STSM REPORT

TITLE:	**Cross-linguistic applicability of the DRD annotation protocol: comparing results for French, Spanish, German and Polish**
DATES:	**18th September - 15th October 2016**
INSTITUTION:	**Université Catholique Louvain-la-Neuve (UCL)**
SUPERVISOR:	**Prof. Liesbeth Degand**
GRANTEE:	**Karolina Grzech**

1. Fulfilment of objectives

I have undertaken the STSM at UCL in September/October 2016, under the supervision of Prof. Liesbeth Degand. The objective of the STSM was to test the protocol for annotation of discourse markers in spoken discourse, developed on the basis of the protocols proposed by Crible (2014), and adapted by Crible, Degand and Simon (2016) . The aims of the STSM were the following:

1) Participant's training in functional annotation of DRDs in spoken corpora.
2) Selection of material from comparable corpora of Spanish, French, German and Polish to be annotated.
3) Carrying out the annotation process of the selected material.
4) Identifying the issues which arise in DRD annotation of comparable spoken corpora.
5) Consulting the UCL research team with the view of resolving the problematic issues.
6) Adapting the DRD annotation protocol to resolve the encountered issues.

During my monthly stay at UCL, all these goals were satisfactorily met, apart from goal 2. Prof. Degand and I have decided that it would be more beneficial from the project to annotate a substantial amount of text in one language, and compare it with the annotation of French data carried out previously by prof. Degand. Therefore, I have concentrated on annotation of DMs in Polish.

2. Research results

During the course of the STSM, I have identified and annotated over 850 DMs in the Polish conversational data, which were then compared with annotated French data. As a result, a number of adjustments were made to the DM annotation protocol with which we have initially worked; the resulting protocol distinguishes between 4 domains and 15 functions. Domain and function labels can be assigned independently (cf. Crible, Degand & Simon 2016). The domains remain the same as in the work of Crible, Degand and Simon (2016, see Figure 1). The inventory of functions (see Figure 2) is based on the work of Crible (2014) and references therein. However, in comparison with previous work, it has been substantially reduced.

Figure 1: DM domains (after Crible, Degand & Simon 2016)

Domain	**Definition**
Ideational	Linked to states of affairs in the world, semantic relations between real events.
Rhetorical	Pragmatic relations applying to subjective claims, implicit assumptions or speech-acts.
Sequential	Linked to the structuring of discourse segments, explicitly signals of the progressing steps of speech and thought.
Interpersonal	Linked to intersubjective and phatic functions in the management of the speaker-hearer relationship.

Figure 2: DM functions[1]

Function	Definition
Addition	The basic function of additive connectives when they "provide additional, discourse new information that is related to the situation described in Segment1 (S1).
Alternative	The arguments are alternative situations, exclusive or not. The choices or the preference given by an alternative relation do not imply the speaker's subjective appreciation of an expression that fits their intention better, unlike reformulative relations, but merely reports competing facts. Paraphrase: "on the one hand... on the other hand"; "instead". Also includes reformulation, i.e. equivalence between two simple units with a change in phrasing.
Cause	"The situations described in S1 and S2 are causally influenced and the two are not in a conditional relation" (Prasad et al. 2007: 28). It includes a pragmatic (epistemic or speech-act) cause that applies to the subjective content of a claim or a speech-act. Paraphrase: "This happened because..."/ "I say this because…"
Closing	The item indicates the intention to close a list, a thematic unit or a turn. It must be in final or autonomous position. Paraphrase: "This topic/ this turn is now closed".
Concession	Markers deny one or several clearly identified expectations explicitly related to the concessive segment. Concession can apply to both events and assumptions as long as the expectation derived from S2 is logical and verbally expressed. Includes pragmatic (epistemic/speech-act) concession, and counter-expectation. Paraphrase: "Although..., yet something else happened". / "Although I said that, actually..."
Condition	Occurs when the situation in S2 is taken to be the condition and the situation in S1 is taken to be the consequence. It includes all possible subtypes identified by the PDTB group (present, past, unreal etc.). It includes pragmatic condition, specifically when S1 and S2 are not causally related (Prasad et al. 2007: 31). The condition is what makes the speech-act relevant to the particular context. Paraphrase: "On this condition only...". / "I can say this only in the context of..."

[1] For the sake of clarity of presentation, the descriptions of functions presented here have been slightly simplified.

Consequence	The situation in S2 is the [logical] effect brought about by the situation described in S1" It includes markers of purpose, result, epistemic/speech-act consequence, summaries with conclusive value (PDTB's "generalization"), but excludes simple paraphrasing (see Alternative). Conclusion usually corresponds to an evaluation or a generalization, when the causal link between the two segments is under-specified other than by the speaker's appreciation. "We can now say that ..."/ "As a consequence of that, this happened"
Contrast	S1 and S2 share a predicate or property and a difference is highlighted with respect to the values assigned to the shared property" (Prasad et al. 2007: 32), either as an opposite (PDTB's subtype "juxtaposition") or as a scalar difference (PDTB's "opposition"). Contrast differs from concession by explicitly referring to a verbally expressed property that is contrasted. It includes pragmatic (epistemic or speech-act) contrast. Paraphrase: "X is this, whereas Y is that"/ "Although I said that, actually..."
Opening	The item opens a new turn, in which case it indicates floor-taking, or a new sequence within the same topic. Apart from turn-taking, it corresponds to any form of opening or engaging which is not covered by topic-shift or any other sequential function. Cuenca (2013) refers to this function as "start".
Punctuation	The item signals the intention to hold the floor while planning the upcoming speech, or for any other reason not mentioned by the other functions related to text-structuring. Used in cases of holding the floor, quoting, 'weak structuring' of text - with markers with no semantic content corresponding of their own. Paraphrase: corresponds to typographical commas.
Resuming	The item signals the intention to link the upcoming segment to previous topic, to come back to the topic after a digression, a hesitation or a non-relevant passage. Formal criteria include anaphora or reference to a previous topic which is taken up.
Specification	Introduces a remark not directly related to the current discourse topic but considered relevant for full understanding: a digression or parenthesis. Paraphrase: "by the way". It "applies when Arg2 describes the situation described in Arg1 in more detail" (Prasad et al. 2007: 34) and instantiates Arg1 with an example. The content of S2 must fall within the informational scope of S1.
Temporal	The situations described in the arguments are related temporally, either synchronically or asynchronically. Temporal bias is suggested in case of conflict with under-specified consequence relations. Paraphrase: "After/before/during/then".
Topic-shift	Signals a change of topic within or between turns. A distant connection to previous context can still remain, with a shift in focus. The new topic can be a subtopic of the previous one, but the latter should be definitely closed and not taken up in upcoming speech. Paraphrase: "Let's move on to..."

During the STSM, this adapted protocol has been tested on French and Polish data, and seems to account satisfactorily for the use of discourse markers in spoken conversation.

3. Implications & future work

The STSM has fulfilled its proposed contribution to the scientific objectives of TextLink. It has also contributed to the 2016/17 research focus of the COST Action 1312 by extending the set of available resources on DRDs. We hope that the refinement of the DM annotation protocol will be used for analysis of spoken discourse in a number of languages, leading to the unification of the annotation systems used to describe and analysed DMs in spoken discourse. If the protocol we propose is deemed cross-linguistically applicable, it will have implications for the cross-linguistic taxonomy of discourse relations.

Prof. Degand and I continue working together beyond the STSM. As the next step in the process, I will undertake coding of the French data, so as to evaluate our inter-annotator agreement. Then the protocol will be applied to another European language. Depending on the activities which will take place within TextLink, we plan to propose a workshop to test the protocol on a larger number of languages. I am also planning to apply the protocol to conversational data from Tena Kichwa (Quechuan, Ecuador).

4. References

CRIBLE, Ludivine. 2014. Reaching cross-linguistic comparability across eight speech situations : a challenge for corpus design.

CRIBLE, Ludvine, Liesbeth DEGAND & Anne-Catherine SIMON. 2016. Interdependence of annotation levels in a functional taxonomy for discourse markers in spoken corpora. 2nd TextLink Action Conference, 11-13 April. Budapest.

CUENCA, Maria Josep. 2013. The fuzzy boundaries between discourse marking and modal marking. In Liesbeth DEGAND, Bert CORNILLIE & Paola PIETRANDREA (eds.), *Discourse markers and modal particles: categorization and description*, vol. 234, 191–216. (Pragmatics & beyond New Series). Amsterdam ; Philadelphia: John Benjamins Publishing Company.

PRASAD, Rashmi, Eleni MILTSAKAKI, Nikhil DINESH, Alan LEE, Aravind JOSHI, Livio ROBALDO & Bonnie WEBBER. 2007. The Penn Discourse Treebank 2.0 Annotation Manual. *IRCS Technical Reports Series*. http://repository.upenn.edu/ircs_reports/203 (accessed 01/11/2016).