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

Liesbeth Degand & Anne Catherine Simon

The Louvain Corpus of Annotated Speech – French is a dataset of spoken French segmented into Basic Discourse Units (BDUs). A Basic Discourse Unit 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)

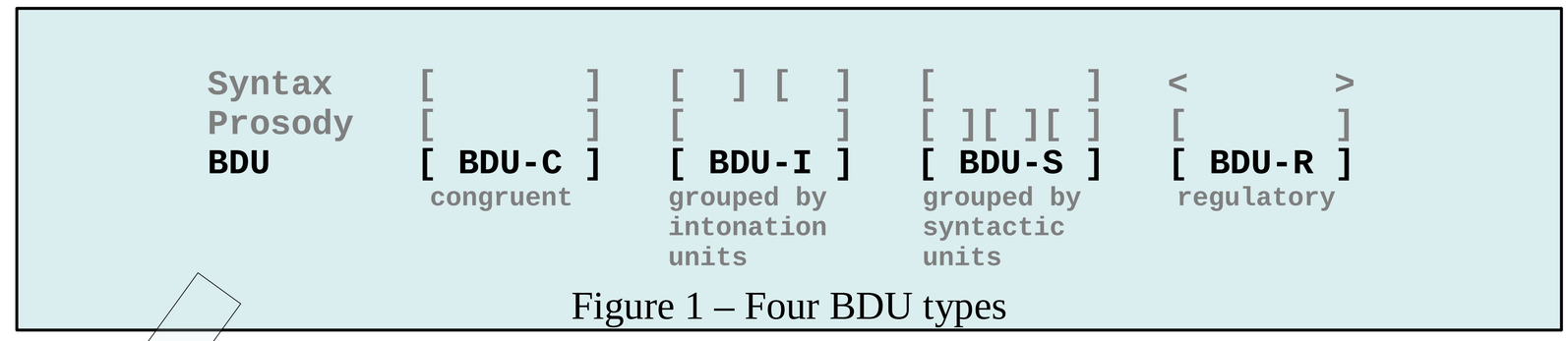

Figure 1 – Four BDU types

- 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**.

BDU TYPE	Examples from LOCAS-F (simplified representation)
BDU-c	<en fait> //C [elle se fait sous la forme d'une grille métrique qui va indiquer les relations de proéminences //S entre les syllabes] ///C
BDU-i	[je sais pas] //T <enfin> <bon> //T [un très chouette film] ///C
BDU-s	[quand euh on interroge Bernadette euh elle répond ///C que si la Vierge l'a choisie //C c'est parce qu'elle était euh la plus ignorante] ///T
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> ///

The Corpus

#	BDUs	Words	Length	Situations	Samples	DM
Total	2875	41322	3:38	14	48	1780

- BDU-c: 43%
- BDU-i: 17%
- BDU-s: 23%
- BDU-r: 9%
- BDU-x: 8%

- Type of elicitation
- Number of speakers
- Degree of preparation
- Degree of interaction
- Degree of broadcasting

Weak clause association
Relation between host utterance and discourse situation:
- textual coherence
- interpersonal meanings
- epistemic meaning

DMs in LOCAS-F

Position	Initial	Medial	Final	Isolated
BDU	697	833	163	87
Clause	1321	114	177	/
Intonation	715	797	181	87

	+ interactive	- interactive
- prepared	39 types **649** tokens *après, au fond, bien que, encore que, en même temps, et tout, par contre, quoique, sinon*	28 types 255 tokens *bref, ensuite ou quoi*
+ prepared	0	37 types **314** tokens *ainsi, au contraire, car, cependant, certes, de ce fait, de même, de plus, en effet, par ailleurs, par conséquent, pour autant, toutefois*

Annotation of DMs in LOCAS-F

Ideational	52 (17%)	et, mais, donc, alors
Rhetorical	94 (32%)	
Sequential	139 (47%)	
Interpersonal	13 (4%)	hein
Total sample	**298**	

Cf. Crible 2015

Round 1	Round 2	Round 3
Kappa: .43 + (abstract) coding scheme	Kappa: .59 + bias ideational > rhetorical > sequential > interpersonal	Kappa: .78 + bias + DM cues/paraphr.

Preliminary Conclusions:

- MACRO-functions are difficult to annotate BUT interesting → distribution across genres, registers, interaction with position → **dimensions?**
- Bias needed to disentangle: ideational/rhetorical, ideational/sequential, rhetorical/sequential
- ? Cross-linguistic validation of macro-functions**:** DMs as (linguistic) **cues** of domains increase agreement (cf. underspecified DMs), but cross-linguistic impact?
- Typical spoken dimensions/domains: **sequential** (including distinction between fluent and disfluent sequences)
- **Questions to solve**: complex DMs (*et puis, et voilà, et alors, …) vs.* repeated DMs (*mais mais; et et; …)*