**REPORT OF RESEARCH VISIT**

**COST STSM Reference Number:** COST-STSM-IS1312-21434

**COST Action:** IS1312

**STSM Type:** Regular (from Germany to Netherlands)

**STSM Applicant:** Fatemeh Torabi Asr, Saarland University, Saarbruecken (DE)

**Period:** 2015-04-09 to 2015-04-18

**STSM Topic:** Structuring Discourse in Multilingual Europe (Text-Link)

**Host:** Prof. Ted Sanders, Utrecht Institute of Linguistics, Universiteit Utrecht, Utrecht (NL

**Perspective (submitted previously in the application):**

Previous work on categorization and formalization of discourse relations has not emerged to a consensus in terms of what types of relations exist or would best explain human understanding of text coherence. One important focus in my PhD project has been to investigate the applicability of the Penn Discourse Treebank for studying information theory at the level of discourse. The overlap between our computational research on discourse relations and discourse markers (in Dr. Vera Demberg's group at Saarland University) with the psycholinguistic approach to the same problem in Prof. Sanders's group at the Utrecht institute of Linguistics encouraged us to have a visit of them in order to exchange ideas on cross-corpora and cross-linguistic studies which sheds some light on the general characteristics of the discourse regardless of a particular taxonomy of relation senses, i.e., via cognitive and empirical validation of relation sense categories. This action includes two separate visits: Dr. Demberg travelled in March in order to present the lab's general approach to the study of discourse relations and discuss possible roadmaps for a collaborative attempt. I travelled in April to present work specifically related to the discourse relations from an information-theoretic perspective and come up with two project proposals with specific emphasis on the goals of STMS regarding standardization of discourse annotation schema and mapping from one to another.

**Report of visit**

**April 10:** in the MODERN Project meeting I presented my computational study of discourse relations and linguistic features extracted form Penn Discourse Treebank, in particular negation features and connectives to investigate the Uniform Information Density hypothesis at the level of discourse. We discussed several ways in which the theory can be investigated for other languages such as Chinese that in particular is being studied in Utrecht by Yipu Wei (who presentation her work in the meeting). Jet Hoek also presented her cross-lingual study on implication (using vs. dropping discourse connectives), which was related to my talk as well as our 2012 paper on "implicitness of discourse relations". Then, I had a meeting with Merel Scholman on her ongoing eye-tracking study regarding processing of the causal relation. Together with Merel we conducted a set of eye-tracking studies at University of Edinburgh last year, which she wanted to present in Saarbruecken.

**April 13:** I met Sergey Avrutin specifically to talk about the information theoretic approach to language processing, an in particular in my work on discourse-level phenomena. Then I had a colloquium talk on the distributional meaning representation for discourse connectives, modeling it based on data from Penn Discourse Treebank and making predictions about how ambiguous or multi-sense discourse markers such as BUT and ALTHOUGH are processed in context where different interpretations are possible. I got very helpful feedback on the coherence judgment and the eye-tracking studies and we next had a meeting with Ted, Jet Hoek, and Jacqueline Evers-Vermeul on the possible interpretations on top of the findings of my study, as well as, follow-up experiments that can be designed in Dutch or English to make sure predictions of the distributional model is in general language-independent. Jet and Jacqueline are specifically working on cross-linguistic corpora, which provides a great framework to establish collaborative research.

**April 14-15:** I had to travel to London to present a paper at the IWCS conference.

**April 16:** I had a second meeting with Jacqueline, Jet and Ted on more focused topics such as the methodology to study objective vs. subjective relations or pragmatic vs. semantic ones. We talked about selecting the best dimensions (causality, polarity, etc.) to define relation senses in different schemes such as PDTB and RST and discussed what dimensions can be empirically tested, i.e., to make sure humans perceive them while reading text or can agree on it e.g., in an annotation task.

**April 17:** In my last meeting with Ted I presented some ideas on the methodology to approach the mapping between discourse relation annotation schemes. In particular I suggested some PDTB relations should be re-annotated given my observations during analysis of the data regarding the heterogeneous and suspicious annotations existing in this corpus. For example, the Conjunction relations constitute 25% of the entire PDTB relations (most frequent label both among explicit and implicit relations) need to be revised. Since, NLP community is using PDTB as a gold standard for automatic identification of discourse relations, and also the data is being used in psycholinguistic modeling such as in my research it would be our first priority. One way of

doing the revision would be to re-annotated relations by a set of cognitive dimensions such as the ones proposed in Ted's previous studies (e.g., Sanders, Spooren and Noordman 1992, Sanders, Vis and Broeder 2014). The bigger perspective is to construct a mapping between PDTB relation senses and the feature-based space. More on this topic has been presented by Ted and Vera in the meeting of WG3/4 in Freiburg. I also met Yipu Wei on the same day for a discussion on her collocation analysis of the linguistic features and relations on Chinese corpora (that is very similar to my feature analysis on PDTB), as well as previous theories about connectives not being enough in some context for facilitation of the right interpretation (e.g., the study of diagnostic relations by Traxler et al. 1997).

**Summary and outcomes with respect to the goals of STMS:**

During the visits Dr. Demberg and I had from Prof. Sanders's group in Utrecht, we discussed a variety of possibilities to approach discourse relation corpora for psycholinguistic treatments. In particular, we came up with the idea of mapping between relation senses proposed for annotation of discourse relations in Penn Discourse Treebank (as well as similar corpora of other languages which adapt the same schema) and cognitive features such as source of coherence, polarity, and temporal order. A preliminary mapping between PDTB and the set of features taken from previous work of Prof. Sanders, as well as some new dimensions, has been obtained in the coarse of our visits from Utrecht. Dr. Demberg and Prof. Sanders have presented the motivation and a piece of work at the recent meeting of WG3/4 in Freiburg. In order to continue with this idea, we will need to design a concrete methodology that includes annotation of a sample of PDTB relations in their context according to the proposed feature-set for empirical validation of the mapping. This will also help enhancing ideas on how/where new features should be added to distinguish between fine-grained relation senses in PDTB hierarchy.