Framework for an Analysis of Slovene Regional Language Variants on Twitter

4th Conference on CMC and Social Media Corpora for the Humanities

Jaka Čibej

Faculty of Arts, University of Ljubljana

Ljubljana, 28 September 2016
Outline

1. Introduction and Motivation
2. Dataset Preparation
3. Typology of Non-Standard Slovene Language Elements on Twitter
4. Dataset Annotation
5. Results
6. Measures of Regional Specificity
7. Conclusion
Introduction

- high degree of regional language variation
- ~48 dialects (7 main dialect groups)
Motivation

- regional language variants in user-generated content
  - in Slovene UGC (Erjavec & Fišer 2013; Zwitter Vitez & Fišer 2015)
- linguistic analysis
**Dataset Preparation**

- 130,000 geotagged tweets from the JANES corpus of Slovene UGC (Fišer et al. 2016)
Regional Subcorpora

- 9 regional subcorpora (7 main dialectal groups + Ljubljana and Maribor)
- 90% threshold
Sampling

- Sampling criteria:
  - 500 tweets per region
  - only private users
  - non-standard tweets (L3) (Ljubešić et al. 2015)
  - all users included
  - max. 30–50 random tweets per user

- This paper: Primorska, Gorenjska, Štajerska
Typology of Non-Standard Slovene Language Elements

- manual analysis, bottom-up approach
- 7 main categories, 105 different tags
  - non-standard vocabulary (ejga, čuj, nanka)
  - reductions and ellipses (čudno \(\rightarrow\) čudn_)
  - alternative graphemes (ne vem \(\rightarrow\) ne wem)
  - non-standard morphology (imate \(\rightarrow\) imaste)
  - spelling variants of frequent standard words (jaz \(\rightarrow\) jz, js, jst, jest, jast...)
  - frequent transformations (-aj- to -ej-, e.g. neka\(j\) \(\rightarrow\) neke\(ej\), včera\(j\) \(\rightarrow\) včere\(ej\))
  - miscellaneous (to je \(\rightarrow\) toj)
@RoganMatevz Ne ta vikend ne, [nasledn]{Rj.nj}{RksPme.i} pa [mogoč]{RksR.e} ... boš ti [nasledn]{Rj.nj}{RksPme.i} vikend doma? :) Rose me čaka ne? :D
## (Preliminary) Annotation Results

<table>
<thead>
<tr>
<th>Category</th>
<th>Primorska</th>
<th>Gorenjska</th>
<th>Štajerska</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-standard vocabulary</td>
<td>394</td>
<td>347</td>
<td>371</td>
</tr>
<tr>
<td>Spelling variants of frequent standard words</td>
<td>233</td>
<td>322</td>
<td>183</td>
</tr>
<tr>
<td>Alternative graphemes</td>
<td>40</td>
<td>54</td>
<td>34</td>
</tr>
<tr>
<td>Reductions and ellipses</td>
<td>588</td>
<td>1122</td>
<td>648</td>
</tr>
<tr>
<td>Non-standard morphology</td>
<td>90</td>
<td>99</td>
<td>67</td>
</tr>
<tr>
<td>Frequent transformations</td>
<td>120</td>
<td>181</td>
<td>68</td>
</tr>
<tr>
<td>Miscellaneous</td>
<td>39</td>
<td>59</td>
<td>24</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>1504</strong></td>
<td><strong>2184</strong></td>
<td><strong>1395</strong></td>
</tr>
</tbody>
</table>
Measures of Regional Specificity and Dispersion

- **relative frequency** \((f_R)\)
  - element occurrences / category occurrences
  - higher \(f_R\) → occurs more often within region
  - lower \(f_R\) → occurs less often within region
Measures of Regional Specificity and Dispersion

- **user ratio** \((u)\)
  - percentage of users in region using the element
  - higher \(u \rightarrow\) more widespread in region
  - lower \(u \rightarrow\) less widespread in region
Measures of Regional Specificity and Dispersion

- **type/token ratio** ($t$)
  - penalises elements occurring often, but with a limited number of types
  - Example: kmal`u ('soon') $\rightarrow$ kmal_ (final u ellipsis)
**Measures of Regional Specificity and Dispersion**

- **annotation ratio** \( (a) \)
  - tags element is used with / all tags in category
  - penalises elements occurring only with a limited number of tags (e.g. only with adverbs)
  - higher \( a \) → occurs in more word positions and with more PoS categories
  - lower \( a \) → fewer positions, fewer PoS categories
Measures of Regional Specificity and Dispersion

- coefficient of regional dispersion ($\delta_R$)
  - summary of all measures
  - higher $\delta_R \rightarrow$ more widespread and regionally specific

$$\delta_R = f_R \times u \times t \times a \times 100$$
**Example: Final -i ellipsis**

<table>
<thead>
<tr>
<th></th>
<th>Gorenjska</th>
<th>Štajerska</th>
<th>Primorska</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_R$</td>
<td>0.52</td>
<td>0.60</td>
<td>0.47</td>
</tr>
<tr>
<td>$u$</td>
<td>0.75</td>
<td>0.42</td>
<td>0.54</td>
</tr>
<tr>
<td>$t$</td>
<td>0.61</td>
<td>0.53</td>
<td>0.67</td>
</tr>
<tr>
<td>$a$</td>
<td>0.43</td>
<td>0.41</td>
<td>0.24</td>
</tr>
<tr>
<td>$\delta_R$</td>
<td>10.25</td>
<td>5.41</td>
<td>4.08</td>
</tr>
</tbody>
</table>
Conclusion

- dataset and typology for an analysis of Slovene regional language variants on Twitter
- measures of regional specificity and dispersion
- Future work:
  - annotation of all regions
  - finalisation of the typology
  - comparison with existing dialectological studies
  - comparison with statistical tests
Thank you.