DisFrEn

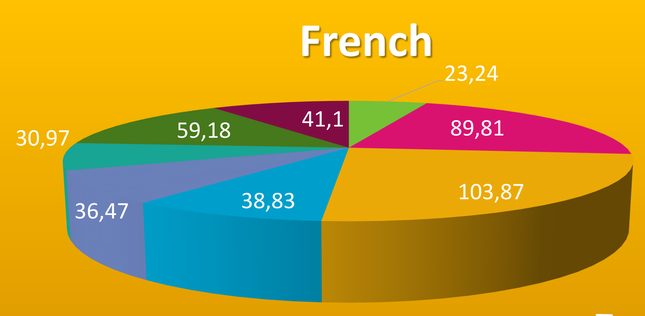
A richly annotated dataset for the contrastive and variationist study of discourse markers in speech

Ludivine Crible

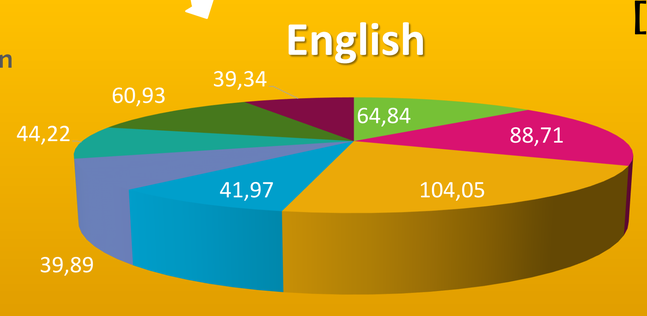
Principles for corpus design

Representativity
[sample size]

- 2 languages, 8 registers
- 15 hours of recordings, 163,620 words



Impromptu speech
[frequency of DMs]



Feasability
[availability of source corpora]
[manual annotation]

Variation
[impact of context on discourse features]

Corpus-based functional taxonomy for spoken DMs

Speech *vs.* writing
- Less types, greater ambiguity
- ➢ Need to **group** values
- Speech-specific functions
- ➢ Need to **add** values

Writing-based models
- PDTB: 3-level **hierarchy**
- CCR: 4 **dimensions**, no end-label
- SDRT/RST: **spans** over whole texts
- Include **implicit** relations

Other frameworks
- **Generic** functions only
- Distinctions **not operational**
- **Incoherent** groupings in categories
- Language and/or genre-specific

several tests on pilot corpus

"**Domains**" as macro-functions (González 2005)

Objective-subjective distinction

Operational definitions (PDTB-style)

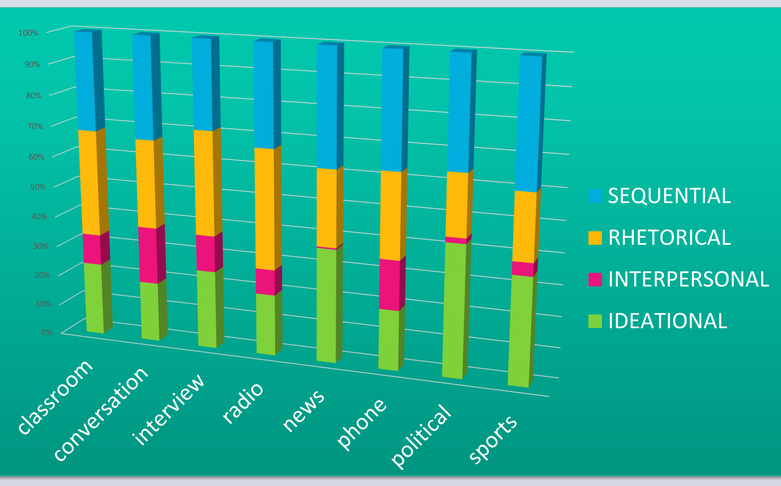
Ideational	Rhetorical	Sequential	Interpersonal
cause	motivation	punctuation	monitoring
consequence	conclusion	opening boundary	face-saving
concession	opposition	closing boundary	disagreeing
contrast	specification	topic-resuming	agreeing
alternative	reformulation	topic-shifting	elliptical
condition	relevance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

Extended to spoken functions (Cuenca 2013)

- 4 domains, 30 functions
- Domains and functions are inter-dependent
- Up to 2 simultaneous functions
- Explicit functions/relations only

Results

- 8743 DM tokens
- Sequential most frequent
- FR more interpersonal
- EN more ideational



Most frequent functions

English	French
Addition	Addition
Specification	Monitoring
Consequence	Opposition
Temporal	Specification
Conclusion	Conclusion

Reliability of the protocol

- Applicability to speech, writing, gestures, sign language
- Domain: K = 0.563, 70.9%
- Function: K = 0.59, 60%

Perspectives

Cross-tabulation of functional and syntactic features of DMs with word-level annotation of local markers of (dis)fluency (filled pauses, repetitions, etc.). Combination with experimental, machine-learning and qualitative methods. Comparison with other annotation frameworks and languages.

Selected References: ASHER, N. & LASCARIDES, A. 2003. *Logics of conversation*. Cambridge: CUP. 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)*, 14-22. CUENCA, M.-J. 2013. The fuzzy boundaries between discourse marking and modal marking. In Degand et al. (eds), *Discourse markers and modal particles. Categorizations and description*, Amsterdam, John Benjamins: 191-216. GONZÁLEZ, M. 2005. Pragmatic markers and discourse coherence relations in English and Catalan oral narrative. *Discourse Studies* 7/1: 53-86. PRASAD, R. et al. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of LREC 2008*: 2961-2968. SANDERS, T., SPOOREN, W. & NOORDMAN, L. 1992. Toward a taxonomy of coherence relations. *Discourse* Processes 15: 1-35.

Contact & acknowledgments: This research benefits from the financial support of the ARC "Fluency and disfluency markers. A contrastive multimodal perspective" project funded by the Fédération Wallonie-Bruxelles, grant nbr. 12/17-044, and the ISCH COST Action IS1312 "TextLink. Structuring Discourse in Multilingual Europe". Please contact the author at ludivine.crible@uclouvain.be for any questions.