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 up 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 end-owed multi-word lexicalizations of discourse relations (Prasad et al. 2010, Rysová and Rysová 2015), we focus on grammatically irrevocable fixed multi-word units.

Gos Corpus
The Gos reference corpus of spoken Slovenian (Verdonik et al. 2015) 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-based 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 (iii) from standardized spelling transcriptions, (ii) spanning within utterances, (iv) with a minimum relative frequency of 5/million, 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

<table>
<thead>
<tr>
<th>Functional Domain</th>
<th>Number of Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>verbal segment</td>
<td>195</td>
</tr>
<tr>
<td>non-verbal segment</td>
<td>201</td>
</tr>
<tr>
<td>propositional move</td>
<td>101</td>
</tr>
<tr>
<td>introducing opinion</td>
<td>81</td>
</tr>
<tr>
<td>relational</td>
<td>62</td>
</tr>
<tr>
<td>posing a question</td>
<td>59</td>
</tr>
<tr>
<td>expressing (dis)agreement</td>
<td>27</td>
</tr>
<tr>
<td>general extender</td>
<td>17</td>
</tr>
<tr>
<td>active listenship</td>
<td>16</td>
</tr>
<tr>
<td>meta-commenting</td>
<td>16</td>
</tr>
<tr>
<td>modal</td>
<td>19</td>
</tr>
</tbody>
</table>

SYNTACTIC COMPLETENESS
1. COMPLETE (63%)
   - one complete constituent
2. INCOMPLETE (24%)
   - (core) head missing
3. it depends (15%)
   - for some units, the interpretation of completeness depends on whether we consider the unit to be compositional or not, e.g. multiword conjunctions.

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.
   - the marker is present in its initial position and otherwise.
2. it depends (%)
   - the interpretation of whether a discourse marker is syntactically optional depends on the syntactic constraints on its arguments.

SYNTACTIC FUNCTION
The syntactic relations have been labelled according to the annotation scheme of Universal Dependencies (Nivre et al. 2013). 33 different types of syntactic labels or their combinations have been identified, although some also depend on the interpretation of compositionality of the most frequent multi-word units. Examples below show the most frequent labels for multi-word discourse markers as complete single constituents.

Identification and Annotation of Discourse Relations in Spoken Language (DiSpol 2015), Saarbrücken, Germany, October 1-2, 2015

References
- Blakemore, S. (2002). Communication depends on... (pp. 71-95).