Sharing data through specialized corpus-based tools: the case of GramatiKat

Dominika Kováříková

Slovko, October 2021
Main points

▶ how to share data (and get something in return)
▶ quantitative analysis of GK is worthy of our attention
▶ experience with creating pre-processed and pre-analyzed data

representative corpus of contemp. written Czech, 100 million words
SYN2005, SYN2015, SYN2020
Reasons to share data (NŘ, March 2021)

1. falsifiability/verifiability
2. incentive to be thorough in your research
3. replicability and possibility to further develop the original idea
4. synergy and opening new research possibilities
5. student training
6. social responsibility (publicly funded research)
7. FAIR data and possibility to re-interpret results
8. publishing the data as an regular research output
9. meta-analysis

(Chromý & Cvrček, 2021)
Reasons to share data (NŘ, March 2021)

1. falsifiability/verifiability
2. incentive to be thorough in your research
3. replicability and possibility to further develop the original idea
4. synergy and opening new research possibilities
5. student training
6. social responsibility (publicly funded research)
7. FAIR data and possibility to re-interpret results
8. publishing the data as an regular research output
9. meta-analysis

(Chromý & Cvrček, 2021)
% OF GRADED ADJECTIVES
(in SYN2020, 100 million words corpus)

- ALL adjectives
- at least one graded form
- comparative and superlative
# GramatiKat: Pre-processed data

<table>
<thead>
<tr>
<th>LEMMA</th>
<th>ROD</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
<th>S6</th>
<th>S7</th>
<th>P1</th>
<th>P2</th>
</tr>
</thead>
<tbody>
<tr>
<td>bača</td>
<td>M</td>
<td>0.48</td>
<td>0.05</td>
<td>0.02</td>
<td>0.07</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>beruška</td>
<td>F</td>
<td>0.23</td>
<td>0.07</td>
<td>0.01</td>
<td>0.09</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>brach</td>
<td>M</td>
<td>0.15</td>
<td>0</td>
<td>0.22</td>
<td>0.05</td>
<td>0.33</td>
<td>0.01</td>
<td>0</td>
<td>0.14</td>
<td>0.05</td>
</tr>
<tr>
<td>čubčka</td>
<td>F</td>
<td>0.46</td>
<td>0.02</td>
<td>0.03</td>
<td>0.09</td>
<td>0.26</td>
<td>0.01</td>
<td>0.05</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>děvenka</td>
<td>F</td>
<td>0.08</td>
<td>0.02</td>
<td>0.01</td>
<td>0.01</td>
<td>0.73</td>
<td>0</td>
<td>0.01</td>
<td>0.03</td>
<td>0.02</td>
</tr>
<tr>
<td>drahoušek</td>
<td>M</td>
<td>0.08</td>
<td>0.01</td>
<td>0.42</td>
<td>0.01</td>
<td>0.42</td>
<td>0.02</td>
<td>0</td>
<td>0.03</td>
<td>0</td>
</tr>
<tr>
<td>hajzl</td>
<td>M</td>
<td>0.26</td>
<td>0.04</td>
<td>0.03</td>
<td>0.12</td>
<td>0.34</td>
<td>0.01</td>
<td>0.01</td>
<td>0.13</td>
<td>0.04</td>
</tr>
<tr>
<td>inspektorka</td>
<td>F</td>
<td>0.47</td>
<td>0.07</td>
<td>0.05</td>
<td>0.07</td>
<td>0.24</td>
<td>0</td>
<td>0.05</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>kámoš</td>
<td>M</td>
<td>0.22</td>
<td>0.04</td>
<td>0.04</td>
<td>0.05</td>
<td>0.3</td>
<td>0</td>
<td>0.04</td>
<td>0.13</td>
<td>0.05</td>
</tr>
<tr>
<td>lhárka</td>
<td>F</td>
<td>0.56</td>
<td>0.02</td>
<td>0</td>
<td>0.1</td>
<td>0.21</td>
<td>0</td>
<td>0.05</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>miláček</td>
<td>M</td>
<td>0.14</td>
<td>0.04</td>
<td>0.3</td>
<td>0.06</td>
<td>0.24</td>
<td>0.01</td>
<td>0.07</td>
<td>0.04</td>
<td>0.02</td>
</tr>
<tr>
<td>mylord</td>
<td>M</td>
<td>0.02</td>
<td>0.01</td>
<td>0</td>
<td>0</td>
<td>0.98</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>pán</td>
<td>M</td>
<td>0.17</td>
<td>0.05</td>
<td>0.03</td>
<td>0.04</td>
<td>0.49</td>
<td>0.01</td>
<td>0.03</td>
<td>0.09</td>
<td>0.03</td>
</tr>
<tr>
<td>píča</td>
<td>F</td>
<td>0.32</td>
<td>0.15</td>
<td>0.04</td>
<td>0.13</td>
<td>0.22</td>
<td>0.06</td>
<td>0.06</td>
<td>0.01</td>
<td>0</td>
</tr>
<tr>
<td>signor</td>
<td>M</td>
<td>0.35</td>
<td>0.09</td>
<td>0.02</td>
<td>0.07</td>
<td>0.46</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>sir</td>
<td>M</td>
<td>0.44</td>
<td>0.15</td>
<td>0.07</td>
<td>0.04</td>
<td>0.26</td>
<td>0.01</td>
<td>0.03</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Standard behavior of a word class

Standard case distribution in Czech nouns (SYN2015!)
Grammatical profiles of lemmas

distribution of grammatical category values within individual lemmas

→ e.g. distribution of case in nouns
Grammatical profiles of lemmas

třpytka ‘glitter’
Grammatical profiles of lemmas

\[ \text{šílenství} \quad \text{‘madness’} \]
Expect the unexpected

- anomaly is quite common (14 ways to go wrong)
- at least one anomalous paradigm cell in 40% of nouns (missing or excess)
- frequency dependent phenomenon!
Expect the unexpected

▶ anomaly is quite common (14 ways to go wrong)
▶ at least one anomalous paradigm cell in 40% of nouns (missing or excess)
▶ frequency dependent phenomenon!

⇒ blurry boundary between lexicon and grammar
Lists of anomalous lemmas

emerging patterns
(semantic fields, grammatical phenomena...)

⇒ Extremely high percentage of Locative plural:

  v džínsách (‘in jeans’)
  v montérkách (‘in ovealls’)
  v kraťasech (‘in shorts’)
  v trenýrkách (‘in boxer shorts’)
  v šortkách (‘in shorts’)
  ve spodkách (‘in long johns’)
  ve slipech (‘in underpants’)
  v teplákách (‘in sweatpants’)
  v boxerkách (‘in boxers’)

APPLICATIONS

▶ morphology
▶ lexicography
▶ teaching on all levels, speech therapy, CzSL
▶ theoretical research of paradigm defectivity and language potentiality
Morphology

- in-depth analysis of grammatical categories
- synergy of grammatical and lexicological research
- comparison of languages
Czech vs Slovak sg

InterCorp parallel corpus, 50 mil word for Czech and for Slovak
Lexicography (academic Czech)

důsledek (N) ‘consequence’
Lexicography (academic Czech)

v důsledku (preposition) ‘as a result of’
Defectivity and language potentiality

Project Feast and Famine: Confrontation of overabundance and defectivity in language

Internetová jazyková příručka
Janda and Tyers (2018) suggest that “learning may be enhanced by focusing only on the word forms most likely to be encountered” (p. 28).

cizina ‘foreign countries, abroad’
Conclusion

- sharing data has its benefits: more ideas, more collaborations
- sharing ideas is easier, if we have too many (↑)
- quantitative research of GK is fascinating
References


