N-gram Frequency Lists for Reference Corpora of Slovenian

Kaja Dobrovoljc
Jožef Stefan Institute
Centre for Language Resources and Tools UL
Motivation

• **frequency lists** a fundamental tool in corpus linguistics
  • investigations into general/specialized vocabulary, glossaries, language modelling etc.

• shift from words to **multi-word expressions**
  • most work on collocations (association-based approaches)

• increase of research on **formulaic language**
  • lexical bundles, strings ... **n-grams** (frequency-based approaches)

• practical **limitations**
  • software/hardware performance, data restrictions
Corpora

• **GOS** reference corpus of *spoken Slovenian*
  • spontaneous speech in everyday situations (1M tokens)

• **IMP** reference corpus of *historical Slovenian*
  • digitized Slovenian texts from the period 1584-1919 (17M tokens)

• **KRES** reference corpus of modern *written Slovenian*
  • books and periodicals from 1990-2011 (120M tokens)

• **JANES** reference corpus of *user-generated Slovenian*
  • tweets, blogs, forums, news comments and wiki user pages (253M tokens)
N-gram extraction

• set of python scripts

• parameters:
  • token type (e.g. lemma)
  • the size of n (e.g. 3 words)
  • ignoring punctuation (e.g. tako da vs. tako, da)

• 3 types of frequency lists:
  • regular (all n-grams sorted by frequency)
  • filtered (n-grams above a given corpus/text frequency threshold)
  • adjusted
Adjusted frequency list

- difficult to compare n-grams of different lengths
  - substrings always more frequent than parent strings
  - e.g. *glede na* (f = 309) > *glede na to* (f = 178)
  - although realistically *glede na to* (178) > *glede na* (131)

- statistical reduction of substrings
  - O'Donnell 2011: selective reduction of counts in a pre-indexed corpus
  - n-gram counted only if not part of a longer relevant n-gram
    - relevancy defined by min. freq. of occurrence
  - a joint list of all n-grams with a more telling list of types, tokens and rankings
Results

- an adjusted frequency list of **normalized** n-grams of **1-5 tokens** with a freq. $\geq 10/\text{mio.}$, occurring in at least **2 texts** ("core vocabulary")

<table>
<thead>
<tr>
<th>No. of words</th>
<th>GOS</th>
<th>IMP</th>
<th>KRES</th>
<th>JANES</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.371</td>
<td>8.460</td>
<td>10.270</td>
<td>8.914</td>
</tr>
<tr>
<td>2</td>
<td>8.860</td>
<td>6.087</td>
<td>4.885</td>
<td>5.984</td>
</tr>
<tr>
<td>3</td>
<td>3.199</td>
<td>1.131</td>
<td>901</td>
<td>1.110</td>
</tr>
<tr>
<td>4</td>
<td>244</td>
<td>43</td>
<td>32</td>
<td>68</td>
</tr>
<tr>
<td>5</td>
<td>47</td>
<td>6</td>
<td>11</td>
<td>171</td>
</tr>
<tr>
<td>SUM</td>
<td>18.721</td>
<td>15.727</td>
<td>16.099</td>
<td>16.247</td>
</tr>
</tbody>
</table>
Structure and frequency

- large no. of multi-word units > formulaicity of human communication
- esp. in spontaneous speech, in which multi-word strings prevail
- not just bigrams!
### Overlap of core vocabulary

<table>
<thead>
<tr>
<th></th>
<th>% of overlapping n-grams</th>
<th>% of unique</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>in GOS</td>
<td>in IMP</td>
</tr>
<tr>
<td>GOS</td>
<td>34,0</td>
<td>42,5</td>
</tr>
<tr>
<td>IMP</td>
<td>40,5</td>
<td>49,1</td>
</tr>
<tr>
<td>KRES</td>
<td>49,5</td>
<td>48,0</td>
</tr>
<tr>
<td>JANES</td>
<td>55,6</td>
<td>42,0</td>
</tr>
</tbody>
</table>

- a large number of unique n-grams in all corpora
  - some differences in annotation principles (e.g. [@per] [URL])
  - distinct vocabulary
Conclusion

• collection of n-gram frequency lists for selected corpora
  • regular, filtered and adjusted
  • 1-5 normalized tokens
  • CLARIN.SI repository (CC-BY-SA)

• first quantitative comparison
  • similar structure and frequency of core vocabulary
  • a substantial amount of formulaic sequences
  • only partial overlap of n-grams
Future work

• open-source tool (NSSS, CLARIN.SI)
  • computationally efficient
  • TEI-friendly
  • GUI
  • new functionalities

• new lists

• linguistic analysis
  • categorization by structure and function
Thanks!