Annotating discourse relations in spoken language: A comparison of the PDTB and CCR frameworks

Ines Rehbein, Merel Scholman, Vera Demberg
Saarland University
Campus C7.4, 66123, Saarbrücken
{rehbein, m.c.j.scholman, vera}@coli.uni-saarland.de

We describe preliminary work on applying the Penn Discourse Treebank framework (PDTB, Prasad et al., 2008) and the Cognitive approach to Coherence Relations (CCR, Sanders, Spooren & Noordman, 1992) to spoken language. Due to register differences, we do not only expect to see a distribution of discourse relations that deviates from the one in written text, but also expect to come across new relations that are not yet covered by the two frameworks (for an application of the PDTB annotation scheme to spoken Italian see Tonelli et al., 2010).

The PDTB proposes a hierarchical scheme with three layers and 43 different sense labels. In CCR, on the other hand, each discourse relation can be described according to four cognitive categories (basic operation, order, polarity, source of coherence). These dimensions have been claimed to function as an intermediate language that can be used to map annotations from different frameworks, such as PDTB, RST or SDRT, onto each other. Such a mapping would be extremely useful, as it would increase the inter-operability of existing resources and tools for discourse analysis.

Annotating the same texts according to both the PDTB and CCR framework will allow us to investigate the adequacy of the two schemes for spoken discourse annotation. Additionally, we will investigate whether the CCR framework is expressive enough to capture the rich information encoded by the PDTB labels.

For the current study, we augmented parts of the SPICE-Ireland corpus (Kallen & Kirk, 2012) with discourse relation annotations. The SPICE corpus contains spoken texts from different (public and private) discourse situations. The data comprises speech-act annotations, and will thus allow us to investigate whether there is a relation between speech-acts and discourse relations: do certain speech acts occur more often in certain types of relations?

So far, we annotated a sample of 10.316 tokens with 193 explicit relations from the text type broadcast interviews, focusing on the connectives *and, because, so, since, but, whereas, although/though*. As expected, we needed to adapt the PDTB scheme to be able to describe our corpus of spoken language better. For example, we have added a new subtype *pragmatic consequence* to the PDTB hierarchy that allows us to annotate instances where the first argument of the relation expresses a claim and the second argument presents the argument justifying the claim. Our goal is to provide annotations for all discourse relations in the broadcast interview data (including additional connectives and implicit relations), and also to add another text type (private telephone conversations), with a final data size of >20.000 tokens.

Our work contributes on different levels. First, the data will be made available as additional test data for evaluating discourse parsers. Second, the data can be used for studying discourse in spoken language. In addition, we will invite researchers to augment the data with additional annotations in different frameworks, thus allowing us to compare and evaluate the usefulness of different annotation schemes for coherence relations. At the workshop, we will present corpus statistics for the annotated discourse relations in spoken dialogues and discuss the mapping to the CCR dimensions.

References

Kallen, J.L. & Kirk, J.M. (2012). *SPICE-Ireland: A User's Guide*. Belfast: Cló Ollscoil na Banríona.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., & Webber, B. (2008). The Penn Discourse Treebank 2.0. *Proceedings of the 6th International Conference of Language Resources and Evaluation (LREC 2008)*, Marrakech.

Sanders, T.J.M, Spooren, W.P.M.S., & Noordman, L.G.M. (1992). Toward a taxonomy of coherence relations. *Discourse Processes,* 15: 1-35.

Tonelli, S., Riccardi, G., Prasad, R., & Joshi, A. (2010). Annotation of discourse relations for conversational spoken dialogs. *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, Malta.