its efficiency in quantitative analyses, or on theoretical considerations such as the validity of the categories and discriminating criteria and the overall cognitive-pragmatic relevance of the whole model.

Against this backdrop, the present protocol aims at addressing a number of these issues, by taking into account the contributions of previous research but also a wealth of theoretical frameworks and data types involved in our concerted research action. The originality of the present approach is to try to overcome the specificities of a particular framework by offering an exhaustive and flexible model, which can potentially adapt to many research questions.

The following sections will first present our theoretical and methodological framework, before turning to operational definitions of all phenomena covered by the annotation.

1.2 Purpose of the manual

The purpose of this document is to provide the annotators with a coding scheme for (dis)fluency markers in line with the componential approach to (dis)fluency as defined by Götz (2013) and flexible enough to account for such markers in French, English (L1 and L2) and French Belgian Sign Language (LSFB). This protocol has been developed in the framework of a collaboration between four PhD students working with different data types in terms of language (French, English, LSFB), modality (speech vs. sign language), speakers (native speakers vs. learners), and, more generally, in their analytical approach (semi-automatic vs. manual, semasiological vs. onomasiological). This plurality of approaches not only vouches for the flexibility of the protocol, but also provides the foundation for a number of theoretical and practical decisions at the core of this work. By adopting this protocol and by using the same labels for the same category across different types of corpora, the different teams will ensure the comparability and inter-operability of their research. We would like to argue that this linguistic, modal and theoretical adaptability strengthens the robustness and relevance of the annotated categories and of the protocol as a whole.

In order to ensure replicability, (i) each phenomenon covered by the coding scheme is systematically defined, (ii) discriminating and/or application criteria are explained, (iii) labels and other technical aspects are considered, (iv) and illustrations with authentic corpus examples are provided.

The definitions and criteria presented here emerged from the comparison of existing models and their application to authentic corpus data. This corpus-based method vouches for the operability of the protocol by ensuring the applicability of the criteria and minimizing inter-annotator disagreement. It remains that some categories may be more prone to individual interpretation than others, given the inevitable – but limited – semantic-pragmatic considerations involved in any approach to (dis)fluency.

Nevertheless, the majority of the phenomena and criteria selected here are related to formal features or surface cues that considerably constrain the analysis, as will be shown for the different categories in section 3.

2 Overview of the protocol: design and applications

2.1 Definitions and terminology

(Dis)fluency can be defined in a holistic perspective as the smooth, fast and effortless use of language (Crystal 1988). In this view, the fluent or disfluent interpretation results from the hearer's perception and the overall impression of the interaction. The componential approach, on the other hand, views fluency and disfluency as a combination of features that, taken in isolation, can either be involved in fluent or disfluent speech depending on their frequency, function, position and combination patterns. In other words, these markers can be considered either as felicitous and deliberate cues to support production and comprehension, or as signals of online planning and production difficulties. These difficulties stem from lexical access trouble ("tip-of-the-tongue" phenomenon), semantic and/or syntactic complexity, cognitive or emotional load, stalling for planning time, or more directly the perception of a "mistake" by the speaker leading to its reformulation. In contrast, positive effects of (dis)fluency markers include anchoring an utterance in discourse and the interaction, informative salience, or stylistic effect. In general, they can be understood as "relevant" in terms of an effort-effect ratio (Sperber and Wilson 1995).

Furthermore, our definition of (dis)fluency is fundamentally situational, i.e. fluency is only evaluated with respect to expectations in particular interaction settings. Thus, the recurrence of the same phenomenon (e.g. frequent silent pauses) can either create a fluent effect in a formal situation such as a political speech, or contribute to an impression of disfluency in a more informal setting. This ambivalence led us to use the term "(dis)fluency" in this manual and other related publications, in order to avoid any interpretative bias.

Apart from this general view of fluency and disfluency, other key notions referring to the different structures observed need to be briefly defined as well, namely *fluenceme* (simple and compound), *phrase* and *sequence*.

Following Götz (2013), we use the term "fluenceme" to refer to any (dis)fluency marker, with no a priori evaluation of its fluent or disfluent character. In the coding scheme, two groups of fluencemes are distinguished according to their internal structure: "simple" fluencemes consist of one part only (discourse markers, different types of pauses, false starts, editing terms); "compound" fluencemes show a two-part structure (repetitions and substitutions). Truncation is either simple if it is left incomplete, or compound when it is completed.

The term "phrase" is used to refer to lexicalized units consisting of several graphic words or several signs. It can for instance be used for some discourse markers (e.g. "tu vois", "in other words"). The term "sequence" refers to any segment of speech containing at least one fluenceme, simple or compound, either isolated, juxtaposed or embedded.

2.2 Technical aspects of the annotation system

Corpus annotation is mainly manual, sometimes combined with automatic annotation, and is not restrictive in technical terms so that different data configurations can be accounted for. Transcriptions need to be segmented at word or sign level and to be aligned with the sound (or video), in a format that is compatible with a multi-layered annotation software (EXMARaLDA (Schmidt and Wörner 2012), ELAN (Sloetjes and Wittenburg 2008), PRAAT (Boersma and Weenink 2014), etc.).

This protocol organizes the annotation in two layers (or "tiers"). Simple and compound fluencemes are both annotated on the first tier, and their annotation can be completed when necessary by

“diacritics”⁷⁵ on the second tier. When a word belongs to several structures, it will receive a double tag, still on the first tier. We argue that the gain in information and efficiency that results from a unique annotation tier is a substantial asset for contextual analysis of fluencemes. However, thanks to the labelling system, it still remains possible to query each individual fluenceme or to extract them on separate tiers if necessary.

Each word or sign within a fluenceme is annotated. All elements included in a fluenceme are tagged with a label containing brackets, two-letter initials and sometimes numbers (cf. section 2.2). For instance, the phrase “you see” receives a tag at each graphic unit (i.e. one for “you” and one for “see”), even though it constitutes only one discourse marker (frequency counts are not altered thanks to the bracketing system). However, within a sequence of fluencemes, only the elements defined in the present protocol are annotated, i.e. fluencemes and related phenomena defined in the protocol such as certain parenthetical asides or inserted words; the annotation thus excludes any other element located in a sequence which is not explicitly defined in this protocol.

The annotation is mainly linear, but back-and-forth movements to the left and right contexts are not excluded if they are required for the interpretation. Moreover, when a sequence includes several compound fluencemes, the annotation focuses on a global interpretation of the phenomena: although every fluenceme is individually accounted for, some compound structures can be identified as supporting the whole sequence if they provide a coherent and representative interpretation of the segment. Concretely, in the following example, the higher-order fluenceme is the partial repetition of “it’s a long process” repeated twice with internal modulations:

*Backbone: Bb_en009*⁷⁶

it's	a	long	process	(345)		
<RM0	RM0	RM0	RM0<SP0	< UP >		
				< WI >		
it	's	a	long	haul		
RM1	RM1	RM1	RM1<SP0	SP1>		
and	it	's	a	very	expensive	process
<DM>	RM2	RM2	RM2	<IL>	SP1>	RM2>
< WI						
>						

A bracketing system is used to delineate the boundaries of a (simple or compound) fluenceme. In other words, brackets are applied to all types of fluencemes, even those containing only one element (e.g. a pause or a false start). The position of the brackets is meaningful: open brackets “<” indicate the beginning of the phenomenon marked to their right (e.g. <RM0), whereas closing brackets “>” show the end of the phenomenon marked to their left (e.g. RM2>).

⁷⁵ By “diacritics” we mean elements that are hierarchically dependent on the main annotation tier. In other words, they form a secondary annotation level giving additional information to previously annotated fluencemes (cf. section 3.4)

⁷⁶ Examples are authentic and their reference corresponds to the name of the text in its corpus. Corpora used for the examples in the present document are: the French component of LINDSEI (Gilquin, De Cock, and Granger 2010), BACKBONE (Kurt 2012), LSFb (Meurant 2015), C-Humour (Grosman in press).

Labels	Fluencemes
<i>Fluenceme tier</i>	
UP	Unfilled pause
FP	Filled pause
S1/S2/S3	Manual stop
PU	Palm Up (LSFB)
DM	Discourse marker
ET	Explicit editing term
FS	False start
TR	Truncation
RI	Identical repetition
RM	Modified repetition
RE	Framing repetition
RG	Grammatical repetition (LSFB)
SP	Propositional substitution
SM	Morphosyntactic substitution
IL	Lexical insertion
IP	Parenthetical insertion
DE	Deletion
<i>Diacritics tier</i>	
AR	Misarticulation
LG	Lengthening
WI	Within
OR	Re-ordering
CF	Functional completion
CT	Total completion

Table 1: Overview of labels in the protocol

A numbering system is used for compound fluencemes, namely repetitions, substitutions and repeated truncations. In case of embedded fluencemes, simple fluencemes are always tagged before others: the labels UP, FP, DM, ET and TR always precede those of repetition and substitution. Diacritics (OR, AR, LG, CT, CF, WI) are likewise also annotated on a separate tier. Examples in the remainder of this document will thus contain two tiers in addition to the transcription tier.

The annotation procedure presented here is neither constraining nor restrictive: not all phenomena need to be annotated if they do not contribute to a particular research question, and additional labels for specific modalities can be appended (e.g. pauses, cf. 3.1.1 or repetitions in LSFB, cf. 3.2.1).

3 Categories of fluencemes covered by the annotation protocol

This section defines and illustrates every annotation label. It summarizes similarities and differences with the phenomena as defined in the literature and specifies where the present approach stands. When no reference is mentioned, the phenomenon has, to the authors' knowledge, not been documented before.

3.1 Simple fluencemes

As a reminder, simple fluencemes are markers consisting of only one element. These phenomena can occur in isolation (e.g. a pause used alone), juxtaposed to other markers (e.g. a pause directly followed by a discourse marker) or embedded in compound fluencemes (e.g. a pause within a repetition).

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

Despite the multimodal perspective of this protocol, pauses in spoken languages and sign languages (henceforth SL) are distinguished in order to avoid any premature mapping between different concepts. More specifically, although hand stops do occur in LSFB between signs by relaxing or fixing hands, the non-manual channel (i.e. eyes, gaze, eyebrows, mouth and head movements) keep on communicating information. These hand stops are therefore filled by the non-manual channel (Notarrigo 2016; Meurant and Notarrigo 2014). Moreover, contrarily to spoken languages where the sound vs. silence distinction is clear, in SL, hands do not “disappear” and the articulators are still visible: they have a specific location and shape in space. It is therefore premature, even perhaps not relevant, to refer to “empty” or “filled” pauses in SL. The analysis will in time shed more light unto the behavior of pauses in LSFB, thus opening the perspective of multimodal comparisons.

Unfilled pause (UP) This category covers any interruption of the sound signal above a certain threshold fixed by the annotation conventions already existing in the corpora⁷⁷ or fixed by taking into account the speaking rate in a given text⁷⁸. No distinction is made *a priori* between different types of pauses based on their duration (Little et al. 2013).

LINDSEI: FR001-F

an	institution	(330)	really	er	necessary	is
		<UP>		< FP >		

Backbone: Bb_fr013

mon	ami	Georget	(315) ⁷⁹	monsieur	Georget	Daumas
<SP0	SP0	<RM0	<UP>	SP1>	RM1>	
			< WI >			

The protocol does not distinguish, in the first tier, pauses (or any other simple fluenceme) located within a compound fluenceme from pauses that are not embedded. However, it is possible to specify this information in the diacritics tier, in order to identify these cases afterwards. In those cases, the “WI” (for “within”) tag is used (cf. section 3.4.3).

Filled pause (FP) This category covers active and conventional vocalizations traditionally seen as supporting or maintaining on-going speech (Duez 2001; Strassel 2003 ; Vasilescu, Candea, and Adda-decker 2004). These vocalizations maintain the audio signal in a neutral position (e.g. *ah*, *eh*, *er*, *uh*, *um*, *euh*, *mh*). In case of hesitation between a filled pause and a lengthening, the latter is used for any post-vocal schwa that is not preceded by a silent pause.

The filled pause category can be transcribed in several forms (*erm*, *er*, *eh*, etc.). However, vocalizations such as *mh*, *mm*, *mhm* are only annotated when they constitute hesitation markers, not backchannelling devices.

Backbone: Bb_fr 003

football	en	France	en	fait	euh	s'appuie
			<DM	DM>	< FP >	

⁷⁷ ICE and VALIBEL have different annotation conventions regarding pauses (variable thresholds from 100 to 250 ms).

⁷⁸ Either using an optional function of the annotation software DisMo (Christodoulides, Avanzi, and Goldman 2014) or with a system similar to Little et al. (2013).

⁷⁹ In the examples, duration (in milliseconds) of pauses is annotated in the transcription tier. Examples from different corpora will thus share the same format.

LINDSEI: FR005-F

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

Hand stop in SL (S1/S2/S3) The three categories used to describe pauses in LSFBE cover: hand stop during a sign (S1), hand stop between signs (S2), hesitation or lexical search moves (S3). The first group includes hands fixation in the starting position (<S1:ST>) or ending position (<S1:EN>) of a sign. The second group covers the moments when the signer is not signing (S2). Hands configuration is not significant in those cases: crossed (<S2:CR>), alongside the body (<S2:BO>) or relaxed in the neutral space in front of the signer (<S2:NE>). The last group (<S3>) includes the conventional sign corresponding to French *eah* as well as more idiosyncratic configurations. Because of overlap with other labels, S1 are annotated on a separate tier distinct from that of other fluencemes and other types of hand stops (see Notarrigo (2016) for more details).

LSFB⁸⁰: CLSFBE CAM1002

“yes it’s better to change for instance (320) there are too many spelling signs for example the sign for USB it’s better to use the sign for key”

AVENIR	MEILLEUR	CHANGER	PAR- EXEMPLE	(320)	TROP
			<RM0	<S2:CR>	< SP 0
				< WI >	
EPELLATION	PAR- EXEMPLE	USB	MIEUX	CLEF	CLEF
<S1:EN>		<S1:EN>			< S1:EN >
SP0	RM1>	SP1	SP1	SP1><RI0	RI1>

LSFB: CLSFBE

« yes it happened in Paris Paris uh three four years ago in Paris»

PARIS	PARIS	EUH	TROIS	QUATRE	ANS	PASSE	PARIS
	<S1: EN >						
<RI0	RI1><RM0	<S3:EU>	<SP0	SP1>			RM1>

3.1.2 Palm-up in SL (PU)

This category covers the following hand configuration: one or two hands are stretched, the five fingers separated from each other, making a rotating movement from the wrist to get the palms up. Palm-ups occur in the neutral space in front of the signer.

Palm-ups can have several functions (Loon 2012): they can indicate the beginning or end of a turn, show the signer’s attitude towards his/her own discourse, invite the hearer to react, establish a relation between two units of discourse, fill a stop while showing the will to keep the turn. Palm-ups are thus

⁸⁰ In all LSFB examples, the sentence above the table is a translation in English without punctuation. The top line is a gloss of the video source, i.e. each sign appearing in the discourse in LSFB is assigned a label corresponding to a lemma.

quite similar to discourse markers (see Bolly, Gabarro-Lopez, and Meurant (2015) and Gabarro-Lopez (2015) for more detail).

LSFB: CLSFBE JMS22076

« that's it I sign now yes uh in my opinion well what is the goal of the deafs world day »

/	PT:PRO1	SIGNER	/	OUI	(220)	EUH	PT: PRO 1
<PU>			<PU>		<S2:CR>	<S3: EU >	
/	JMS	VRAI	SOURD	MONDE	JOUR	POUR	QUOI
< PU >							

LSFB: CLSFBE JMS22076

« deaf people oh they are not really involved enough in the world of deafness »

SOURD	/	PAS	ASSEZ	DANS	PLUS	MONDE	SOURD
	< PU >						

3.1.3 Discourse marker (DM)

This category covers grammatically heterogeneous and multifunctional elements that are syntactically optional and oriented towards interpretation processes. They function at the metadiscursive level as cues to situate their host-unit in a dialogically co-built representation of speech (Brinton 1996; Hansen 2006; Crible in press). They can either signal a discourse relation between the host-unit and its context, make the structural organizations of segments explicit, express the speaker's metacomments on their speech, or contribute to interpersonal collaboration.

Discourse markers thus cover very diverse elements such as coordinating and subordinating conjunctions (“done”, “parce que”), adverbs (“enfin”), particles (“ben”), etc. A limited number of interjections can also be considered as discourse markers when they fulfill a text-structuring function (e.g. “oh” to introduce reported speech, “eh ben” as turn-opener). Backchannelling⁸¹ devices are excluded from this category given their nonlexical form (“mh mh”) and specific function.

To keep a word-level annotation, each element in a discourse marker (for instance “je veux dire”) receives the two-letter label. This does not mean that each separate element is granted the status of discourse markers on its own. Thanks to the bracketing system (< on the first and > on the last word of a phrase), it is possible to concatenate the labels assigned to each word in order to get a unique unit in post-treatment: the phrase “je veux dire” will be annotated <DM DM DM>.

The difference between discourse markers and explicit editing terms can sometimes be subtle. Therefore, the following criteria borrowed from Crible (2014) will be used to distinguish them and exclude some pragmatic expressions from the DM category:

⁸¹ We define backchannel on an interactional criterion: they are the signal of understanding, vocalized by the interlocutor, i.e. not the current main speaker.

- explicit reference to lexical access trouble (thus excluding expressions such as “comment”, “comment dire”, “comment dirais-je”, “si je peux dire”)
- low degree of grammaticalization (free juxtapositions and semantic transparency)
- presence of propositional content (thus excluding “if I may say so” for instance).

Borderline cases are, for instance, “si vous voulez”, “si tu veux”, “if you will”, “je dirais”, “on va dire”, “I don’t know”, “I suppose”, which present a high degree of fixation but remain explicit references to the act of speaking or thinking. In some contexts, they can still be considered as discourse markers. We acknowledge that the identification of discourse markers is never free from language-specific considerations and remains partly subjective when dealing with items at the border of the category.

LINDSEI: FR014-F

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

Backbone: Bb_en019

which	is	/	sort	of	/	I	suppose	would	be
	<SM0	<UP>	<DM	DM>	<UP>	<DM	DM>	SM1	SM1>
		<WI>	<WI	WI>	<WI>	<WI	WI>		

Backbone: Bb_fr021

loc/	euh	enfin	locales	euh	enfin	du	coin	enfin
<TR	<FP>	<DM>	TR><SP0	<FP>	<DM>	SP1	SP1>	<DM>
	<WI>	<WI>		<WI>	<WI>			

Backbone: Bb_fr013

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

LSFB: S041 CLSFBI1905

« first ASL it's signs with iconicity it's visual »

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

LSFB: S041 CLSFBI1905

« if you didn't understand I can sign »

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

LSFB: CLSFBE-JMS20064

« they learn ne/ it's as if they discover »

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

3.1.6 Truncation (TR)

This category covers any word fragment, either completed (immediately or with delay) or abandoned. This phenomenon reveals formal incompleteness of a morpheme or word that can be left incomplete or be taken up and modified (Pallaud and Henry 2004). Similarly, in LSFB, a truncation is the onset of an interrupted sign. The sign must be recognizable by the hands configuration and the initial localization (albeit incomplete). The truncated word/sign can be directly completed with the repetition of the fragment, or completed after the insertion of a few elements, or be abandoned (Henry and Pallaud 2003: 78). In this last case only, truncations are simple fluencemes; otherwise in all the other cases, they are compound.

The fragment is annotated along with its completed form, when applicable. An abandoned fragment takes opening and closing brackets. Without counter-evidence, we assume that the next word is the complete form of the fragment as soon as the first phoneme is in common between the fragment and its restart. When a truncation is combined with a false start, the former fluenceme is annotated before the latter, as in the LSFB example in section 3.1.5 (CLSFBE-JMS 20064).

LSFB: S070 CLSFBI3406

« sometimes a person is mad so they sign fast »

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

LSFB: CLSFBE-JMS22-068

« to meet members of the ASBL and get lessons, explanations⁸² »

ASBL	RENCONTRER	ENSEIGNER	ENSEIGNER	EXPLIQUER
		<TR	TR>	
		< AR >		

In accordance with our definition of identical repetition, according to which a repetition only applies to a lexical segment (see section 3.2.1), a truncation cannot be affected by a repetition. Successive fragments will be annotated by a numbering system, as in the following example:

⁸² The signer begins the standard sign “ENSEIGNER”. The sign starts from the signer’s body to the interlocutor. Then, the signer stops and changes the standard sign “ENSEIGNER” (in the sense of “I teach”) by the derived version of the verb (“I am taught”). The signer’s hands change their direction and the sign is completed from the interlocutor towards the signer.

(Built example)

la	m/	m/	maman
	<TR0	TR1	TR>

When a fragment is completed, it can be immediately or after a lexical insertion or another fluenceme.

Backbone: Bb_fr018

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

The following two cases are examples where the audio recording is necessary in order to determine the status of the truncated word, i.e. whether it is a misarticulation or a propositional substitution.

Backbone: Bb_fr003

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

Backbone: Bb_en009

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

3.2 Compound fluencemes

As mentioned above (section 2.1), the structure of a compound fluenceme consists in at least two parts. It is not excluded that these phenomena can apply to elements previously annotated as simple fluencemes, namely discourse markers that can be repeated or substituted.

For the annotation of compound fluencemes, a numbering system is used: identical numbers correspond to words from the same part of the fluenceme while increasing numbers represent the number of times the segment has been repeated and/or substituted, as will become clear with the examples of the following sections.

3.2.1 Repetitions (RI, RM, RE, RG)

Bearing in mind that this annotation system makes no a priori judgment between potentially fluent or disfluent markers, all types of repetitions are annotated, including “rhetorical” repetitions and repetitions that are caused by external interventions in the interaction, provided they occur in the same speech turn. By extension, this restriction also applies to reported speech, thus excluding from the annotation cases where the repetition spreads over both direct and reported speech.

The protocol accounts for cases where the repeated elements appear in a different order than their first occurrence: in the diacritics tier, the label <OR> can be added to mark a re-ordering. Moreover, other labels on the diacritics tier, namely <CF> for functional completion and <CT> for total completion, signal the degree of syntactic completion of the repeated or substituted segment (see sections 3.2.1 and 3.2.2).

Identical repetition (RI) This category covers cases where one or several (quasi) contiguous words are repeated in their exact same form and without any semantic addition (Candéa 2000). “Quasi-

contiguity” refers to the possibility to insert an element with low or no propositional content between the two parts of the repetition, namely an UP, FP, DM, ET or an IP (since the content of a parenthesis does not affect the content of the repeated elements). As mentioned above, repetitions only apply to complete propositional elements, thus excluding truncations and filled pauses (which are not lexical).

Backbone: Bb_fr004

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

LINDSEI: FR002-F

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

LSFB: S050 CLSFB2406

« or or there a lot of times, too often, when I meet oralists ⁸³ »

OU	OU	/	BEAUCOUP	/	MOMENT
<S1:EN>					<S1: EN >
<DM><RI0	<DM>RI1>	<S2:NE>	<RM0	<S2: NE >	
/	BEAUCOUP	/	ORAL	RENCONTRER	RENCONTRER
			<S1: ST >		
<S2:NE>	RM1>	<S2:NE>		<RG0	RG1>

Modified repetition (RM) This category covers cases where one or several (quasi) contiguous words are repeated either partially or with a change in content. This phenomenon is based on a less restrictive definition than that of identical repetitions since it also includes the possibility of having syntactico-semantic modification. This modification can be a lexical insertion, a deletion or a substitution. In other words, a RM differs from a RI by the fact that its content is modified by one or several propositional elements. This category, however, cannot be applied to discourse markers and explicit editing terms.

During the analysis of sequences, formal redundancy primes over semantic interpretation. In other words, if the semantic change (caused by an insertion for instance) is “strong” as in “I ate some chocolate, some chocolate cake” (change of referent), the sequence is annotated as a repetition (and not as a propositional substitution) since formal parameters are prioritized over semantic considerations, in an attempt to enhance reliability and objectivity.

Clitics are also excluded from this category in order not to overload the annotation: a partial repetition where only a clitic (e.g. a subject pronoun) is repeated and the other words (e.g. the verb) are substituted will not be annotated, as in “they come they go”.

⁸³ Deaf people communicating in a spoken language.

Backbone: Bb_en021

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

Backbone: Bb_en009

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

Backbone: Bb_fr008

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

LSFB: S041 CLSFBII905

« here in Europe more in the region of Belgium and France we are against dactylogy we don't like dactylogy »

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

LSFB: S007 CLSFBII306

« and moreover if there is a little mistale they blame me for it and moreover the mistake is related to a communication problem »

PLUS	APPARAÎTRE	PETITE	ERREUR	ACCUSER	ERREUR	ACCUSER
<DM>			<RM0 ¹⁰	RM0	RM1	RM1<RG0
ACCUSER	PLUS	ERREUR	COMM_PAS	COMM_PAS	ERREUR	COMM_PAS
< S1:EN >						
RM1>RG1>	<DM>	<RM0	RM0<RG0	RG1>	RM1	RM1>

Framing repetition in SL (RE) This category covers repetitions referred to in the SL literature as “doubling” (Kimmelman 2013) or “reduplication” (Vermeerbergen and Vriendt 1994). According to these authors, this type of repetition belongs to syntactic structures available in SL (for instance, there could be two subject slots for one verb (SVS)) or has discourse functions (for example in the topic-comment structure, in contrast or parenthetical constructions). They appear in a X Y X form, following the model by Kimmelman (2013). This type of repetition can apply to any class of signs (nouns, verbs, question words, adjectives, pronouns, adverbs etc.) and take scope over a sign within a unit or over the

unit itself if it frames another unit. There are thus two parts of a repetition framing a propositional element. The repeated components act as symmetrical “braces” around a central element made up of one sign to one or several clauses

For practical reasons, we do not annotate insertions (either lexical or parenthetical) between the two parts of a framing repetition since it can cover very large segments, in which case the whole text would receive a label <IL>, rendering the annotation counterproductive.

LSFB: S007 CLSFBI306

« we used to pretend to fight for play »

RIRE	JOUER	BOXE	RIRE
<RE0			RE1>

Grammatical repetition in SL (RG) This category covers so-called “morphological” repetitions (Filpczak and Mostowski 2013) or “morphosyntactic reduplication” (Wilbur 2005; Pfau and Steinbach 2006). Those have grammatical functions such as marking a change in number or gender, marking the scope of the verb or marking aspect. For example, to express the fact that there are several houses, the signer can repeat *n* times in a row the sign for “HOUSE” either at the same location or placing them in the space in front of him/her. Another case is that of an action that is performed repeatedly with a sense of boredom. For instance, the signer can produce *n* times in a row the sign for “WORK” to express the idea that they work a lot.

Some degree of variation between the two parts of the repetition is accepted: “repeated signs are not necessarily identical; one occurrence can present some modulation of the citation form or be accompanied by non-manual features that are absent in the other occurrence” (Vermeerbergen and Vriendt 1994). Kimmelman (2013) also observes phonological, phonetic and grammatical variations between two parts of a repetition. For example, the second occurrence can be articulated in a more relaxed way and with a shorter and weaker move than the first one. The author further notes a difference of placement of the sign in the second occurrence in order to add some information of aspect, modality or other (see Notarrigo (2016) and Notarrigo, Meurant, and Van Herreweghe (2016) for more details).

LSFB: S041 CLSFBI1905

« but the structure is the same as in French “De, la, à” they are taken out the structure of signs follows that of French »

MAIS	(830)	STRUCTURE	AUSSI	FRANÇAIS	FS:ADELA	RIEN	ENLEVER
				<S1: EN >			
<DM>	<SE:NE>	<RE0	RE0	RE0			< RG 0
ENLEVER	ENLEVER	MOT	STRUCTURE	AUSSI	LS	FRANCAIS	
						<S1: EN >	
RG1	RG2>		RE1	RE1	<IL>	RE1>	

3.2.2 Substitutions (SM, SP)

Morphosyntactic substitution (SM) This category covers any morphosyntactic modification. Any modification of a full lemma (thus excluding truncations) is annotated. In other words, any addition,

change or elision of morpheme from the same lemma is annotated. Morphosyntactic substitutions often involve partial repetitions.

Backbone: Bb_fr004

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

Backbone: Bb_en022

you	can	make	have	anything	made	in	Birmingham
		<SM0	SM1	<IL>	SM1 ⁸⁴ >		

Propositional substitution (SP) This category covers any take-up of a discourse segment by another which includes a semantic nuance or modification. Substitution can include cases of modified repetitions (Duez 2001). In other words, this category concerns any element being replaced, including in an interrupted sequence. Appositions and reference chains (e.g. “this man he went there”) are however excluded from propositional substitutions in this protocol.

As for the different categories of repetition, different types of substitution are distinguished in this protocol, namely substitutions taking scope over complete segments (see diacritics tier) and substitutions taking scope over incomplete segments. However, this incompleteness is not annotated: only completion gets a label. This makes it possible to distinguish a sentence such as “he goes uh we go to the market” (substitution of an incomplete structure by a complete one) from “he goes to the market uh we go to the market” (substitution of complete structures). Note that cases of abandon with no formal or semantic take-up are analyzed as false starts (not as substitutions).

To determine the end of a substitution (right boundary), we use the grammatical criterion of word-to-word equivalence as scope of the annotation. If such an equivalence cannot be found in context, the interpretation of the end of the substitution is left to the annotator’s judgment on the basis of semantic and/or syntactic criteria, with a minimal bias (anything forming a complete, coherent content, a functional sequence). This choice aims at standardizing the annotation and prevents over-annotations of large segments.

Backbone: Bb_fr008

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

Backbone: Bb_en021

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

⁸⁴ The two forms “make” and “have made” can be considered as a change of structure rather than change of lemma (SP). In case of hesitation, a bias towards SM is preferred.

LSFB: S007 CLSFBI306

« in order for me to work well I need to be controlled »

BIEN	TRAVAILLER	BIEN	ENCADRER	ENCADRER
<RM0	<SP0	RM1>	SP1><RG0	RG1>

3.3 Insertions

3.3.1 Lexical insertion (IL)

This category covers the cases when a lexical element is inserted within a compound fluenceme, i.e. a partial repetition or a truncation. This phenomenon is not considered as a fluenceme *per se*, but its annotation is useful for the analysis in context of repetitions and truncations. Lexical insertions modify the co-text by adding some propositional information. In an attempt not to overload the annotation, all elements in an insertion are framed by one opening and one closing bracket, although this grouping might not be meaningful in a phraseological perspective.

Lexical insertions are only annotated in the second part of a repetition, between the two parts of a repetition, or between a word fragment and its completion. If a word appears in the first part of a repetition and is no longer present in the second part, it is rather annotated as a deletion (Shriberg 1994) (see section 3.3.3). Any element appearing after the closing bracket of a compound fluenceme, thus outside the sequence, is not annotated.

Lexical insertions semantically modify a pre-existing content, and are thus different from simple fluencemes embedded in a sequence of compound fluencemes. Insertions do not get a “WI” diacritic label (see section 3.4.3).

Backbone: Bb_en009

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

Backbone: Bb_fr008

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

LSFB: S007 CLSFBI306

« you see as if for instance as if we would write in French »

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

LSFB: S007 CLSFB1306

« it's normal even the oldest think hard to search I think about the answers and then I prepare before I type them »

NORMAL	PLUS	AGE	CONCENTRER	CHERCHER	PT:PRO1	CHERCHER
			<S1: EN >			
				<RM0		< TR
PENSER	CHERCHER	REPONDRE	AUSSI	PREPARER	CLAVIER	
<IL>	TR> RM1>		< DM >			

3.3.2 Parenthetical insertion (IP)

This category covers any propositional segment functioning as a “parenthetical aside” (Shriberg (1994: 61) and Strassel (2003: 42)). Parentheses are not considered as fluencemes *per se* and they are only annotated when they are placed within a sequence (i.e. in the immediate right or left context of a fluenceme, or embedded within a compound fluenceme). The content of the segment, although contextually related to the adjacent utterance, only indirectly affects this utterance by providing additional background information. It remains independent and is not syntactically integrated. As for lexical insertions, the brackets open and close on the first and last word of the parenthesis.

Backbone: Bb_en009

normally	would	take	you	(257)	before	you	're	a	fully
<RI0	RI0	RI0	RI0	<UP>	<IP	IP	IP	IP	IP
<OR	OR			< WI >					
qualified	solicitor	(222)	would	normally	take	you	a	minimum	of
IP	IP>	<UP>	RI1	RI1	RI1	RI1>			
		<WI>	OR	OR>					

Backbone: Bb_fr019

nous	avons	je	l'	ai	pas	dit	dans	la	présentation
		<IP	IP	IP	IP	IP	IP	IP	
au	début	mais	euh	un	euh	un	sous-sol		
IP	IP>	<DM>IP>	<FP>	<RI0	<FP>	RI1>			

LSFB: S040 CLSFB11905

« before I me my true way of expressing it's LSFB I think it's good and enjoyable why because I used to talk before »

AVANT	PT:PRO1	PT:PRO1	VRAI	EXPRESSION	LSFB	BIEN
<RI0	RI0	<IP	IP	IP	IP	IP
<OR	OR					
AGREABLE	POURQUOI	PARCE QUE	PT:PRO1	AVANT	PARLER	
IP	<DM> IP	<DM>IP>	RI1	RI1>		
				OR	OR>	

3.3.3 Deletion (DE)

This category covers the cases where a lexical element present in the first part of a modified repetition is absent in the second part. Deletions are conceptually the opposite of insertions, and as such do not constitute fluencemes either. Brackets apply to the first and last word of the deleted segment.

C- Humour: chfAR3

et	Rocard	qui	a	été	nommé	par	Sarkozy	ambassadeur
<DM>	<RM0	<DE	DE	DE	DE	DE	DE	DE>
au	pôle	nord	Rocard	au	pôle	nord		
RM0	RM0	RM0	RM1	RM1	RM1	RM1>		

3.4 Diacritic signs

A number of local phenomena are not treated as proper fluencemes but add some information to another element in the co-text. All diacritics are annotated in the second tier.

3.4.1 Misarticulation (AR)

This category covers any element identified as different from a “standard” or “correct” pronunciation and identified as such by the speaker him/herself. This misarticulation must be explicitly noticed by the speaker through a fluenceme (ET, DM, TR, etc.) otherwise it is not annotated (this is to avoid any reference to a linguistic “norm”).

Backbone: Bb_en009

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

3.4.2 Lengthening (LG)

This category covers any lengthening of a phoneme at the beginning, middle or end of a word, relatively to its expected duration. This annotation is perceptive and therefore arguably subjective. The label is applied to the word affected by the lengthening regardless of the position of the lengthening within the word.

LINDSEI: FR002-F

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

3.4.3 Embedding of simple fluenceme (WI)

A simple fluenceme framed by a compound fluenceme – for example, a pause in the middle of a repetition (between the two parts of the repetition) – gets a <WI> label (for *within*) on the diacritics tier. This information can be useful to compare the contexts of annotated fluencemes. Deletions, lexical and parenthetical insertions, however, never get the <WI> label given their “intrinsic” embeddedness.

Backbone: Bb_fr004

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

Backbone: Bb_en019

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

LSFB: S007 CLSFBI306

« I was sad and frustrated I felt frustrated. But it's as if I had felt that with a lot of delay [with respect to the other members of his family] »

FRUSTRER	TRISTE	FRUSTRER	/	SENTIR	FRUSTRER
<RE0		RE1><RM0	<SE:NE>	<IL>	RM1>
			< WI >		
/	MAIS	PT: PRO1	TITRE	RETARD	RETARD
		<S1: EN >			
<PU>	<DM>			<RG0	RG1>

3.4.4 Re-ordering (OR)

This category covers any element repeated identically – formally and semantically – but in a different syntagmatic order. Since the repeated elements are identical (except for the change of order), these cases will always be tagged “RI” in the first tier. The diacritic OR applies to all elements affected by the change of order in both parts of the fluenceme (RI0 and RI_n). Brackets open on the first word (first part of the repetition) and close on the last (last part of the repetition). *Backbone: Bb_en009*

normally	would	take	you	before	you're	a	fully	qualified	solicitor
<RI0	RI0	RI0	RI0	<IP	IP	IP	IP	IP	IP>
<OR	OR								
would	normally	take	you	a	minimum	of			
RI1	RI1	RI1	RI1>						
OR	OR>								

LSFB: S007 CLSFBI306

« when my uncle died my mother my brother and I were sitting side by side »

PT:PRO1	maman	frere	a-cote	maman	a-cote	PT:PRO1	PT:PRO3	mort
					<S1:EN>			<SI:ST><S1:EN>
<RI0	RI0		RI0	RI1	RI1	RI1>		
<OR	OR		OR	OR	OR	OR>		

3.4.5 Syntactic completion - functional and total (CF et CT)

This category covers two levels of completion: (i) a minimal level corresponding to a full functional sequence, for instance “a lot of money a lot of money” (as opposed to “a lot of a lot of money”), and (ii) a total completion corresponding to a whole utterance, a dependency structure with a predicate and its governed elements, typically a clause. These two labels are applied to repeated or substituted elements and are only given to the last word of the first part of the compound fluenceme. Incompletion (a case where the repetition or substitution takes scope over incomplete segments) is not annotated in order not to overload the annotation.

Backbone: Bb_fr021

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

Backbone: Bb_en021

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

4 Queries

The annotation protocol presented here is designed with the perspective of efficient extraction and query by the researcher, following the principle that when one type of fluenceme is queried, all relevant items are extracted except for the internal elements of a compound or phrasal structure (e.g. “parce que”, “in other words” will be counted as one DM and not respectively 2 and 3). Similarly, a segment repeated three times counts as one occurrence of a repetition. Since all labels contain two letters, querying by label should never retrieve unwanted occurrences.

Combining labels, brackets and numbers makes it possible to automatically extract groups and subgroups of phenomena without complex queries. Concretely, to extract all fluencemes of a specific type, the simple formula [<TAG] will suffice. For instance, querying [<RM] will return all occurrences of modified repetitions and identify the first element of this type of fluenceme.

Other queries can also be made on both tiers or on double labels in the same tier to specify particular contexts. For instance, all silent pauses in a “within” position, i.e. embedded in a compound fluenceme, can be extracted by filtering the <WI> tags applying to <UP>. Similarly, to extract all repeated discourse markers, the following formula can be used: [<DM<R]. The use of automatic scripts can help with more complex queries that are beyond the scope of this protocol.

5 Conclusion

This protocol aims at offering a methodological framework for the annotation of more than twenty (dis)fluency phenomena. It was tested on aligned multimodal and spoken corpora in English, French and LSFB.

Its exhaustivity pairs up with its flexibility: it is likely to adapt to many research questions while providing operational definitions of the phenomena under scrutiny. By avoiding any preconceived judgment on the fluency or disfluency of some markers, it substantially broadens the coverage of annotated phenomena compared to previous annotation systems, without additional difficulty. We hope

to have stressed the usefulness of using formally replicable criteria for the analysis that do not partake in particular theories of syntax, prosody or semantics, thus limiting the impact of subjective semanticpragmatic considerations in our approach to (dis)fluency. This protocol is still being tested on corpus, and inter-rater agreement analysis is in progress.

References in Appendix 2

- Biber, D., S. Johansson, G. Leech, S. Conrad, and E. Finegan. 2000. *Longman Grammar of Spoken and Written English*. Harlow: Pearson Education.
- Boersma, P. and D. Weenink. 2014. *Praat: doing phonetics by computer [Computer program]*. Version 5.3.80.
- Bolly, C., S. Gabarro-Lopez, and L. Meurant. 2015. "Pragmatic gestures at the gesturesign interface. Nonmanuals and palm-up gestures among older Belgian French speakers and French Belgian Sign Language signers". In: *Workshop Nonmanuals at the Gesture Sign Interface (NaGSI)*, Göttingen, Germany.
- Brinton, L. J. 1996. *Pragmatics markers in English. Grammaticalization and discourse functions*. New York: Mouton de Gruyter.
- Candéa, M. 2000. "Contribution à l'étude des pauses silencieuses et des phénomènes dits d'hésitation en français oral spontané. Etude sur un corpus de récits en classe de français." PhD thesis. Université Paris III Sorbonne Nouvelle.
- Christodoulides, G., M. Avanzi, and J.-P. Goldman. 2014. "DisMo: A Morphosyntactic , Disfluency and Multi-Word Unit Annotator An Evaluation on a Corpus of French Spontaneous and Read Speech Presentation of DisMo". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC) 2014, Reykjavik, Iceland, 26-31 May 2014*, 3902–3907.
- Crible, L. in press. "Towards an operational category of discourse markers: A definition and its model". In: *Discourse markers, pragmatic markers and modal particles: New perspectives*. Ed. by C. Fedriani and A. Sanso. Amsterdam: John Benjamins.
- Crible, L. 2014. *Identifying and describing discourse markers in spoken corpora. Annotation protocol v.8*. Tech. rep. Université catholique de Louvain.
- Crystal, D. 1988. "Another look at, well, you know..." In: *English Today* 4.1, 47–49.
- Dister, A. 2007. "De la transcription à l'étiquetage morphosyntaxique. Le cas de la banque de données textuelles orales VALIBEL". PhD thesis. Université catholique de Louvain.
- Duez, D. 2001. "Signification des hésitations dans la parole spontanée". In: *Revue Parole* 17-18-19.33, 113–138.
- Eklund, R. 2004. "Disfluency in Swedish human-human and human-machine travel booking dialogues". PhD thesis. Linköping Studies in Science and Technology.
- Filpczak, J. and P. Mostowski. 2013. "Repetition in Polish Sign Language (PHM). Discourse grammar information structure". In: *Theoretical Issues in Sign Language Research Conference (TISLR 11)*. London, United Kingdom.
- Gabarro-Lopez, S. 2015. "Les marqueurs du discours en langue des signes de Belgique francophone (LSFB) et en langue des signes catalane (LSC): les "balise-liste" et les "palm-ups". In: *4th International Symposium "Discourse markers in Romance languages: A contrastive approach"*, Heidelberg, Germany.
- Gilquin, G., S. De Cock, and S. Granger. 2010. *LINDSEI Louvain International Database of Spoken English Interlanguage. Handbook and CD-ROM*. Louvain-la-Neuve, Belgique: Presses Universitaires de Louvain.
- Götz, S. 2013. *Fluency in native and nonnative English speech*. Amsterdam: John Benjamins.
- Grosman, I. in press. "How do French humorists manage their persona across situations? A corpus study on their prosodic variation". In: *Metapragmatics of humor: Current research trends*. Ed. by L. Ruiz-Gurillo. Amsterdam: John Benjamins.
- Hansen, M.-B. M. 2006. "A dynamic polysemy approach to the lexical semantics of discourse markers (with an exemplary analysis of French toujours)". In: *Approaches to discourse particles*. Ed. by K. Fischer. Amsterdam: Elsevier, 21–41.

- Henry, S. and B. Pallaud. 2003. "Word fragments and repeats in spontaneous spoken French". In: *Gothenburg Papers in Theoretical Linguistics 90*. Ed. by R. Eklund, 77 – 80.
- Kimmelman, V. 2013. "Doubling in RSL and NGT: a Pragmatic Account". In: *Interdisciplinary Studies on Information Structure* 17, 99–118.
- Kurt, K. 2012. "Pedagogic corpora for content and language integrated learning. Insight from the BACKBONE Project". In: *The Eurocall Review* 20.2, 3–22.
- Little, D. R., R. Oehmen, J. Dunn, K. Hird, and K. Kirsner. Mar. 2013. "Fluency Profiling System: An automated system for analyzing the temporal properties of speech". In: *Behavioral Research Methods* 45.1, 191–202.
- Loon, E. van. 2012. "What's in the palm of your hands? Discourse functions of Palm-up in Sign Language of the Netherlands". MA thesis. University of Amsterdam.
- Meteer, M. 1995. "Dysfluency annotation stylebook for the Switchboard corpus. Linguistic Data Consortium."
- Meurant, L. 2015. *Corpus LSFB. A digital open access corpus of movies and annotation in French Belgian Sign Language (LSFB)*. Laboratoire de Langue des signes de Belgique francophone (LSFB-Lab), FRS-F.N.R.S et Université de Namur, www.corpus-lsfb.be.
url: www.corpus-lsfb.be.
- Meurant, L. and I. Notarrigo. 2014. "Nonmanuals and markers of (dis)fluency in French Belgian Sign Language (LSFB)". In: *Proceedings of the 6th Workshop on the Representation and Processing of Sign Languages: Beyond the Manual Channel*, 135–142.
- Moniz, H., F. Batista, A. I. Mata, and I. Trancoso. 2014. "Speaking style effects in the production of disfluencies". In: *Speech Communication* 65, 20–35.
- Notarrigo, I. 2016. "Les marqueurs de (dis)fluence en Langue des Signes de Belgique Francophone (LSFB)". PhD thesis. University of Namur, Belgium.
- Notarrigo, I., L. Meurant, and M. Van Herreweghe M. and Vermeerbergen. 2016. "Repetition of signs in French Belgian Sign Language (LSFB) and Flemish Sign Language (VGT): Typology and Annotation protocol". In: *Poster présenté à TISLR 12 - 12th International Conference on Theoretical Issues in Sign Language Research, Melbourne, Australie*.
- Pallaud, B. and S. Henry. 2004. "Troncations de mots, reprises et interruption syntaxique en français parlé spontané". In: *Le poids des mots*. Ed. by G. Prunelle, C. Fairon, and A. Dister. Louvain-la-Neuve 10-12 mars 2004 / March 10-12, 2004 Gérard: Presses Universitaire de Louvain, 707–715.
- Pallaud, B., S. Rauzy, and P. Blache. 2013. "Auto-interruption et disfluences en français parlé dans quatre corpus du CID". In: *TIPA. Travaux interdisciplinaires sur la parole et le langage* 29.
- Pfau, R. and M. Steinbach. 2006. "Pluralization in sign and in speech: A cross-modal typological study". In: *Linguistic Typology* 10.2, 135–182.
- Schmidt, T. and K. Wörner. 2012. "EXMARaLDA". In: *Handbook on Corpus Phonology*. Ed. by J. Durand, G. Ulrike, and G. Kristoffersen. Oxford: Oxford University Press, 402–419.
- Shriberg, E. E. 1994. "Preliminaries to a Theory of Speech Disfluencies." PhD thesis. University of California at Berkeley.
- Sloetjes, H. and P. Wittenburg. 2008. "Annotation by category – ELAN and ISO DCR". In: *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*. url: <http://tla.mpi.nl/tools/tla-tools/elan/>.

- Sperber, D. and D. Wilson. 1995. *Relevance. Communication and Cognition [2nd Edition]*. Oxford: Blackwell.
- Strassel, S. 2003. *Simple Metadata Annotation Specification*, 1–43.
- Vasilescu, I., M. Candea, and M. Adda-decker. 2004. “Hésitations autonomes dans 8 langues: une étude acoustique et perceptive”. In: *Colloque MIDL 2004, Modélisations pour l’identification des langues et des variétés dialectales, 29-30 novembre*, 29–30.
- Vermeerbergen, M. and S. de Vriendt. 1994. “The repetition of signs in Flemish Sign Language”. In: *Perspectives on sign language usage. Proceedings of The Fifth International Symposium on Sign Language Research*. Ed. by I. Ahlgren, B. Bergman, and M. Brennan. England: University of Durham, 201–214.
- Wilbur, R. B. 2005. “A Reanalysis of Reduplication in American Sign Language”. In: *Studies on Reduplication*. Ed. by B. Hurch. Empirical Approaches to Language Typology 28. Berlin: Mouton de Gruyter, 595–623.

Appendix 3: Top-five most frequent discourse markers by register

	1 st	2 nd	3 rd	4 th	5 th
English					
conversation	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					
conversation	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 4: List of discourse markers in *DisFrEn* and their functions

English discourse markers (4249)	
and (1140)	<i>addition</i> (651), <i>specification</i> (180), <i>consequence</i> (101), <i>topic-shift</i> (41), <i>temporal</i> (27), <i>punctuation</i> (24), <i>conclusion</i> (20), <i>topic-resuming</i> (16), <i>contrast</i> (13), <i>opening boundary</i> (13), <i>enumeration</i> (10), <i>comment</i> (9), <i>concession</i> (8), <i>emphasis</i> (6), <i>quoting</i> (4), <i>addition-punctuation</i> (3), <i>opposition</i> (3), <i>punctuation-conclusion</i> (2), <i>topic-resuming-conclusion</i> (2), <i>specification-comment</i> (1), <i>motivation</i> (1), <i>enumeration-topic-resuming</i> (1), <i>punctuation-consequence</i> (1), <i>topic-shift-specification</i> (1), <i>contrast-addition</i> (1), <i>topic-resuming-specification</i> (1)
but (477)	<i>opposition</i> (203), <i>concession</i> (142), <i>contrast</i> (38), <i>topic-resuming</i> (22), <i>topic-resuming-opposition</i> (10), <i>closing boundary</i> (9), <i>topic-shift</i> (9), <i>closing boundary-opposition</i> (9), <i>topic-shift-opposition</i> (6), <i>punctuation</i> (6), <i>opening boundary</i> (6), <i>opening boundary-opposition</i> (4), <i>topic-resuming-motivation</i> (1), <i>addition</i> (1), <i>specification</i> (1), <i>enumeration-opposition</i> (1), <i>addition-opposition</i> (1), <i>topic-resuming-conclusion</i> (1), <i>cause-topic-resuming</i> (1), <i>disagreeing</i> (1), <i>cause-topic-shift</i> (1), <i>emphasis</i> (1), <i>punctuation-opposition</i> (1), <i>exception</i> (1), <i>reformulation</i> (1)
so (429)	<i>conclusion</i> (198), <i>consequence</i> (123), <i>specification</i> (25), <i>topic-shift</i> (17), <i>topic-resuming</i> (16), <i>topic-resuming-conclusion</i> (16), <i>closing boundary-conclusion</i> (10), <i>closing boundary</i> (7), <i>punctuation</i> (3), <i>reformulation</i> (3), <i>addition</i> (2), <i>opening boundary-conclusion</i> (2), <i>consequence-specification</i> (1), <i>punctuation-conclusion</i> (1), <i>topic-resuming-consequence</i> (1), <i>emphasis</i> (1), <i>enumeration</i> (1), <i>motivation</i> (1), <i>opening boundary</i> (1)
well (304)	<i>opening boundary</i> (177), <i>reformulation</i> (26), <i>punctuation</i> (20), <i>disagreeing</i> (15), <i>quoting</i> (12), <i>topic-shift</i> (10), <i>disagreeing-opening boundary</i> (7), <i>agreeing</i> (6), <i>emphasis</i> (4), <i>comment</i> (4), <i>specification</i> (3), <i>conclusion</i> (3), <i>punctuation-conclusion</i> (3), <i>topic-resuming</i> (2), <i>disagreeing-reformulation</i> (2), <i>face-saving</i> (2), <i>opening boundary-motivation</i> (2), <i>motivation-reformulation</i> (1), <i>topic-resuming-reformulation</i> (1), <i>disagreeing-punctuation</i> (1), <i>punctuation-reformulation</i> (1), <i>opening boundary-specification</i> (1), <i>comment-reformulation</i> (1)
you know (196)	<i>monitoring</i> (180), <i>quoting</i> (3), <i>monitoring-specification</i> (3), <i>monitoring-closing boundary</i> (2), <i>monitoring-quoting</i> (2), <i>monitoring-topic-shift</i> (1), <i>reformulation</i> (1), <i>monitoring-punctuation</i> (1), <i>monitoring-reformulation</i> (1), <i>face-saving</i> (1), <i>face-saving-monitoring</i> (1)

if (195)	<i>condition (132), relevance (55), motivation (3), concession (2), temporal (1), cause-relevance (1), contrast-motivation (1)</i>
because (190)	<i>cause (98), motivation (89), specification (1), topic-resuming-motivation (1), condition (1), motivation-specification (1)</i>
I mean (174)	<i>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)</i>
when (129)	<i>temporal (120), relevance (3), cause (3), condition (2), concession (1)</i>
actually (97)	<i>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)</i>
then (94)	<i>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)</i>
and then (70)	<i>temporal (41), addition (12), enumeration (10), consequence (3), topic-shift (1), temporal-concession (1), concession (1), contrast-enumeration (1)</i>
or (65)	<i>alternative (45), reformulation (15), alternative-enumeration (2), alternative-punctuation (1), alternative-closing boundary (1), alternative-ellipsis (1)</i>
sort of (60)	<i>approximation (53), punctuation (2), emphasis (2), punctuation-approximation (1), face-saving (1), face-saving-approximation (1)</i>
now (40)	<i>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)</i>
as (32)	<i>temporal (22), cause (4), cause-temporal (3), motivation (2), condition (1)</i>
right (31)	<i>monitoring, agreeing (6), closing boundary (4), quoting (2), opening boundary (1), agreeing-punctuation (1), monitoring-closing boundary (1)</i>

kind of (31)	<i>approximation (26), face-saving (2), punctuation-approximation (2), approximation-specification (1)</i>
though (30)	<i>opposition (13), concession (12), contrast (4), topic-shift-opposition (1)</i>
in fact (29)	<i>specification (8), comment (7), reformulation (3), opposition (3), emphasis (3), topic-shift (1), comment-motivation (1), topic-shift-conclusion (1), concession (1), addition (1)</i>
yeah (27)	<i>agreeing (13), monitoring (3), topic-resuming (3), agreeing-topic-resuming (3), agreeing-closing boundary (2), agreeing-opening boundary (2), closing boundary (1)</i>
okay (34)	<i>monitoring (18), agreeing (3), monitoring-closing boundary (7), agreeing-closing boundary (2), closing boundary (2), opening boundary (1), agreeing-topic-resuming (1)</i>
and so on (19)	<i>ellipsis (14), closing boundary (4), ellipsis-closing boundary (1)</i>
like (16)	<i>approximation (9), specification (3), face-saving-approximation (2), punctuation (2)</i>
although (16)	<i>concession (15), opposition (1)</i>
therefore (16)	<i>consequence (12), conclusion (4)</i>
for example (16)	<i>specification (16)</i>
anyway (15)	<i>topic-resuming (8), closing boundary (3), topic-shift (1), emphasis (1), reformulation (1), opposition (1)</i>
so that (14)	<i>consequence (12), temporal (1), conclusion (1)</i>
indeed (13)	<i>motivation (3), specification (3), comment (2), opposition (1), reformulation (1), comment-specification (1), agreeing (1), emphasis (1)</i>
yes (13)	<i>agreeing-topic-resuming (4), topic-resuming (3), agreeing (3), monitoring (1), agreeing-opening boundary (1), closing boundary (1)</i>
if you like (12)	<i>approximation (7), face-saving (2), monitoring (2), reformulation (1)</i>
since (11)	<i>temporal (8), cause (3)</i>
before (11)	<i>temporal (11)</i>
while (10)	<i>concession (6), temporal (2), contrast-temporal (1), contrast (1)</i>
even if (9)	<i>concession (5), relevance (3), opposition (1)</i>
unless (9)	<i>exception (8), alternative (1)</i>
oh (9)	<i>quoting (9)</i>

you see (8)	<i>monitoring (7), monitoring-specification (1)</i>
for instance (8)	<i>specification (8)</i>
after (8)	<i>temporal (8)</i>
once (8)	<i>temporal</i>
say (8)	<i>specification (6), approximation (2)</i>
however (7)	<i>opposition (3), contrast (3), concession (1)</i>
until (7)	<i>temporal (7)</i>
whereas (7)	<i>contrast (6), opposition (1)</i>
for (6)	<i>cause (5), motivation (1)</i>
etcetera (6)	<i>ellipsis (4), ellipsis-approximation (1), closing boundary (1)</i>
meanwhile (6)	<i>temporal (3), temporal-topic-shift (2), topic-shift (1)</i>
in other words (4)	<i>reformulation (4)</i>
yet (4)	<i>concession</i>
look (4)	<i>quoting (2), face-saving (2)</i>
as soon as (4)	<i>temporal (4)</i>
by the way (4)	<i>topic-shift (2), comment (2)</i>
alright (4)	<i>monitoring (4)</i>
whilst (3)	<i>contrast (2), temporal (1)</i>
either (3)	<i>alternative (2), alternative-enumeration (1)</i>
first (3)	<i>enumeration (3)</i>
and things (3)	<i>ellipsis (3)</i>
as it were (3)	<i>approximation (3)</i>
otherwise (3)	<i>alternative (3)</i>
nevertheless (3)	<i>concession (2), topic-shift-opposition (1)</i>
see (3)	<i>monitoring (3)</i>
no (3)	<i>agreeing (2), disagreeing-topic-resuming (1)</i>
listen (2)	<i>monitoring (2)</i>
even though (2)	<i>concession (2)</i>
as long as (2)	<i>temporal (1), condition (1)</i>

I suppose (2)	<i>approximation (1), agreeing (1)</i>
plus (2)	<i>addition (2)</i>
provided (2)	<i>condition (2)</i>
first of all (2)	<i>enumeration (2)</i>
or something (2)	<i>approximation (2)</i>
having said that (2)	<i>opposition (2)</i>
in addition (1)	<i>addition (1)</i>
finally (1)	<i>enumeration (1)</i>
where (1)	<i>contrast (1)</i>
considering (1)	<i>motivation (1)</i>
but then (1)	<i>opposition (1)</i>
second (1)	<i>enumeration (1)</i>
whenever (1)	<i>temporal (1)</i>
and that kind of stuff (1)	<i>ellipsis (1)</i>
only (1)	<i>exception (1)</i>
secondly (1)	<i>enumeration (1)</i>
I don't know (1)	<i>specification (1)</i>
insofar as (1)	<i>motivation (1)</i>
after all (1)	<i>specification (1)</i>
on the other hand (1)	<i>opposition (1)</i>
and still (1)	<i>concession (1)</i>
albeit (1)	<i>concession (1)</i>
till (1)	<i>temporal (1)</i>
instead (1)	<i>alternative (1)</i>

French discourse markers (4494)	
et (869)	<i>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)</i>
mais (540)	<i>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)</i>
donc (291)	<i>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)</i>
alors (271)	<i>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)</i>
hein (260)	<i>monitoring (256), face-saving (2), disagreeing (1), ellipsis (1)</i>
quoi (239)	<i>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)</i>
parce que (216)	<i>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)</i>

ben (183)	<i>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)</i>
enfin (157)	<i>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)</i>
quand (133)	<i>temporal (116), relevance (8), condition (4), motivation (2), specification (1), cause (1), opposition (1)</i>
si (119)	<i>condition (80), relevance (34), cause (2), concession (1), temporal (1), motivation (1)</i>
bon (98)	<i>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)</i>
et puis (97)	<i>temporal (40), addition (32), topic-shift (11), enumeration (7), consequence (2), specification (2), conclusion (1), contrast-enumeration (1), opposition (1)</i>
tu vois (58)	<i>monitoring (53), face-saving (2), opening boundary (1), specification (1), monitoring-specification (1)</i>
voilà (50)	<i>closing boundary (39), agreeing (2), punctuation (2), emphasis (2), quoting (1), topic-resuming (1), conclusion (1), opening boundary (1), face-saving (1)</i>
en fait (45)	<i>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)</i>
par exemple (43)	<i>specification (43), topic-shift (2), closing boundary-specification (1)</i>
etcetera (41)	<i>ellipsis (35), approximation (2), closing boundary (2), monitoring (1), face-saving (1)</i>
puisque (32)	<i>motivation (20), cause (12)</i>
d'ailleurs (32)	<i>comment (24), specification (3), topic-shift (2), comment-specification (1), opposition (1), emphasis (1)</i>
bon ben (31)	<i>punctuation (18), opening boundary (6), face-saving (2), opposition (1), conclusion (1), emphasis (1), ellipsis-conclusion (1), opening boundary-consequence (1)</i>
eh bien (30)	<i>punctuation (16), opening boundary (6), topic-resuming (2), consequence (2), punctuation-conclusion (1), disagreeing (1), conclusion (1), emphasis (1)</i>

oui (30)	<i>agreeing</i> (28), <i>monitoring</i> (1), <i>agreeing-opening</i> (1)
puis (26)	<i>temporal</i> (12), <i>addition</i> (6), <i>enumeration</i> (3), <i>specification</i> (2), <i>consequence</i> (1), <i>topic-shift</i> (1), <i>temporal-addition</i> (1)
ou (26)	<i>alternative</i> (20), <i>reformulation</i> (6)
je veux dire (26)	<i>reformulation</i> (12), <i>specification</i> (8), <i>approximation</i> (3), <i>punctuation</i> (1), <i>emphasis</i> (1), <i>face-saving</i> (1)
et tout ça (25)	<i>ellipsis</i> (25)
alors que (25)	<i>temporal</i> (8), <i>concession</i> (6), <i>concession-temporal</i> (4), <i>contrast</i> (4), <i>opposition</i> (2), <i>consequence</i> (1)
comme (21)	<i>cause</i> (17), <i>motivation</i> (4)
eh ben (21)	<i>opening boundary</i> (8), <i>punctuation</i> (6), <i>topic-shift</i> (2), <i>topic-resuming</i> (2), <i>conclusion</i> (1), <i>agreeing-opening boundary</i> (1), <i>specification</i> (1)
écoutez (19)	<i>monitoring</i> (16), <i>quoting</i> (1), <i>face-saving</i> (1), <i>face-saving-quoting</i> (1)
au fond (17)	<i>specification</i> (8), <i>conclusion</i> (3), <i>opposition</i> (2), <i>comment</i> (1), <i>reformulation</i> (1), <i>consequence</i> (1), <i>emphasis</i> (1)&
je dirais (16)	<i>approximation</i> (15), <i>face-saving</i> (1)
voilà quoi (15)	<i>closing boundary</i> (12), <i>ellipsis</i> (1), <i>face-saving</i> (1), <i>consequence</i> (1)
vous savez (15)	<i>monitoring</i> (13), <i>monitoring-topic-shift</i> (1), <i>face-saving</i> (1)
c'est-à-dire (14)	<i>specification</i> (10), <i>conclusion</i> (2), <i>reformulation</i> (2)
car (14)	<i>motivation</i> (7), <i>cause</i> (7)
pour que (11)	<i>consequence</i> (11)
ouais (11)	<i>agreeing</i> (10), <i>monitoring</i> (1)
je vais dire (11)	<i>approximation</i> (10), <i>approximation-reformulation</i> (1)
ou bien (11)	<i>alternative</i> (11)
tandis que (10)	<i>contrast</i> (8), <i>temporal</i> (2)
pourtant (10)	<i>concession</i> (9), <i>opposition</i> (1)
c'est-à-dire que (10)	<i>specification</i> (6), <i>emphasis</i> (2), <i>opening boundary-reformulation</i> (1), <i>face-saving-emphasis</i> (1)
enfin bon (9)	<i>closing boundary</i> (4), <i>opposition</i> (3), <i>topic-shift-opposition</i> (1), <i>conclusion</i> (1)
dès que (9)	<i>temporal</i> (9)
par contre (9)	<i>opposition</i> (5), <i>addition-opposition</i> (2), <i>emphasis</i> (1), <i>contrast</i> (1)
disons (9)	<i>approximation</i> (4), <i>topic-resuming</i> (1), <i>reformulation</i> (1), <i>specification</i> (1°), <i>emphasis</i> (1), <i>punctuation-approximation</i> (1)
d'abord (9)	<i>enumeration</i> (9)
non (9)	<i>monitoring</i> (4), <i>topic-resuming</i> (2), <i>disagreeing-topic-resuming</i> (2), <i>agreeing-topic-resuming</i> (1)
sinon (8)	<i>alternative</i> (5), <i>exception</i> (2), <i>topic-shift</i> (1)
tiens (8)	<i>quoting</i> (7), <i>topic-shift</i> (1)
même si (8)	<i>concession</i> (8)
soit (7)	<i>alternative</i> (7)
si tu veux (7)	<i>monitoring-approximation</i> (4), <i>approximation</i> (2), <i>monitoring</i> (1)
lorsque (7)	<i>temporal</i> (6), <i>cause</i> (1)

okay (8)	<i>agreeing</i> (8)
bien (7)	<i>topic-shift</i> (2), <i>opening boundary</i> (2), <i>topic-resuming</i> (1), <i>closing boundary</i> (1), <i>quoting</i> (1)
tu sais (6)	<i>monitoring</i> (6)
et tout (6)	<i>ellipsis</i> (5), <i>ellipsis-approximation</i> (1)
tout ça (5)	<i>ellipsis</i> (5)
à ce moment-là (5)	<i>consequence</i> (3), <i>exception</i> (1), <i>conclusion</i> (1)
en conséquence (5)	<i>consequence</i> (5)
maintenant (5)	<i>opposition</i> (2), <i>topic-shift</i> (1), <i>concession</i> (1), <i>enumeration</i> (1)
d'accord (5)	<i>monitoring</i> (4), <i>agreeing-closing boundary</i> (1)
si vous voulez (5)	<i>approximation</i> (3), <i>reformulation</i> (1), <i>monitoring-specification</i> (1)
entre guillemets (5)	<i>approximation</i> (4), <i>emphasis</i> (1)
en tout cas (5)	<i>reformulation</i> (4), <i>emphasis</i> (1)
après (4)	<i>enumeration</i> (2), <i>topic-shift</i> (1), <i>opposition</i> (1)
écoute (4)	<i>monitoring</i> (3), <i>face-saving</i> (1)
autrement (3)	<i>exception</i> (2), <i>alternative</i> (1)
à ce propos (3)	<i>comment</i> (2), <i>topic-shift</i> (1)
un (3)	<i>enumeration</i> (3)
en effet (3)	<i>specification</i> (3)
du coup (3)	<i>consequence</i> (3)
en plus (3)	<i>addition</i> (2), <i>enumeration</i> (1)
vu que (3)	<i>motivation</i> (2), <i>cause</i> (1)
mais bon (3)	<i>opposition</i> (2), <i>punctuation</i> (1)
savez (3)	<i>monitoring</i> (2), <i>specification</i> (1)
vous voyez (3)	<i>monitoring</i> (3)
du moins (3)	<i>emphasis</i> (2), <i>reformulation</i> (1)
ainsi (3)	<i>specification</i> (2), <i>consequence</i> (1)
or (3)	<i>concession</i> (3)
seulement (2)	<i>opposition</i> (1), <i>concession</i> (1)
quoique (2)	<i>reformulation</i> (1), <i>concession</i> (1)
bien que (2)	<i>opposition</i> (1), <i>concession</i> (1)
ah (2)	<i>quoting</i> (2)
ou sinon (2)	<i>exception</i> (1), <i>alternative</i> (1)
d'un autre côté (2)	<i>opposition</i> (1), <i>contrast</i> (1)
enfin bref (2)	<i>ellipsis</i> (1), <i>closing boundary</i> (1)
encore que (2)	<i>reformulation</i> (2)
ou quoi (2)	<i>ellipsis</i> (2)

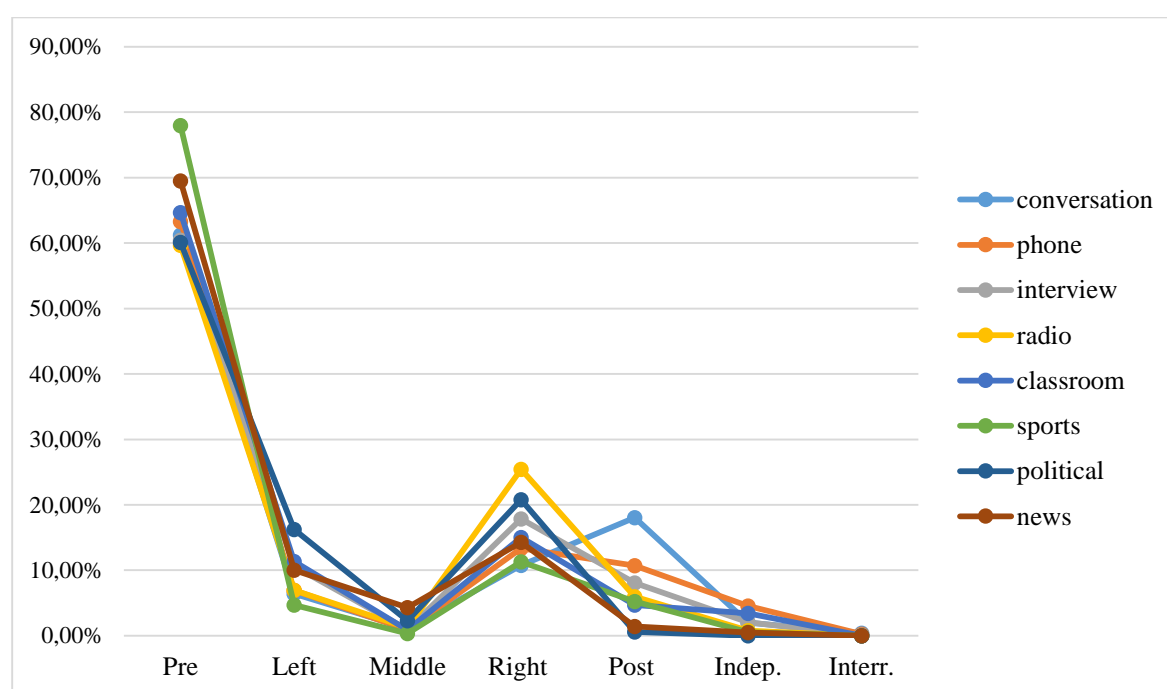
déjà (2)	<i>enumeration (1), comment (1)</i>
sauf que (2)	<i>exception (2)</i>
ça va (2)	<i>agreeing (2)</i>
voyez (2)	<i>monitoring (2)</i>
tant que (2)	<i>temporal (1), condition-temporal (1)</i>
ou alors (2)	<i>alternative (1)</i>
m'enfin (2)	<i>opposition (2)</i>
comme ça (1)	<i>approximation (1)</i>
d'autre part (1)	<i>topic-shift (1)</i>
de un (1)	<i>enumeration (1)</i>
au contraire (1)	<i>opposition (1)</i>
bref (1)	<i>topic-resuming-conclusion (1)</i>
du reste (1)	<i>comment (1)</i>
dès lors (1)	<i>temporal (1)</i>
du temps où (1)	<i>temporal (1)</i>
genre (1)	<i>specification (1)</i>
maintenant que (1)	<i>cause-temporal (1)</i>
de même (1)	<i>addition (1)</i>
à part cela (1)	<i>topic-shift (1)</i>
par conséquent (1)	<i>consequence (1)</i>
dis (1)	<i>monitoring (1)</i>
on va dire (1)	<i>approximation (1)</i>
quand même (1)	<i>emphasis (1)</i>
à propos (1)	<i>topic-shift (1)</i>
mais enfin (1)	<i>opposition (1)</i>
deuxièmement (1)	<i>enumeration (1)</i>
depuis que (1)	<i>temporal (1)</i>
cependant (1)	<i>opposition (1)</i>
de sorte que (1)	<i>consequence (1)</i>
bah (1)	<i>agreeing (1)</i>
néanmoins (1)	<i>néanmoins (1)</i>
ouais ouais (1)	<i>agreeing (1)</i>
boh (1)	<i>opening boundary (1)</i>
par ailleurs (1)	<i>topic-shift (1)</i>
autrement dit (1)	<i>reformulation (1)</i>
du moment que (1)	<i>condition (1)</i>
effectivement (1)	<i>comment (1)</i>

Appendix 5: Macro-syntactic position of DMs by register

Appendix 5.1: Distribution of DMs across macro-syntactic slots by register

	Pre-field	Left	Middle	Right	Post-field	Indep.	Interr.
conversation	1515	159	28	266	447	49	10
phone	721	79	10	152	122	52	3
interview	1431	251	22	423	191	48	2
radio	549	64	11	234	55	7	0
classroom	456	80	6	106	33	24	0
sports	449	27	2	65	30	3	0
political	211	57	8	73	2	0	0
news	146	21	9	30	3	1	0
Total	5478	738	96	1349	883	184	15

Appendix 5.2: Proportions of macro-syntactic slots by register



Appendix 6: Mapping of domains and (non-)relational type by register

	Sequential	Rhetorical	Ideational	Interpersonal
conversation	755	701	475	430
RDM	334	614	475	0
NRDM	421	87	0	430
phone	367	312	224	163
RDM	131	282	224	0
NRDM	236	30	0	163
interview	648	799	594	251
RDM	420	727	594	0
NRDM	228	72	0	251
radio	297	356	182	69
RDM	209	314	182	0
NRDM	88	42	0	69
classroom	211	238	168	58
RDM	131	213	168	0
NRDM	80	25	0	58
sports	215	121	200	23
RDM	189	118	200	0
NRDM	26	3	0	23
political	112	70	148	5
RDM	103	70	148	0
NRDM	9	0	0	5
news	74	52	73	0
RDM	60	52	73	0
NRDM	14	0	0	0
Total	2679	2649	2064	999

Appendix 7: List of functions in *DisFrEn* and their discourse markers⁸⁵

	English DMs	French DMs
Sequential functions (2680)		
Addition (1238)	<i>and</i> (651), <i>and then</i> (12), <i>now</i> (8), <i>so</i> (2), <i>plus</i> (2), <i>in fact</i> (1), <i>but</i> (1), <i>in addition</i> (1), <i>I mean</i> (1)	<i>et</i> (498), <i>et puis</i> (32), <i>alors</i> (12), <i>mais</i> (7), <i>puis</i> (6), <i>en plus</i> (2), <i>de même</i> (1), <i>parce que</i> (1)
Opening boundary (404)	<i>well</i> (177), <i>and</i> (13), <i>I mean</i> (11), <i>but</i> (6), <i>now</i> (4), <i>right</i> (1), <i>so</i> (1), <i>okay</i> (1)	<i>ben</i> (85), <i>alors</i> (40), <i>mais</i> (25), <i>eh ben</i> (8), <i>bon</i> (6), <i>et</i> (6), <i>bon ben</i> (6), <i>eh bien</i> (6), <i>bien</i> (2), <i>parce que</i> (1), <i>donc</i> (1), <i>tu vois</i> (1), <i>voilà</i> (1), <i>boh</i> (1), <i>en fait</i> (1)
Punctuation (284)	<i>I mean</i> (26), <i>and</i> (24), <i>well</i> (20), <i>but</i> (6), <i>so</i> (3), <i>sort of</i> (2), <i>like</i> (2), <i>now</i> (2), <i>actually</i> (2)	<i>ben</i> (41), <i>bon</i> (38), <i>et</i> (23), <i>quoi</i> (21), <i>mais</i> (20), <i>bon ben</i> (18), <i>eh bien</i> (16), <i>alors</i> (6), <i>eh ben</i> (6), <i>voilà</i> (2), <i>donc</i> (2), <i>mais bon</i> (1), <i>je veux dire</i> (1), <i>en fait</i> (1), <i>enfin</i> (1)
Topic-shift (271)	<i>and</i> (41), <i>so</i> (17), <i>now</i> (12), <i>well</i> (10), <i>but</i> (9), <i>then</i> (5), <i>actually</i> (2), <i>by the way</i> (2), <i>in fact</i> (1), <i>and then</i> (1), <i>meanwhile</i> (1), <i>anyway</i> (1)	<i>et</i> (102), <i>mais</i> (19), <i>alors</i> (12), <i>et puis</i> (11), <i>bon</i> (3), <i>par exemple</i> (2), <i>en fait</i> (2), <i>bien</i> (2), <i>d'ailleurs</i> (2), <i>eh ben</i> (2), <i>ben</i> (1), <i>sinon</i> (1), <i>après</i> (1), <i>à propos</i> (1), <i>à ce propos</i> (1), <i>à part cela</i> (1), <i>par ailleurs</i> (1), <i>puis</i> (1), <i>maintenant</i> (1), <i>d'autre part</i> (1), <i>tiens</i> (1), <i>enfin</i> (1)
Topic-resuming (167)	<i>but</i> (22), <i>and</i> (16), <i>so</i> (16), <i>anyway</i> (8), <i>now</i> (7), <i>then</i> (5), <i>I mean</i> (4), <i>yes</i> (3), <i>yeah</i> (3), <i>well</i> (2)	<i>mais</i> (21), <i>donc</i> (20), <i>et</i> (13), <i>bon</i> (7), <i>alors</i> (7), <i>ben</i> (2), <i>eh ben</i> (2), <i>non</i> (2), <i>eh bien</i> (2), <i>enfin</i> (2), <i>disons</i> (1), <i>voilà</i> (1), <i>bien</i> (1)

⁸⁵ 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 4 so that the information is not lost.

Closing boundary (166)	<i>but</i> (9), <i>so</i> (7), <i>and so on</i> (4), <i>right</i> (4), <i>anyway</i> (3), <i>okay</i> (2), <i>etcetera</i> (1), <i>yeah</i> (1), <i>yes</i> (1), <i>now</i> (1)	<i>quoi</i> (46), <i>voilà</i> (39), <i>bon</i> (14), <i>voilà quoi</i> (12), <i>donc</i> (6), <i>enfin</i> (5), <i>enfin bon</i> (4), <i>mais</i> (3), <i>bien</i> (1), <i>okay</i> (2), <i>enfin bref</i> (1)
Enumeration (83)	<i>and then</i> (10), <i>and</i> (10), <i>then</i> (6), <i>first</i> (3), <i>first of all</i> (2), <i>secondly</i> (1), <i>finally</i> (1), <i>second</i> (1), <i>so</i> (1)	<i>et</i> (13), <i>d'abord</i> (9), <i>et puis</i> (7), <i>alors</i> (4), <i>puis</i> (3), <i>un</i> (3), <i>après</i> (2), <i>donc</i> (1), <i>déjà</i> (1), <i>deuxièmement</i> (1), <i>en plus</i> (1), <i>de un</i> (1), <i>enfin</i> (1), <i>maintenant</i> (1)
Quoting (66)	<i>well</i> (12), <i>oh</i> (9), <i>and</i> (4), <i>you know</i> (3), <i>right</i> (2), <i>look</i> (2)	<i>ben</i> (10), <i>mais</i> (8), <i>tiens</i> (7), <i>bon</i> (3), <i>ah</i> (2), <i>bien</i> (1), <i>et</i> (1), <i>écoutez</i> (1), <i>voilà</i> (1)
Emphasis	<i>well</i> (1)	
Rhetorical functions (2650)		
Specification (626)	<i>and</i> (180), <i>I mean</i> (64), <i>so</i> (25), <i>actually</i> (24), <i>for example</i> (16), <i>for instance</i> (8), <i>in fact</i> (8), <i>say</i> (6), <i>like</i> (3), <i>well</i> (3), <i>indeed</i> (3), <i>but</i> (1), <i>I don't know</i> (1), <i>after all</i> (1), <i>then</i> (1), <i>because</i> (1)	<i>et</i> (78), <i>alors</i> (42), <i>par exemple</i> (40), <i>donc</i> (25), <i>en fait</i> (18), <i>c'est-à-dire</i> (10), <i>mais</i> (9), <i>enfin</i> (9), <i>au fond</i> (8), <i>je veux dire</i> (8), <i>ben</i> (6), <i>c'est-à-dire que</i> (6), <i>en effet</i> (3), <i>d'ailleurs</i> (3), <i>puis</i> (2), <i>bon</i> (2), <i>ainsi</i> (2), <i>et puis</i> (2), <i>savez</i> (1), <i>tu vois</i> (1), <i>eh ben</i> (1), <i>genre</i> (1), <i>parce que</i> (1), <i>disons</i> (1), <i>quoi</i> (1)
Opposition (546)	<i>but</i> (203), <i>actually</i> (20), <i>though</i> (13), <i>in fact</i> (3), <i>and</i> (3), <i>however</i> (3), <i>now</i> (2), <i>having said that</i> (2), <i>but then</i> (1), <i>then</i> (1), <i>anyway</i> (1), <i>even if</i> (1), <i>whereas</i> (1), <i>on the other hand</i> (1), <i>although</i> (1), <i>indeed</i> (1)	<i>mais</i> (235), <i>enfin</i> (7), <i>par contre</i> (5), <i>bon</i> (4), <i>en fait</i> (4), <i>et</i> (3), <i>enfin bon</i> (3), <i>mais bon</i> (2), <i>maintenant</i> (2), <i>alors</i> (2), <i>m'enfin</i> (2), <i>alors que</i> (2), <i>au fond</i> (2), <i>ben</i> (2), <i>après</i> (1), <i>et puis</i> (1), <i>quand</i> (1), <i>cependant</i> (1), <i>au contraire</i> (1), <i>d'ailleurs</i> (1), <i>néanmoins</i> (1), <i>pourtant</i> (1), <i>donc</i> (1), <i>seulement</i> (1), <i>bien que</i> (1), <i>bon ben</i> (1), <i>d'un autre côté</i> (1), <i>mais enfin</i> (1)

Conclusion (510)	<i>so</i> (198), <i>then</i> (36), <i>and</i> (20), <i>I mean</i> (5), <i>therefore</i> (4), <i>well</i> (3), <i>so that</i> (1), <i>now</i> (1)	<i>donc</i> (146), <i>alors</i> (42), <i>quoi</i> (18), <i>et</i> (12), <i>enfin</i> (11), <i>au fond</i> (3), <i>c'est-à-dire</i> (2), <i>eh bien</i> (1), <i>bon ben</i> (1), <i>bon</i> (1), <i>à ce moment-là</i> (1), <i>et puis</i> (1), <i>voilà</i> (1), <i>eh ben</i> (1), <i>enfin bon</i> (1)
Motivation (259)	<i>because</i> (89), <i>I mean</i> (4), <i>indeed</i> (3), <i>if</i> (3), <i>as</i> (2), <i>and</i> (1), <i>insofar as</i> (1), <i>so</i> (1), <i>for</i> (1), <i>considering</i> (1)	<i>parce que</i> (113), <i>puisque</i> (20), <i>car</i> (7), <i>comme</i> (4), <i>quand</i> (2), <i>vu que</i> (2), <i>si</i> (1), <i>enfin</i> (1), <i>quoi</i> (1), <i>ben</i> (1), <i>mais</i> (1)
Reformulation (248)	<i>I mean</i> (41), <i>well</i> (26), <i>or</i> (15), <i>in other words</i> (4), <i>so</i> (3), <i>in fact</i> (3), <i>actually</i> (3), <i>but</i> (1), <i>indeed</i> (1), <i>if you like</i> (1), <i>you know</i> (1), <i>anyway</i> (1)	<i>enfin</i> (99), <i>je veux dire</i> (12), <i>ou</i> (6), <i>en tout cas</i> (4), <i>quoi</i> (4), <i>en fait</i> (3), <i>mais</i> (3), <i>donc</i> (3), <i>bon</i> (2), <i>encore que</i> (2), <i>c'est-à-dire</i> (2), <i>au fond</i> (1), <i>ben</i> (1), <i>si vous voulez</i> (1), <i>du moins</i> (1), <i>quoique</i> (1), <i>autrement dit</i> (1), <i>alors</i> (1), <i>disons</i> (1)
Approximation (154)	<i>sort of</i> (53), <i>kind of</i> (26), <i>like</i> (9), <i>if you like</i> (7), <i>as it were</i> (3), <i>say</i> (2), <i>or something</i> (2), <i>I suppose</i> (1)	<i>je dirais</i> (15), <i>je vais dire</i> (10), <i>entre guillemets</i> (4), <i>disons</i> (4), <i>quoi</i> (3), <i>si vous voulez</i> (3), <i>je veux dire</i> (3), <i>si tu veux</i> (2), <i>etcetera</i> (2), <i>enfin</i> (2), <i>on va dire</i> (1), <i>ben</i> (1), <i>comme ça</i> (1)
Emphasis (108)	<i>actually</i> (9), <i>and</i> (6), <i>well</i> (3), <i>I mean</i> (3), <i>in fact</i> (3), <i>sort of</i> (2), <i>then</i> (2), <i>so</i> (1), <i>anyway</i> (1), <i>indeed</i> (1), <i>but</i> (1)	<i>alors</i> (16), <i>mais</i> (12), <i>ben</i> (9), <i>enfin</i> (7), <i>et</i> (6), <i>en fait</i> (5), <i>donc</i> (3), <i>c'est-à-dire que</i> (2), <i>voilà</i> (2), <i>du moins</i> (2), <i>bon ben</i> (1), <i>parce que</i> (1), <i>disons</i> (1), <i>je veux dire</i> (1), <i>eh bien</i> (1), <i>par contre</i> (1), <i>d'ailleurs</i> (1), <i>quand même</i> (1), <i>au fond</i> (1), <i>entre guillemets</i> (1), <i>bon</i> (1), <i>en tout cas</i> (1)
Relevance (103)	<i>if</i> (55), <i>when</i> (3), <i>even if</i> (3)	<i>si</i> (34), <i>quand</i> (8)

Comment (96)	<i>actually (18), and (9), in fact (7), I mean (4), well (4), by the way (2), indeed (2), now (1)</i>	<i>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)</i>
Ideational functions (2064)		
Temporal (500)	<i>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)</i>	<i>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)</i>
Consequence (478)	<i>so (123), and (101), then (25), so that (12), therefore (12), and then (3), actually (1),</i>	<i>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)</i>
Concession (368)	<i>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)</i>	<i>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)</i>
Cause (247)	<i>because (97), for (5), as (4), since (3), when (3)</i>	<i>parce que (94), comme (17), puisque (12), car (7), si (2), vu que (1), lorsque (1), quand (1)</i>
Condition (223)	<i>if (132), provided (2), when (2), because (1), as long as (1), as (1)</i>	<i>si (79), quand (4), du moment que (1)</i>
Contrast (129)	<i>but (38), and (13), whereas (6), though (4), however (3), whilst (2), then (2), where (1), while (1), now (1)</i>	<i>mais (25), et (18), tandis que (8), alors que (4), d'un autre côté (1), par contre (1), alors (1)</i>
Alternative (101)	<i>or (45), otherwise (3), either (2), instead (1), unless (1), actually (1)</i>	<i>ou (20), ou bien (11), soit (7), sinon (5), ou alors (2), ou sinon (1), autrement (1), et (1)</i>

Exception (18)	<i>unless</i> (8), <i>but</i> (1), <i>only</i> (1)	<i>sauf que</i> (2), <i>autrement</i> (2), <i>sinon</i> (2), <i>ou sinon</i> (1), <i>à ce moment-là</i> (1)
Interpersonal functions (999)		
Monitoring (718)	<i>you know</i> (180), <i>okay</i> (18), <i>right</i> (16), <i>you see</i> (7), <i>alright</i> (4), <i>yeah</i> (3), <i>see</i> (3), <i>if you like</i> (2), <i>listen</i> (2), <i>yes</i> (1)	<i>hein</i> (256), <i>quoi</i> (112), <i>itu vois</i> (53), <i>écoutez</i> (56), <i>vous savez</i> (13), <i>tu sais</i> (6), <i>non</i> (4), <i>d'accord</i> (4), <i>écoute</i> (3), <i>donc</i> (3), <i>vous voyez</i> (3), <i>voyez</i> (2), <i>savez</i> (2), <i>si tu veux</i> (1), <i>etcetera</i> (1), <i>dis</i> (1), <i>ouais</i> (1), <i>oui</i> (1)
Ellipsis (99)	<i>and so on</i> (14), <i>etcetera</i> (4), <i>and things</i> (3), <i>and that kind of stuff</i> (1)	<i>etcetera</i> (35), <i>et tout ça</i> (25), <i>tout ça</i> (5), <i>et tout</i> (5), <i>ou quoi</i> (2), <i>hein</i> (1), <i>enfin bref</i> (1), <i>voilà quoi</i> (1), <i>et</i> (1), <i>enfin</i> (1)
Agreeing (94)	<i>yeah</i> (13), <i>well</i> (6), <i>right</i> (6), <i>yes</i> (3), <i>okay</i> (3), <i>no</i> (2), <i>I suppose</i> (1), <i>indeed</i> (1)	<i>oui</i> (28), <i>ouais</i> (10), <i>okay</i> (8), <i>bon</i> (3), <i>ben</i> (3), <i>voilà</i> (2), <i>ça va</i> (2), <i>ouais ouais</i> (1), <i>bah</i> (1), <i>mais</i> (1)
Face-saving (53)	<i>look</i> (2), <i>well</i> (2), <i>kind of</i> (2), <i>if you like</i> (2), <i>sort of</i> (1), <i>actually</i> (1), <i>you know</i> (1), <i>I mean</i> (1)	<i>quoi</i> (15), <i>bon</i> (7), <i>enfin</i> (3), <i>tu vois</i> (2), <i>bon ben</i> (2), <i>hein</i> (2), <i>vous savez</i> (1), <i>voilà</i> (1), <i>je dirais</i> (1), <i>etcetera</i> (1), <i>je veux dire</i> (1), <i>voilà quoi</i> (1), <i>écoute</i> (1), <i>donc</i> (1), <i>écoutez</i> (1), <i>alors</i> (1)
Disagreeing (35)	<i>well</i> (15), <i>actually</i> (1), <i>but</i> (1)	<i>ben</i> (8), <i>mais</i> (5), <i>quoi</i> (2), <i>hein</i> (1), <i>eh bien</i> (1), <i>enfin</i> (1)

Appendix 8: Top-five most frequent functions by register in *DisFrEn*

	1 st	2 nd	3 rd	4 th	5 th
English					
conversation	<i>addition</i>	<i>opening</i>	<i>opposition</i>	<i>monitoring</i>	<i>conclusion</i>
phone	<i>opening</i>	<i>addition</i>	<i>opposition</i>	<i>monitoring</i>	<i>conclusion</i>
interview	<i>specification</i>	<i>addition</i>	<i>consequence</i>	<i>conclusion</i>	<i>monitoring</i>
radio	<i>addition</i>	<i>specification</i>	<i>opposition</i>	<i>temporal</i>	<i>motivation</i>
classroom	<i>addition</i>	<i>conclusion</i>	<i>temporal</i>	<i>monitoring</i>	<i>opposition</i>
sports	<i>addition</i>	<i>consequence</i>	<i>temporal</i>	<i>concession</i>	<i>opposition</i>
political	<i>addition</i>	<i>temporal</i>	<i>concession</i>	<i>opposition</i>	<i>condition</i>
news	<i>addition</i>	<i>concession</i>	<i>temporal</i>	<i>opposition</i>	<i>topic-shift</i>
French					
conversation	<i>monitoring</i>	<i>addition</i>	<i>specification</i>	<i>opposition</i>	<i>reformulation</i>
phone	<i>opening</i>	<i>monitoring</i>	<i>punctuation</i>	<i>conclusion</i>	<i>addition</i>
interview	<i>addition</i>	<i>monitoring</i>	<i>specification</i>	<i>opposition</i>	<i>temporal</i>
radio	<i>addition</i>	<i>specification</i>	<i>opposition</i>	<i>motivation</i>	<i>monitoring</i>
classroom	<i>addition</i>	<i>consequence</i>	<i>monitoring</i>	<i>punctuation</i>	<i>closing</i>
sports	<i>addition</i>	<i>monitoring</i>	<i>opposition</i>	<i>consequence</i>	<i>concession</i>
political	<i>addition</i>	<i>condition</i>	<i>consequence</i>	<i>cause</i>	<i>temporal</i>
news	<i>addition</i>	<i>concession</i>	<i>opposition</i>	<i>topic-shift</i>	<i>specification</i>
Total					
conversation	<i>monitoring</i>	<i>addition</i>	<i>specification</i>	<i>opposition</i>	<i>opening</i>
phone	<i>opening</i>	<i>monitoring</i>	<i>addition</i>	<i>conclusion</i>	<i>opposition</i>
interview	<i>addition</i>	<i>specification</i>	<i>monitoring</i>	<i>consequence</i>	<i>conclusion</i>
radio	<i>addition</i>	<i>specification</i>	<i>opposition</i>	<i>motivation</i>	<i>monitoring</i>
classroom	<i>addition</i>	<i>conclusion</i>	<i>monitoring</i>	<i>temporal</i>	<i>consequence</i>
sports	<i>addition</i>	<i>consequence</i>	<i>temporal</i>	<i>concession</i>	<i>opposition</i>
political	<i>addition</i>	<i>temporal</i>	<i>condition</i>	<i>consequence</i>	<i>concession</i>
news	<i>addition</i>	<i>concession</i>	<i>opposition</i>	<i>temporal</i>	<i>topic-shift</i>

Appendix 9: Proportions of sequence types by position and domain

	initial	medial	final
Pauses (P)	86.62%	3.58%	9.80%
sequential	93.11%	2.04%	4.86%
rhetorical	89.46%	6.08%	4.46%
ideational	98.81%	0.83%	0.36%
interpersonal	25.57%	8.24%	66.19%
DMs (D)	78.86%	7.77%	13.37%
sequential	90.69%	1.80%	7.51%
rhetorical	73.81%	15.93%	10.26%
ideational	94.86%	2.83%	2.31%
interpersonal	21.27%	11.33%	67.40%
Repetitions (R)	84.79%	6.80%	8.41%
sequential	93.12%	1.06%	5.82%
rhetorical	85.25%	12.44%	2.30%
ideational	99.29%	0.71%	0.00%
interpersonal	33.33%	16.67%	50.00%
Interruptions (F)	83.19%	6.55%	10.26%
sequential	92.38%	3.81%	3.81%
rhetorical	87.20%	8.80%	4.00%
ideational	94.74%	3.51%	1.75%
interpersonal	50.00%	9.38%	40.63%
Mixed (Z)	81.22%	8.29%	10.50%
sequential	92.73%	1.82%	5.45%
rhetorical	81.82%	14.29%	3.90%
ideational	100.00%	0.00%	0.00%
interpersonal	33.33%	12.50%	54.17%
Substitutions (S)	85.63%	8.62%	5.75%
sequential	95.92%	2.04%	2.04%
rhetorical	81.97%	14.75%	3.28%
ideational	93.75%	6.25%	0.00%
interpersonal	43.75%	12.50%	43.75%
Total	83.01%	5.88%	11.11%

Marqueurs du Discours et (Dis)fluence à travers les Genres : Une Etude Contrastive Basée sur l'Usage en Anglais et en Français

Le langage oral est caractérisé par la présence d'éléments linguistiques tels que les marqueurs du discours (p. ex. *donc, alors, tu vois, parce que, quoi*) et autres phénomènes dits « disfluents » qui témoignent de la nature temporelle et non-linéaire des mécanismes cognitifs de production et de perception. Ces marqueurs, aussi appelés « fluencèmes », sont ambivalents dans leur usage en cela qu'ils peuvent être produits et interprétés tantôt comme des *signaux* remplissant une fonction stratégique (une pause ou un *mais* marquant une frontière thématique-discursive majeure), tantôt comme des *symptômes* reflétant un trouble passager (une pause ou un *mais* exprimant une hésitation). L'objectif de cette thèse est de distinguer les emplois plus ou moins stratégiques ou symptomatiques des fluencèmes sur base de leur combinaison et de leur distribution à travers différents genres de parole en anglais et en français. L'analyse porte plus spécifiquement sur les propriétés syntaxiques (catégorie grammaticale, position dans l'énoncé) et fonctionnelles des marqueurs du discours, repérés et annotés manuellement dans un corpus bilingue comparable d'environ 15 heures de parole. Notre recherche poursuit un double objectif : (i) établir un portrait exhaustif des marqueurs du discours en anglais et en français et (ii) proposer une échelle de (dis)fluence sur laquelle différentes configurations de marqueurs du discours, combinés avec d'autres fluencèmes, pourraient être diagnostiquées comme plutôt fluides ou disfluides, en suivant l'hypothèse que certains emplois sont plus ambivalents que d'autres et que cette variation dépend à la fois de facteurs linguistiques (co-texte) et de facteurs situationnels (contexte). L'analyse statistique et qualitative du corpus annoté révèle en effet de fortes associations entre forme et fonction des fluencèmes ainsi qu'une influence du genre de parole, nous permettant de distinguer, entre autres, deux types d'usages: d'une part, les marqueurs du discours à fonction structurante, en position initiale et en combinaison fréquente avec des pauses, sont utilisés même dans les registres les plus formels et occupent l'extrémité fluente de l'échelle ; d'autre part, les marqueurs à fonction interpersonnelle, en position finale et en combinaison fréquente avec des interruptions (faux départs et fragments de mots), sont caractéristiques des registres informels et correspondent à des emplois plus disfluides. Cette étude met ainsi en lumière le rôle primordial des périphéries (début et fin d'énoncé) dans l'organisation et la structure du discours oral.

Discourse Markers and (Dis)fluency across Registers: A Contrastive Usage-Based Study in English and French

Spoken language is characterized by the occurrence of linguistic devices such as discourse markers (e.g. *so, well, you know, because, I mean*) and other so-called “disfluent” phenomena, which reflect the temporal and non-linear nature of the cognitive mechanisms of language production and comprehension. These markers, also called “fluencemes”, are ambivalent in their use insofar as they can be produced and interpreted either as *signals* performing a strategic function (a pause or a *but* marking a major thematic-discursive boundary) or as *symptoms* manifesting temporary trouble (a pause or a *but* expressing a hesitation). The purpose of this thesis is to distinguish between more or less strategic or symptomatic uses of fluencemes on the basis of their combination and distribution across several registers in English and French. The analysis specifically targets the syntactic (grammatical category, position in the unit) and functional properties of discourse markers, manually identified and annotated in a bilingual, comparable corpus of about 15 hours of recordings. The present research agenda is two-fold: (i) to draw an exhaustive portrait of discourse markers in English and French and (ii) to suggest a scale of (dis)fluency against which different configurations of discourse markers, combined with other fluencemes, could be diagnosed as rather fluent or disfluent, following the hypothesis that some uses are more ambivalent than others and that this variation depends both on linguistic (co-text) and situational (context) factors. The statistical and qualitative analysis of the annotated corpus reveals strong associations between the form and function of fluencemes as well as an effect of register, which allows us to distinguish, among others, two types of uses: on the one hand, discourse markers with a structuring function, in initial position and frequently co-occurring with pauses, are frequent even in the most formal settings and occupy the fluent end of the scale; on the other hand, discourse markers with an interpersonal function, in final position and frequently co-occurring with interruptions (false starts and word fragments) are specific to informal registers and correspond to more disfluent uses. This study thus uncovers the prominent role of peripheries (beginning and end of utterances) in the organization and structure of spoken discourse.

Ludivine CRIBLE

Institut Langage & Communication
Place Blaise Pascal, 1 boîte L3.03.11
1348 Louvain-La-Neuve, Belgique
www.uclouvain.be/ilc