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

Kaja Dobrovoljc
Department of Slovenian Studies, Faculty of Arts, University of Ljubljana
Trojina, Institute for Applied Slovenian Studies

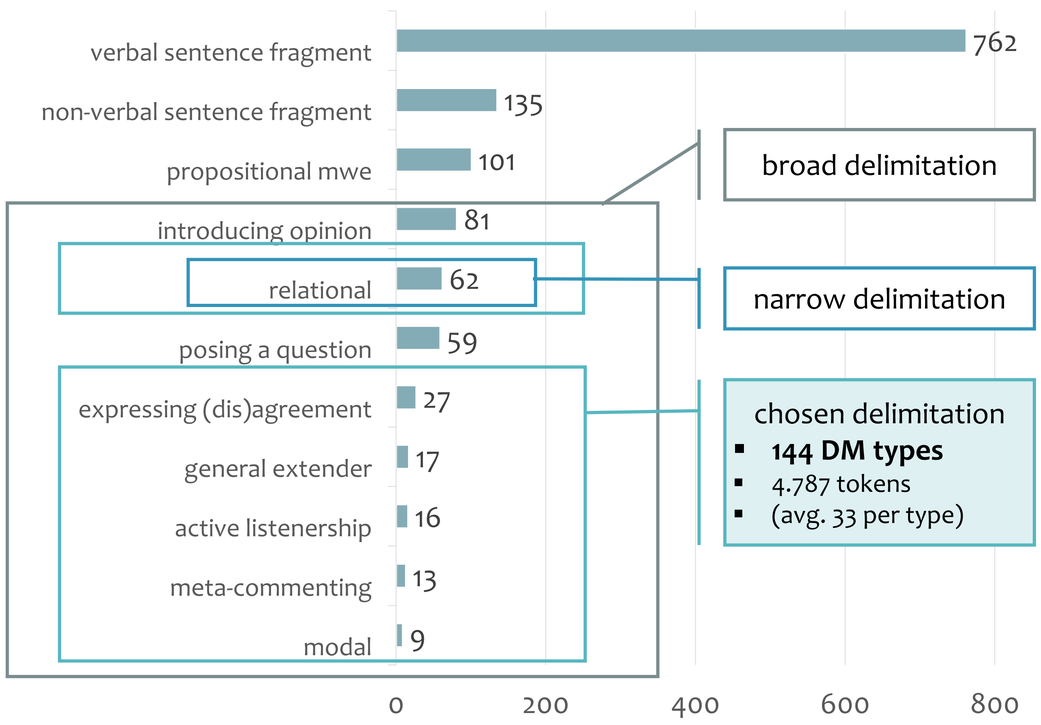
Abstract
With increasing empirical evidence that a considerable amount of spoken communication is made of prefabricated lexical chunks, stored and retrieved as a whole, there is a growing need to move the multi-word discourse relational devices from the periphery to the centre of discourse structuring research. To explore both functional and formal particularities of multi-word discourse markers in speech relevant to their corpus identification and annotation, we present a corpus-driven analysis of the 144 most frequent discourse marking lexical bundles in the reference corpus of spoken Slovene. The results confirm the significant number of fixed multi-word units in the role of discourse structuring devices, constituting a syntactically heterogeneous group of expressions. If we consider multi-word discourse markers to be compositional constructions spanning over verbal predicates, they are both complete and incomplete syntactic constituents, performing various syntactic functions. However, regardless of the degree of their syntactic embedment, their removal does not affect the grammaticality of the host utterance and are thus always syntactically optional.

Multi-Word Discourse Markers
We define multi-word discourse markers as continuous strings of two or more words that have been lexicalized into a semantically non-compositional multi-word unit denoting procedural meaning (Blakemore, 2002). In contrast to complementary research of open-end multi-word lexicalizations of discourse relations (Prasad et al. 2010, Rysová and Rysová 2015), we focus on grammatically invariable **fixed multi-word units**.

Gos Corpus
The Gos reference corpus of spoken Slovenian (Verdonik et al. 2013) is a balanced and representative collection of transcripts of approximately 120 hours (**1 million words**) of spontaneous speech in different everyday situations, such as radio and TV shows, school lessons and lectures, private conversations between friends or family, work meetings, consultations, sales and services, etc., transcribed in pronunciation-based and standardized spelling.

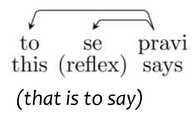
N-gram Extraction Method
The list of most frequent lexical bundles in Gos corpus has been extracted by adapting a cluster-sensitive statistical substring reduction method (O'Donnell 2011), which adjusts the frequency of items of various lengths when they are part of a larger unit occurring at or above a given frequency or statistical threshold. The method was further adapted to extract a list of (i) [1-6]-grams (ii) from standardized spelling transcriptions, (iii) spanning within utterances, (iv) with a minimum relative frequency of 5/mil., and (v) excluding non-lexical tokens, such as silent and filled pauses, vocal and non-vocal sounds. We obtained an adjusted list of 8.301 types of [3-6]-grams, among which **1.282 most frequent [3-6]-grams** were included in the following analysis.

Functional Domain Classification and DM Delimitation



SYNTACTIC COMPLETENESS

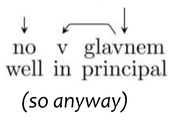
1. COMPLETE (63%)
one complete constituent



(that is to say)

na neki način (*in a way*), v tem primeru (*in this case*), kaj je že (*what's it called*), po eni strani (*on the one hand*), a veš kaj (*you know what*) etc.

2+ complete constituents

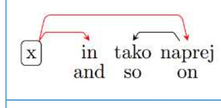


(so anyway)

DM Collocations

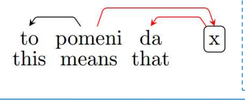
ja ja ja (*yes yey yes*), ja v bistvu (*well in fact*), in zaradi tega (*and due to this*), ja tako je (*yes that's right)*, no tako da (*so this means)*, v redu ja (*ok then)* etc.

2. INCOMPLETE (24%)
(core) head missing



ali pa kaj takega (*or something like that*), ali kaj je (*or what is it)*, pa ne vem kaj še vse (*and I don't know what else)* etc.

(core) dependant missing



kar se tiče (*as far as x is concerned*), kar pomeni da (*which means that*), je pa res da (*it is true however*), v tem smislu da (*in the sense that*) etc.

3. it depends (13%)
For some units, the interpretation of completeness depends on whether we consider the unit to be **compositional or not**, e.g. multiword conjunctions.

(despite the fact that)

MWUs within MWUs

zaradi tega ker (*due to the fact that*), s tem da (*given that),* se pravi da (*that is to say*), glede na to da (*given the fact that)*, v zvezi z (*in relation to*), ker če ne (*otherwise),* s tem ko (*by*), ne samo da *(not only that)* etc.

SYNTACTIC OPTIONALITY

1. OPTIONAL (99%)
Discourse marker detachment does not affect the grammaticality of the host unit, regardless of whether it is complete or not.

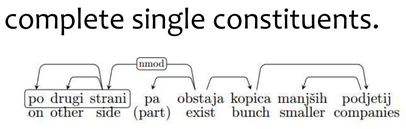
- mhm ~~to se pravi~~ nihče vas še ni učil pravilnih vaj
- *(mhm ~~that is to say~~ you have not been properly trained yet)*

2. it depends(1%)
The interpretation of whether a discourse marker is syntactically optional depends on the **syntactic constraints on its arguments**.

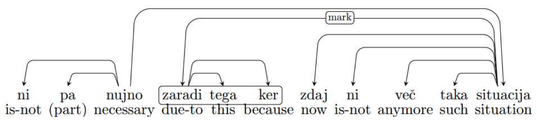
- v zvezi z (*in relation to* + NOUN-instr)
- kar se tiče (*as for* + NOUN-gen)
- ne glede na (*regardless of* + NOUN-acc)

Regardless of how you ***behave***$_{VERB}$ …
Regardless of your ***behaviour***$_{NOUN\text{-}nom}$ …
Regardless of the ***weather***$_{NOUN}$ …

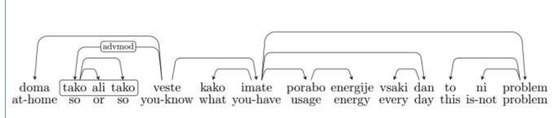
SYNTACTIC FUNCTION
The syntactic relations have been labelled according to the annotation scheme of **Universal Dependencies** (Nivre et al. 2015). **33 different types** of syntactic labels or their combinations have been identified, although some also depend on the interpretation of compositionality of the most grammaticalized multi-word units. Examples below shot the most frequent labels for multi-word discourse markers as complete single constituents.



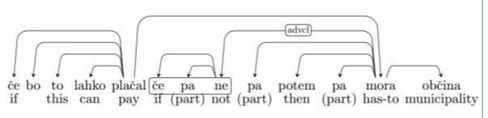
(On the other hand, there is also a bunch of smaller companies.)



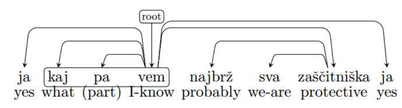
It is not necessary, however, given that the situation has changed.



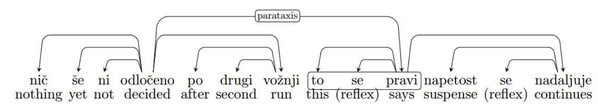
(You know what your domestic energy consumption is anyway; this is not a problem.)



(If he can pay; if not, the municipality has to.)



(Well, I don't know, we're probably being protective, yes.)



(Nothing has been decided in the second run, that is to say, the suspension continues .)

Future Work
We aim to complement this initial quantitive analysis of discourse marking lexical bundles in speech with a qualitative analysis of multi-word discourse markers in the manually annotated dependency **treebank** of spontaneous spoken Slovenian, thus their investigation to two-word markers and context-dependent features, such as **utterance position, situational context** and **prosodic clues.**

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.
- Blakemore, D. (2002). *Relevance and linguistic meaning.* Cambridge: CUP.
- 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*, pp. 1023–1031.
- Rysová, M., Rysová, K. (2015): Secondary Connectives in the Prague Dependency Treebank. In: *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pp. 291-299.
- 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.