Multi-word Discourse Markers in the Spoken Slovenian Treebank

Kaja Dobrovoljc, [1]Institute for Applied Slovene Studies Trojina, [2]University of Ljubljana

Poster proposal related to the activities of Parseme WG4: Annotating MWEs in Treebanks

1 Introduction

Multi-word discourse marking devices, such as discourse connectives, are essential to natural language understanding and generation above sentence level. However, despite their relevance in natural language research, they represent a relatively under-researched group of expressions both in the field of multi-word expressions in general and discourse structuring devices in particular. This poster thus focuses on multi-word discourse markers in a manually annotated dependency treebank of spoken Slovenian, by presenting the identification and functional annotation of discourse-related multi-word expressions and their initial analysis in terms of their functional and formal characteristics. We define multi-word discourse markers as propositionally optional contiguous sequences of two or more words with an institutionalized function of discourse organisation and structuring. In contrast to alternative lexicalizations (Prasad et al. 2010) and other similar concepts (Biber et al. 2004, Siepmann 2005, Nesi and Basturkmen 2006, Rysová and Rysová 2014) that take into account any frequent combination of words performing discourse-organizing functions (e.g. *the reason for this is*), our work therefore focuses only on institutionalised, structurally complete and syntactically optional discourse-related multi-word expressions.

2 Construction, lemmatization and syntactic annotation on the treebank

The Spoken Slovenian Treebank (SST) has been compiled as a representative sample of the GOS reference corpus of Spoken Slovenian (Verdonik et al. 2013), a collection of audio recordings and transcripts of spontaneous monologic, dialogic and multi-party speech in different everyday situations. In the first step of the treebank annotation process, the sampled corpus (amounting to approx. 3,200 utterances/30,000 tokens) has been manually lemmatized and annotated using the Universal Dependencies annotation scheme (Nivre et al. 2016), a one-layer morphological and syntactic annotation scheme with a high degree of cross-modality, cross-framework and cross-language interoperability. In this original application of the scheme to spoken language transcripts, the scheme has been extended to address a wide spectrum of syntactic particularities in speech, including disfluencies, discourse markers and other speech-specific phenomena (Dobrovoljc and Nivre 2016).

3 Identification and annotation of multi-word discourse markers

In the second step of the annotation process, an additional annotation layer was introduced in the WebAnno online annotation tool (Yimam et al. 2013), in which multi-word discourse markers were identified and annotated in terms of their discourse organising function. As a point of departure, we followed the annotation scheme proposed by Crible (2014; Crible and Zufferey 2015), aimed at a unified annotation of discourse structuring devices in both written and spoken language. The functional taxonomy proposes 30 different functions, grouped into *ideational* domain, i.e. semantic relations between realword events (e.g. cause, contrast or condition), *rhetorical* domain, i.e. pragmatic relations between speech-acts (e.g. conclusion, specification or reformulation), sequential domain, i.e. structuring relations between segments or topic (e.g. opening/closing, topic shifting or resuming), and *interpersonal* domain, linked to the interaction management (e.g. monitoring, agreeing or disagreeing).

4 Initial analysis of formal and functional properties of multi-word DMs

The initial analysis of the 125 identified types (444 tokens) of multi-word discourse markers shows that most multi-word discourse markers appear in clause initial and final positions, however, unlike in some other fixed word-order Indo-European languages, such as English or French, clause-medial conjunctive adverbials and discourse particles (e.g. *a ne* '~ right') are also quite frequent (e.g. *v bistvu* 'in fact'). Secondly, multi-word discourse markers display a various array of (UD-defined) syntactic functions, from the category of coordinating (e.g. *tako da* 'so that') and subordinating conjunctions (e.g. *kot da* 'as if') to adverbial (e.g. *še posebej* 'especially') and nominal modifiers (e.g. *na primer* 'for example'), paratactical comment clauses (e.g. *bi rekel* '~sort of'), subordinate comment clauses (*kot rečeno* '~as I already said') and conjuncts (e.g. *ali nekaj* 'or something'). In terms of the core functional domain, most multi-word discourse markers in the SST treebank mostly perform rhetorical (i.e. pragmatic) discourse relating functions, such as emphasis, conclusion, hedging and specification, with other markers being equally distributed among other three domains.

5 Conclusions and acknowledgment

The newly developed SST treebank, manually annotated on several linguistic layers, including multi-word discourse markers, represents a valuable language resource for future work on formal, functional and distributional properties of this under-researched type of multi-word expressions, including machine-learning experiments in their automatic identification and disambiguation in larger corpora. The work presented in this paper has been partially supported by the Young Researcher Programme of the Slovenian Research Agency and the Parseme ICT COST Action IC1207 STSM grant.

6 Key references

Crible, L. and S. Zufferey (2015). Using a unified taxonomy to annotate discourse markers in speech and writing. Proceedings of the 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-11).

Dobrovoljc, K. and Nivre, J. (2016). The Universal Dependencies Treebank of Spoken Slovenian. In Proceedings of LREC'16. Portorož, Slovenia.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal Dependencies v1: A multilingual treebank collection. In Proceedings of LREC'16. Portorož, Slovenia.

Prasad, R., A. K. Joshi and B. L. Webber (2010). Realization of discourse relations by other means: alternative lexicalizations. In: *Proceedings of Coling 2010,* Beijing, China, 1023-1031.

Verdonik, D., I. Kosem, A. Zwitter Vitez, S. Krek and M. Stabej (2013). Compilation, transcription and usage of a reference speech corpus: The case of the Slovene corpus GOS*. Language resources and evaluation 47(3)*, 1031-1048.