Variation of Discourse Markers across a multi-genre corpus of spoken French

Liesbeth Degand, Anne Catherine Simon
Université catholique de Louvain, Institute for Language & Communication
Place Blaise Pascal 1, bte L3.03.33, 1348 Louvain-la-Neuve Belgium
Email: liesbeth.degand@uclouvain.be, anne-catherine.simon@uclouvain.be

This contribution presents an empirical study of Discourse Markers (DM) in a multi-genre corpus of spoken French (LOCAS-F), a dataset of spoken French segmented into Basic Discourse Units (BDUs). A BDU results from the mapping of a syntactic clause and a major intonation unit, giving rise to different types of discourse units (congruent, syntax-bound, intonation-bound, regulatory, mixed; Table 1). Thus, we claim that the prosody-syntax interface gives rise to a distinctive discursive level of analysis contributing to the unfolding (linear) discourse. It follows that we expect to find an interaction between BDUs and other typical linguistic expressions working at the discourse level. A case in point are Discourse Markers.

Type	Examples taken from LOCAS-F
bdu-c	[dans la majorité politique libanaise beaucoup de voix s'étaient élevées contre sa venue] ///
bdu-s	[pourquoi ce rôle majeur n'est-il pas dévolu /// à la religion] ///
bdu-i	[euh on est devenu bien potes] [tout le monde se connaissait] ///
bdu-r	<mais> <bon> ///
bdu-x	<donc> euh [je veux dire] euh [l'employeur a le plus la possibilité de /// de choisir vraiment euh la personne qu'il lui faut] <et> ///

Table 1: Types of BDUs (syntactic clauses in square brackets; major intonation boundaries delimited with ///): -c = congruent, -s = syntax-bound, -i = intonation-bound, -r = regulatory, -x = mixed.

The LOCAS-F corpus comprises 48 samples of speech, taken from 14 different speech activities and amounting to 3h38 / 41 322 tokens. The communicative situations have been characterized in terms of 5 scalar features: type of elicitation, number of speakers, degree of preparation, of interactivity, of professionalism, and broadcasting. In most cases, one syntactic unit does not correspond to one prosodic unit: out of 2875 BDUs, 43% are congruent, 23% are syntax-bound, 17% are intonation-bound, 9% are regulatory and 8% are "mixed". Corpus analysis supports the hypothesis that discourse production results from strategies varying across situational features. Indeed, the distribution of BDU types varies significantly according to the degree of preparation and the degree of interactivity. For example, the more prepared, the more syntax-bound BDUs, the less prepared, the more intonation-bound BDUs (Degand & Simon, 2009; Degand, Martin & Simon, 2014; Martin, Degand & Simon, 2014).

In LOCAS-F, 1780 occurrences of DMs were identified (all in weak clause association with their host utterance, Schourup, 1999), of 73 different types (*alors*, *ben/bien*, *donc…*). Unsurprisingly, the number of tokens and types varies with the discourse situation. A pilot study showed that highly interactive, non-prepared and non-broadcasted discourse favors DM use, both in terms of types and tokens (39 types, 418 tokens), compared to non-interactive, prepared and broadcasted discourse (19 types, 117 tokens). More interestingly seems to be the observation that certain DM types are restricted to certain discourse situations, that is, they are situation-specific. Thus, while DMs have been described as "a thing of speech", a more fine-grained description of the context of speech is required when it comes to analyzing DM use more thoroughly.

Another contextual factor that needs to be taken into account, when describing DM use, is the positional slot the DMs occupy. As shown in Table 2, DM position varies according to the host unit taken into account, i.e. BDU, syntactic clause, or intonation unit.

	Initial	Medial	Final	Isolated
BDU	697	833	163	87
Syntactic clause	1321	114	177	/
Intonation unit	715	797	181	87

Table 2: DM position in LOCAS-F

In ongoing work, DM function is being annotated making use of an annotation protocol developed specifically for spoken language (Crible & Zufferey, 2015). We expect to find significant variation of the distribution of DM function according to both contextual factors described so far, namely situational features ('genre') and position in the host unit.

References

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), pp. 14—22.

Degand, L., Simon, A.C. (2009). Mapping prosody and syntax as a strategic choice. In D. Barth-Weingarten, N. Dehé & A. Wichmann (Eds.). *Where Prosody Meets Pragmatics*. Bangalore: Emerald, pp. 79—105.

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 *CMLF 2014 – 4^{e} Congrès Mondial de Linguistique Française*, Berlin: EDP Sciences, pp. 2613—2626. doi:10.1051/shsconf/20140801211.

Martin, L.J., Degand, L., Simon, A.C. (2014). Forme et fonction de la périphérie gauche dans un corpus oral multigenres annoté. *Corpus* 13, pp. 243—265.

Schourup, L. (1999). Discourse markers. *Lingua* 107 (3-4), pp. 227—265.