A Corpus-Driven Approach to Identifying Features of Multi-Word Discourse Markers in Spoken Slovene

Kaja Dobrovoljc
Trojina, Institute for Applied Slovene Studies
Dunajska cesta 116, SI-1000 Ljubljana, Slovenia
Email: kaja.dobrovoljc@trojina.si

Within the traditional discourse structure research, multi-word discourse markers have usually been explored as one of the possible structural realizations for expressing discourse relations in texts, accordingly named either alternative lexicalizations (Prasad et al., 2010), second-level discourse markers (Siepmann, 2005) or secondary connectives (Rysová and Rysová, 2014). With increasing empirical evidence that a considerable amount of human communication is made of prefabricated lexical chunks, stored and retrieved as a whole (Pawley and Syder, 1983; Sinclair, 1991; Wray, 2002), there is a growing need to move the multi-word discourse relational devices from the periphery to the centre of discourse structuring research. This is especially true for spoken language, as corpus-based research proves that spoken communication is substantially more formulaic than written communication (Brazil, 1995; Biber et al., 1999; Erman and Warren, 2000; Leech, 2000) and that formulaic lexical bundles in speech mostly perform pragmatic or discourse functions (Wray and Perkins, 2000; Biber et al., 2004).

To explore the functional and formal particularities of multi-word discourse markers in speech relevant both to their corpus identification and annotation, we present a corpus-driven analysis of the most frequent multi-word units (lexical bundles) in the reference corpus of spoken Slovene (Verdonik et al., 2013), a balanced and representative collection of transcripts of approx. 120 hours (1 million words) of spontaneous speech in various everyday situations.

The list of most frequent multi-word units has been obtained by adapting the statistical substring reduction method proposed by O'Donnell (2011) that adjusts the frequency of items of various lengths when they are part of a larger unit that occurs above a given statistical threshold. The results show that almost half of the 1.282 most frequent lexical bundles bear some sort of procedural meaning, however, their delimitation in terms of whether they actually constitute a class of discourse markers or not shows the need for specific identification criteria in addition to those usually applicable to their one-word counterparts.

From the formal point of view, multi-word discourse markers are not necessarily structurally complete, syntactically independent units, such as prepositional phrases (e.g. *v vsakem primeru,* 'in any case'), complex conjunctions (e.g. *s tem da,* 'wherein'), clauses (e.g. *to se pravi,* 'that is to say') or complex sentences (*veš kaj je,* 'you know what'), as they often take the form of open-end clause segments with subordinating conjunctions (e.g. *to pomeni da,* 'this means that') or open valency slots (e.g. *kar se tiče* _, 'as far as _ is concerned'). Similarly, multi-word discourse markers take on a variety of different syntactic functions, from adverbial modifiers (e.g. *tako ali tako,* 'either way'), coordinating or subordinating conjunctions (e. g. *zaradi tega ker,* 'due to') to independent clauses (e.g. *to je tako,* 'it's like this') or clause segments (e.g. *se pravi da,* 'this means that'). The third important structural observation to make is the large amount of lexical bundles consisting of two or more co-occurring discourse markers: although some are fairly compositional (e.g. *tako da ja*, 'so yes'), others are more grammaticalized (e.g. *no v glavnem,* 'so anyway'), so the distinction between multi-word discourse markers and discourse marker collocations should be properly addressed.

The need for strictly defined formal criteria is even more important given the difficult functional delimitation of multi-word discourse markers in speech, where many frequent expressions, extending to whole utterances or even turns, perform different interpersonal, sequential and other interactive functions, in particularly clause segments introducing opinion (e.g. *jaz mislim da* 'I think that', *moram reči da* 'I have to say that', *po mojem mnenju* 'in my opinion'), meta-commenting clauses (*jaz ne vem*, 'I don't know'; *kaj je že*, 'what's it called'; *kako bi rekel*, 'how should I say this'), signals of active listenership (*ja ja ja*, 'yes yes yes'; *mhm mhm mhm*), responses (*to je res*, 'that's right'), and general extenders (*in tako naprej*, 'and so on', *ali pa kaj* 'or something').

In the context of developing a unified discourse annotation framework, applicable to both spoken and written communication, we thus advocate that special attention should also be given to multi-word discourse relational devices in speech, in particularly to: (i) the relationship between fixed multi-word units and other expressions denoting discourse relations, (ii) the role of syntactic completeness, optionality and structure, and (iii) the relationship between discourse relation markers and other sequential and interpersonal expressions frequent in spoken interaction.

References

Biber, D., Conrad, S., Cortes, V. (2004). If you look at…: Lexical bundles in university teaching and textbooks. *Applied Linguistics 25*, pp. 371–405.

Biber, D., Johansson, S., Conrad, S., Edward, F. (1999). *Longman grammar of spoken and written English*. Harlow: Longman.

Brazil, D. (1995). *A grammar of speech.* Oxford: Oxford University Press.

Erman, B., Warren, B. (2000). The idiom principle and the open choice principle. *Text 20(1)*, pp. 29–62.

Leech, L. (2000). Grammars of spoken English: New outcomes of corpus-oriented research. *Language Learning 50*, pp. 675–724.

Nivre, J. (2015). Towards a universal grammar for natural language processing. In: A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing*, pp. 3–16.

O'Donnell, M. (2011). The adjusted frequency list: A method to produce cluster sensitive frequency list. *ICAME Journal 35*, pp. 135–169.

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

Rysová, M., Rysová, K. (2014). The centre and periphery of discourse connectives. In: *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computation, Cape Panwa Hotel, Phuket, Thailand*, pp. 452–459.

Siepmann, D. (2005). *Discourse markers across languages: a contrastive study of second-level discourse markers in native and non-native text with implications for general and pedagogic lexicography.* London, New York: Routledge.

Sinclair, J. M. (1991). *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

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

Wray, A. (2002). *Formulaic language and the lexicon.* Cambridge: Cambridge University Press.

Wray, A., Perkins, M. (2000). The functions of formulaic language: an integrated model. *Language & Communication 20(1),* pp. 1–28.