DisFrEn : a richly annotated dataset for the contrastive and variationist study of discourse markers in speech

Ludivine Crible
Université Catholique de Louvain
Place Blaise Pascal, 1, 1348 Louvain-la-Neuve, Belgium
Email: ludivine.crible@uclouvain.be

This paper presents the data and method of a crosslinguistic and variationist approach to discourse markers (DMs) in speech. DMs are here broadly defined as "fulfilling structuring functions with respect to local and global content and structure of discourse" (Fischer, 2000: 20). In order to reach full coverage of this category and following corpus-based definitions (Crible, 2014), manual annotation of numerous functional and surface features was performed upon a comparable corpus of native French and English, balanced across eight contextual settings (e.g. conversation, interview) with over seven hours of speech in each language and about 160.000 words in total. Linguistic variation (language, register, modality) and the heterogeneity of the DM class were accounted for in a robust annotation scheme designed through careful corpus-based testing.

The presentation will focus on the elaboration of a functional taxonomy which builds on existing categorizations (Halliday & Hasan, 1976; Sweetser, 1990) and definitions (Gonzalez, 2005; Prasad et al., 2008). The novelty of the present proposal lies in the integration and extension of traditional writing-based sets of discourse relations to additional (non-)relational functions, and in the revision of certain values to better grasp the specificity of the spoken mode. After operationalization, the taxonomy comprises thirty functions grouped in four domains:

- ideational (semantic relations);
- rhetorical (pragmatic relations and speaker's attitude),
- sequential (topic structure and interaction management)
- interpersonal (hearer-orientation).

These domains partially take up well-established distinctions in the literature on DMs and other pragmatic phenomena: semantic vs. pragmatic source of coherence (Sanders, 1997), the objective-subjective-intersubjective continuum (Traugott, 2007), and the scale of relationality of DMs (Degand & Simon-Vandenbergen, 2011).

The resulting dataset *DisFrEn*, with over 8000 DM tokens identified and annotated, offers many valuable insights into the interface between form and function of discourse markers. The analytical potential of the present categorical and corpus-based approach to DMs will be briefly illustrated by overall distribution results focusing on the functional variables. Our data shows that in both languages and most situations, the sequential domain is the most frequent, with a few exceptions due to the impact of context: more rhetorical (subjective) relations in classroom lessons (cf. transmission of information through epistemic relations) and more ideational (objective) relations in political speeches, where DMs are mostly used to highlight content-based connections. As far as position is concerned, the sequential and interpersonal domains are almost exclusively found in non-governed positions (outside the syntactic boundaries of the dependency structure, either left or right periphery), while the ideational and rhetorical domains show non negligeable frequencies in more integrated slots, which is expected from their connective status.

Finally, a further window into the cognitive processes of online speech production is provided by the annotation of various local marks of (dis)fluency (e.g. pauses, repetitions, false-starts) in the direct co-text of DMs. Hence, DMs in *DisFrEn* are studied for their contribution to the relative (dis)fluency of their utterance, through a combination of functional and surface features that should cluster in relevant patterns for the cognitive-pragmatic modelling of this complex category.

References

Crible, L. (2014). *Identifying and describing discourse markers in spoken corpora. Annotation protocol v.8.* Unpublished working draft, Université Catholique de Louvain.

Degand, L., Simon-Vandenbergen, A.-M. (2011). Grammaticalization and (inter-)subjectification of discourse markers. *Linguistics* 49, pp. 287—294.

Fischer, K. (2000). *From cognitive semantics to lexical pragmatics*. Berlin: Mouton de Gruyter.

González, M. (2005). Pragmatic markers and discourse coherence relations in English and Catalan oral narrative. *Discourse Studies* 77(1): 53—86.

Halliday, M. A. K., Hasan, R. (1976). *Cohesion in English*. London: Longman.

PDTB Research Group. (2008). *The PDTB 2.0 Annotation Manual*. Technical Report IRCS-08-01, Institute for Research in Cognitive Science, University of Pennsylvania.

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.

Sweetser, E. (1990). *From etymology to pragmatics*. Cambridge: CUP.

Traugott, E. C. (2007). Inter-subjectification and unidirectionality. *Journal of Historical Pragmatics* 8: 295—309.

Acknowledgments

This research benefits from the support of the ARC project "A Multi-Modal Approach to Fluency and Disfluency Markers" granted by the Fédération Wallonie-Bruxelles (grant nr.12/17-044).