**DisFrEn**

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

Ludivine Crible

---

**Principles for corpus design**

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

- **Variation**
  - Impact of context on discourse features

- **Feasibility**
  - Availability of source corpora (manual annotation)

---

**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**

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

---

**Results**

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

---

**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**


---

**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.