Analysis of DRD-related Contrasts in Spoken Czech, English and German

Kerstin Anna Kunz*, Ekaterina Lapshinova-Koltunski, Anna Nedoluzhko*****
*University of Heidelberg, **Saarland University, ***Charles University in Prague
kerstin.kunz@iued.uni-heidelberg.de, e.lapshinova@mx.uni-saarland.de, nedoluzko@ufal.mff.cuni.cz

The aim of the present paper is to analyse contrasts in Czech, English and German in terms of discourse-relational devices (DRDs). The novelty of our approach lies in the nature of the resources we are using. Advantage is taken of existing resources, which are, however, annotated on the basis of two different frameworks. We use an interoperable scheme unifying DRDs of both frameworks in more abstract categories and considering only those phenomena that have a direct match in the three languages.

Our aim is two-fold: On the one hand, we intend to compare similarities and differences between spoken and written discourse regarding the use of DRDs and test the interoperable scheme that was originally designed for the analysis of written discourse, as described in Lapshinova et al. (2015). This will help us to check if this scheme is sufficient, or if additional classes of discourse-relational devices (DRDs) have to be added. On the other hand, we are interested in the contrasts existing between typologically close (English vs. German) and typologically more distant (Czech vs. German/English) languages.

Our assumption is that the conditions of spoken language production influence the creation of discourse-relational devices. This has already been postulated for lexico-grammar in various works (e.g., Miller, J. & R. Weinert (2009) and shown by corpus-based works comparing English and German (Amoia et al. 2012, for coreference, and Kunz & Lapshinova 2014, for several cohesive types). For instance, the number of semantically vague devices should be relatively high, because of mutual presence of the speech participants, immediacy and spontaneity of the communication. In addition, the number of devices used should be higher, due to reduced short-term memory capacity and revisions. Although the conditions of spoken language production should find their reflex in all three languages under investigation, we expect some features of spoken discourse to be language-specific.

Quantitative contrastive analyses on the level of discourse require annotated corpora involving time-consuming compilation and annotation, especially in a multilingual context. Therefore, we have decided to take advantage of the existing resources reflecting systemic peculiarities and realisational options of the languages under analysis. We use Czech, English and German data annotated on the basis of two different frameworks: Functional Generative Description, as described in Sgall et al. (1986) for Czech, and textual cohesion, see Halliday and Hasan (1976), for German and English. Lapshinova et al. (2015) have shown that annotations of the involved resources are comparable if abstract categories are used as a starting point and only those phenomena are taken into consideration that have a direct match in the languages under analysis. Such an interoperable scheme allows for capturing the same discourse phenomena across the three languages. For this analysis, we select comparable texts, i.e. transcribed interviews available in both corpora. Although not being clearly representative spoken data (and not capturing prosodic information), these will allow us to get first results on the differences not only between languages, but also between different genres (written and spoken), and thus provide insights for future analysis.

The scheme used for the analysis includes relations of contingency, contrast, expansion and temporal relations. In the Czech data, these categories are further classified into subcategories, e.g., purpose, explication, semantic and pragmatic reason-result and condition for contingency relations. The scheme for German and English contains annotations of the general categories only. However, it also integrates modal adverbs (such as *well*, *certainly* or *of course*) which although not connecting two propositions directly, play an important role for cohesion in spoken language.

In our presentation, we provide more information on our hypotheses, the resources and the scheme applied, as well as the results of our analysis.

References

Amoia, M., K. Kunz and E. Lapshinova-Koltunski. 2012. Coreference in Spoken vs. Written Text: a Corpus-based Analysis. Proceedings of the LREC2012, Istanbul.

Halliday, M.A.K. and Ruqaiya Hasan. 1976. Cohesion in English. Longman, London, New York.

Lapshinova, E., A. Nedoluzhko, and K. Kunz. 2015. Across languages and genres: Creating a universal annotation scheme for textual relations. In Ines Rehbein and Heike Zinsmeister, editors, Proceedings of the Workshop on Linguistic Annotations, NAACL-2015, Denver, USA.

Kunz, K. and E. Lapshinova-Koltunski. 2015. Cross-linguistic analysis of discourse variation across registers. Ajmer, K. and Hassegard, H. (eds.). Nordic Journal of English Studies, 14(1), pp. 258-288.

Miller, J. & R. Weinert. 2009. Spontaneous Spoken Language.Syntax and Discourse. Oxford: OUP.

Sgall, P., E. Hajicova, and J. Panevova. 1986. The meaning of the sentence in its semantic and pragmatic aspects. Reidel, Dordrecht.