CRIBLE Ludivine
Université catholique de Louvain
Place Blaise Pascal, 1, box L3.03.33
1348 Louvain-la-Neuve (Belgium)
ludivine.crible@uclouvain.be

October 24, 2017

STSM Report

To whom it may concern,

I have just completed a five-week Short Term Scientific Mission at the Universität des Saarlandes (Germany) under the supervision of Prof. Vera Demberg. The goal of this research stay was to train on corpus-based and experimental methods in the study of underspecified connectives such as "and", "but" and "so". The expected outcomes were to apply Asr & Demberg's (2012) statistical measure of cue strength to my annotated data of spoken language, to elaborate a typology of cooccurring cues in the vicinity of underspecified connectives, and to design psycholinguistic experiments for the study of the production and comprehension of underspecified connectives. At the end of this STSM, the corpus study was completed and the experimental plan is well advanced with one pilot study already launched. Although interesting trends have been observed regarding the cooccurrence typology, its precise definition and implementation remains for future work.

During my stay in Saarbrücken, Prof. Demberg and I have worked intensively on two specific aspects of underspecified connectives: (1) their distribution across registers in an annotated corpus of spoken English; (2) their disambiguation, elicitation and perception through crowdsourcing experiments.

Corpus study

For this first part of the STSM, we used the annotations of DRDs available in the *DisFrEn* corpus of spoken English (Crible 2017). The eight registers of the corpus were grouped in three genres, namely informal, semiformal and formal speech, in order to obtain larger frequencies. We extracted the sense annotations for "and", "but" and "so" according to the hypothesis that the uses of these high-frequency polysemous connectives would vary across registers. More specifically, we expected them to show a larger functional spectrum (more different senses) in informal speech than in formal speech, where speakers might choose less ambiguous connectives. The results show that "and" confirms this hypothesis: "and" expresses fewer different senses in formal speech, where 80% of its uses express its basic meaning of addition, against around 60% in informal and semiformal speech. "So" shows interesting effets of register variation, but the available data is not large enough to strongly conclude anything besides a larger functional spectrum in informal and semiformal speech, where "so" is used for structuring (e.g. topic-shift), specifying and reformulating functions. Similarly, the connective "but" is used for structuring purposes in informal and semiformal speech only. Apart from this observation, it is not much affected by register variation in terms of connective strength.

Prof. Demberg then explained to me how to compute the statistical measure of cue strength as introduced in Asr & Demberg (2012). I applied it to the same corpus data. The results for "and" and "so" can be found in the following table (the figures in red represent missing data):

	informal	**semiformal**	**formal**	**DisFrEn**
and				
addition	0.6214	1.0001	0.8181	0.571
contrast	0.0116	0.0317	0.0101	0.012
concession	0.0154	0.0127	0.0101	0.007
consequence	0.0617	0.1048	0.0606	0.0886
so				
consequence	0.2709	0.2875	0.125	0.2838

On the basis of these results, computed on rather low frequencies, we can provisionally conclude that our hypothesis only predicted correctly the cue strength of "and" for the additive relation (i.e. the scores decrease with formality) and that there seems to be more differences between semiformal speech on the one hand and informal and formal genres on the other hand, than between the two extremes of the continuum. These scores only illustrate the type of analysis that can be made but cannot be taken as reliable measures of cue strength, given that I did not have enough data points in the corpus for robust statistics.

Overall, none of these connectives are "strong", in the sense of Asr & Demberg (2012): the relations of addition, consequence, contrast and concession are marked by many more connectives besides "and", so" and "but", and these connectives express many other senses as well.

Experiments

Our experimental work has very much progressed during these five weeks. Prof. Demberg and I have prepared a coherent and detailed plan for four crowdsourcing experiments, which need to be carried out one after the other in a predefined order. They are all offline tasks using partly authentic stimuli collected from the Loyola CMC corpus (written chats and blog posts) and mostly focus on the production and perception of "and" as an underspecified connective. The first one, for which the pilot study is ready launched, is a disambiguating task where participants have to select a strong connective (e.g. "therefore", "by contrast") to fill a blank between two sentences which were originally connected by "and". The goal of this first study is to gather indirect annotations (or rather, disambiguations) of discourse relations which are compatible with "and", to be used in later experiments.

The second experiment is highly similar to the first one, except that the stimuli are now presented with the original "and", and the task is for the participants to replace "and" by a more specific connective such as "therefore" or "by contrast". This second design targets potential differences in disambiguations with and without the connective and aims at identifying the specific role of "and" in such crowdsourcing tasks.

The third experiment is a connective elicitation study making use of partly different stimuli which originally contained "and" but also "but", "so" and stronger connectives such as "however" or "therefore". In this design and the next one, we add the variable of discourse genre by alternating two text types with different degrees of formality: chats (informal, conversational) and comments to online press articles (formal). We expect that the participants will use weaker and/or underspecified connectives more often in the informal genre of chats, and more so in consequence relations than in contrastive relations.

The last experiment currently included in our work plan is a perception study where the participants are presented with two versions of the same stimulus, which only differ by the connective (a stronger and a weaker alternative, e.g. "so" vs. "and" or "therefore" vs. "so"). The participants have to select which version they prefer or are likely to have produced themselves. Again, we expect an effect of discourse genre (same two text types as in the previous study) and of discourse relation (consequence vs. contrast).

For all four studies, the design, instructions, stimuli and fillers are completely ready (except for the stimuli with "and" that await validation from the first experiment). They will be launched on the Prolific platform for crowdsourcing experiments. I take this opportunity to thank Florian Pusse who is programming the interface. These studies fill a gap in the literature, which was so far not so much focused on "and" and which did not address genre variation as a potential factor. Prof. Demberg and I have also discussed further designs using spoken stimuli, which require more time to be implemented. In particular, these additional studies will be built on the basis of the results of the corpus study on cooccurring cues, to be carried out in the future.

This STSM was also an opportunity to exchange ideas about experimental designs for underspecified connectives with Frances Yung, a postdoctoral researcher currently working at the Host institution, who has recently designed a crowdsourcing task with a game design. We came to the conclusion that connective elicitation (without the original connective) strongly differs from both natural production and comprehension, and that these tasks should be complemented with online inferential tasks.

Cooccurring cues

In the process of collecting stimuli for the experiments, I was able to note the following trends, which will be relevant in the elaboration of the typology of cooccurring cues: when "and" is used in a contrastive relation, the segments tend to contain antonym pairs (e.g. "men" vs. "women"); in a consequence relation, typical cues include future tense and modal verbs in the second argument; in both relations, the subject of the second argument is sometimes elided. These observations are likely to vary across languages: for instance in French, we can expect that the two past tenses *imparfait* and *passé composé* will be used respectively in the first and second arguments to compensate for an underspecified "et" ('and').

Discussion with Dr Frances Yung revealed that computational approaches to cooccurring cues are limited in their contribution to the building of such a typology. In her 2017 paper with colleagues, she found that, while discourse parsers can automatically predict which relations will be explicit or implicit, the method is mute with respect to which cues were actually used for the prediction of each specific case. There is therefore some room for complementary, qualitative (manual) approaches to cooccurring cues.

Networking

This STSM also represented an opportunity for me to present my past and current research to the linguists of the Host institution, during an internal seminar (FEAST talk) on October 18th. Interesting questions arose during the discussion, regarding the distinction between crowdsourcing tasks, annotation and online interpretation. We also mentioned the annotation strategy to use double tags. Ideas for further experimental designs were suggested to me.

Finally, I attended a scientific presentation (another FEAST talk) in the field of linguistics at the host institution: Evan Brown presented his work on linguistic tools for applied purposes (education and law).

Future collaboration

Parts of this STSM were submitted to be presented at the DiscourseNet conference "Exploring Fuzzy Boundaries in Discourse Studies" to be held in Budapest in May 2018. Once all experiments are run and analyzed, we will publish the results in a journal such as *Dialogue and Discourse* or *Discourse Processes*. Our collaboration will continue after this STSM, to carry out the rest of the experiments, and possibly design others.

We are confident that this STSM will be highly relevant to the discussions and deliverables of WG2 and WG3 by focusing the attention on the complex phenomenon of underspecified connectives, by suggesting different annotation strategies for these cases and by assessing the merits of crowdsourcing tasks in their analysis and interpretation.

Yours truly,

Ludivine Crible

References

Asr, F. & Demberg, V. 2012. Measuring the strength of linguistic cues for discourse relations. In *Proceedings of ADACA.*

Crible, L. 2017. Discourse markers and (dis)fluencies in English and French: Variation and combination in the DisFrEn corpus. *International Journal of Corpus Linguistics* 22(2): 242-269.

Yung, F., Duh, K., Komura, T. & Matsumoto, Y. 2017. A psycholinguistic model for the marking of discourse relations. *Dialogue & Discourse* 8(1): 106-131.