Framework for an Analysis of Slovene Regional Language Variants on Twitter

Jaka Čibej

Department of Translation, Faculty of Arts, University of Ljubljana

Aškerčeva 2, 1000 Ljubljana

E-mail: jaka.cibej@ff.uni-lj.si

Abstract

The rapid rise of computer-mediated communication has allowed regional language variation to flourish in written form, opening new doors both for dialectological studies as well as natural language processing. In this paper, we present the methodology and framework for a linguistic analysis of Slovene regional language variants on Twitter. We describe the creation and sampling of a dataset stratified by region, present a preliminary typology of non-standard Slovene language elements on Twitter, and propose an approach to measure regional specificity and dispersion of non-standard language elements in computer-mediated communication.

Keywords: regional language variants, non-standard Slovene, Twitter, computer-mediated communication

1. Introduction

In the last two decades, the rapid rise of computer-mediated communication and social media has allowed language to spread into digital communication platforms, lending a voice to a plethora of different languages traditionally present only in spoken varieties: from sociolects to dialects and everything in between. In addition, due to their ever increasing quantities, internet texts have become an important source of information, and there is an increasing demand for tools and resources to help process them, as shown by the proliferation of different areas within internet linguistics and natural language processing. One of the problematic aspects to be tackled in this regard is regional language variation.

The main goal of this paper is to present a methodology and framework for a linguistic analysis of regional language variants on Twitter. The paper is structured as follows: first, we present a brief overview of related work, which is followed by the description of our dataset and the sampling methods used. We then provide an overview of the preliminary typology of non-standard Slovene language elements on Twitter and the measures of regional specificity and dispersion to be used in further analyses, and conclude with the preliminary results of the analysis of three regional samples.

2. Related Work

Studies on regional variation of various languages on Twitter have been conducted with different purposes, mostly as part of development of NLP tools, e.g. diacritic restoration (Harrat et al., 2013) and POS-tagging (Bernhard & Ligozat, 2013), but also within sociological studies of language variation (Jørgensen et al., 2015; Eisenstein, 2015).

Slovene regional language variation on Twitter (and in social media in general), however, is currently still an under-researched area that cannot be neglected, especially considering the rich dialectal variation of Slovene (Ramovš, 1931), the numerous dialectological studies conducted on spoken Slovene (Kenda Jež, 2002), as well as the fact that regional variation has already been documented in Slovene tweets (Fišer et al., 2015b).

3. Dataset Preparation

The dataset presented in this paper consists of tweets extracted from the JANES corpus of Slovene user-generated content (Fišer et al., 2015a). The tweets were sampled by taking into account a number of criteria.

First, only tweets sent from private accounts were included, while tweets from corporate accounts (e.g. those managed by press agencies and companies) were eliminated.[1] This was done for two reasons: corporate accounts contain many automatically generated tweets, while the overwhelming majority of their original tweets are written in standard Slovene, which makes them irrelevant for our study.

Second, the dataset only includes L3 tweets, i.e. those with a high level of linguistic non-standardness (Ljubešić et al., 2015). L3 tweets contain a high degree of non-standard spelling and vocabulary and as such provide the most material for the study of regional language variants.

Third, the tweets were sampled by taking into account the metadata on the users' regional origin (Čibej & Ljubešić, 2015). This metadata was determined by collecting Slovene geotagged tweets over a period of eighth months (from January 2015 to September 2015), then assigning each user with geotagged tweets to one of 9 regions corresponding to the 7 main dialectal groups of Slovene as well as Ljubljana and Maribor, the two largest cities, which we decided to treat separately as melting pot areas. In order to exclude users with ambiguous origin, only users that sent more than 90% of their tweets from a single region were taken into account. A certain amount of noise is to be expected in the dataset despite this criterion, but should not prove too prominent and will be further penalised during the analysis (see Section 6).

---

[1] Twitter users included in the JANES corpus were manually annotated as corporate or private.

Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities, Ljubljana, Slovenia, 27–28 September 2016

17

Regional subcorpus	Number of tokens	Number of tweets	Number of users
Gorenjska	37,683	22,070	48
Dolenjska	17,364	6,922	22
Štajerska	41,712	9,284	42
Panonska	5,020	2,512	14
Koroška	6,207	4,203	5
Primorska	13,917	5,748	31
Rovtarska	4,823	2,348	7
Ljubljana	92,104	43,018	116
Maribor	4,789	4,340	14

Table 1: Size of regional subcorpora in the JANES corpus of Internet Slovene (v0.3).

As shown in Table 1, some of the regional subcorpora are very small both in terms of the number of tokens as well as the number of users included. However, geotagged tweets are still being collected, and more users and tweets will be added when the corpus is updated. For the purposes of this paper, we focus on three of the best represented regions: Primorska, Gorenjska, and Štajerska.

3.1. Samples of Regional Subcorpora

For each region, a sample containing 500 L3 tweets was created. First, all tweets were extracted from the relevant regional subcorpus. The tweets were then shuffled and sampled by user in order to avoid overrepresentation of very prolific Twitter users. In some cases, the most active users provided more than 2,000 tweets to a regional subcorpus, while the least active provided less than 10.

The samples included all users from the relevant regional subcorpus, while the number of tweets each user contributed was limited to a maximum of 40–50 tweets (depending on the total number of users).

4. Typology of Non-Standard Slovene Language Elements on Twitter

Small subsets of 100–150 tweets were manually analysed in each sample in order to design a typology of non-standard Slovene language elements on Twitter. The typology was created with a bottom-up approach and so far includes 7 main categories: non-standard vocabulary, reductions and ellipses, non-standard morphology, spelling variants of frequent standard words, alternative graphemes, frequent transformations, and miscellaneous.[2] Currently, the typology consists of 105 different tags, but is flexible and allows for the addition of new elements as certain rare or regionally specific elements (especially those concerning morphology and syntax) may yet arise during annotation. In the following subsections, we present the main categories in further detail.

4.1. Non-Standard Vocabulary

Non-standard vocabulary includes all lexical elements that are considered non-standard, i.e. those that would not be expected in standard Slovene texts and/or are not included in existing standard language resources such as dictionaries or lexicons. Examples include regionally specific words (e.g. particles *ejga* for Gorenjska, *čuj* for Štajerska, *nanka* for Primorska), standard words with new meanings (*hudo* meaning 'awesome' instead of 'bad'), and non-standard words/phrases of foreign language origin, either in their original spelling (e.g. *web app*) or fully/partially adapted to Slovene spelling and morphology (e.g. *ekskjuz*, from English 'excuse'; *učelini*, from Italian 'uccellini').

A subcategory of non-standard vocabulary also included certain CMC-specific abbreviations, either English (*wtf*, *lol*, *omg*) or Slovene (*jbg* „fuck that", *bmk* „I don't give a fuck"), and alphanumerical spellings (*ju3* for *jutri*, 'tomorrow').

4.2. Reductions and Ellipses

With 69 different tags, reductions and ellipses are by far the most prolific category. Most often, they involve vowel drops in different positions in a word. A common example is the ellipsis of the final *-i* in the infinitive (*delati* → *delat*, 'to work') or the final *-o* in adverbs (*čudno* → *čudn*, 'weirdly, oddly'). As for consonants, a common example is the ellipsis of *-j* in the *-lj-* or *-nj-* consonant clusters (*peljem* → *pelem*, 'I drive'; *zadnji* → *zadni*, 'the last').

4.3. Alternative Graphemes

This category encompasses alternative, non-standard spellings of graphemes, most often in cases when it is pronounced differently in spoken language. Examples include the spelling of *g* as *h* (*bog* → *boh*, 'god') or *v* as *w* (*ne vem* → *ne wem*, 'I don't know').

4.4. Non-Standard Morphology

This category included words that exhibited non-standard morphological characteristics such as alternative case endings (e.g. the non-standard locative ending *-i* of singular masculine nouns, *na šihtu* → *na šihti*, 'at work') or other regionally specific suffixes (e.g. the non-standard second-person plural verb suffix *-ste* instead of *-te*, *imate* → *imaste* „you have").

4.5. Spelling Variants of Frequent Standard Words

The category of spelling variants includes common standard (mostly function) words with numerous spelling variants that are unequally distributed between different regions. A good example is the word *toliko* ('this much, so'), which can also be spelt as *tok*, *tolk*, *tolko*, *telko*, *tuk*, *tulk*, etc. Similarly, the word *jaz* (personal pronoun, first person singular, 'I') can also be encountered as *jz*, *js*, *jst*, *jest*, *jes*, etc. Although these spelling variants often also include other non-standard elements (e.g. vowel ellipses), they are also annotated as a separate category in order to produce an exhaustive list so that their regional distribution can be tested on the entire geolocated JANES subcorpus.

---

[2] Initially, a syntactic category was included, but was later omitted as syntactic elements were much too scarce in the samples. However, potential regionally specific syntactic features encountered during the analysis will be researched on larger amounts of data in the JANES corpus of Internet Slovene.

Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities, Ljubljana, Slovenia, 27–28 September 2016
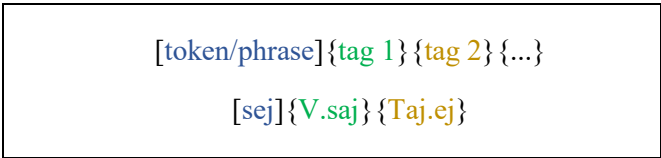
18

4.6. Frequent Transformations

This is an additional category that included spelling transformations in non-standard word spellings that were perceived during annotation as frequently occurring. Similar to frequent spellings of non-standard words, these transformations were annotated separately in order to allow for a comparison of their distributions in different regions. A prevalent example is the transformation of *-aj-* to *-ej-* (*nekaj → nekej*, 'something'; *včeraj → včerej*, 'yesterday') or *-aj-* to *-j-* (*zdajle → zdjle*, 'now'; *kaj → kj*, 'what').

4.7. Miscellaneous

The final category included miscellaneous non-standard language elements that could not be categorised in any of the previous categories. These mainly consisted of joint spellings, i.e. instances where two words should be spelt separately in standard Slovene, but are written together in their non-standard form (e.g. *ne vem → nevem*; 'I don't know, I dunno'), or amalgams of two adjacent words, most often function words (*to je → toj*, 'this is', *če je → čej*, 'if it is').

5. Dataset Annotation

The samples were manually annotated in .txt format to enable flexible post-processing and analysis with Python regular expressions. Relevant tokens or phrases were annotated as shown in Figure 1.

```
[token/phrase]{tag 1}{tag 2}{...}

[sej]{V.saj}{Taj.ej}
```

Figure 1: Annotations.

The upper line shows the general pattern of annotation. A single token or phrase may be annotated with multiple tags. The bottom line shows an example of the annotated word *sej* ('because'), annotated both as a spelling variant of *saj* (*V.saj*) and as a frequent spelling transformation of *-aj-* to *-ej-*. (*Taj.ej*).

Several language elements were excluded from annotation. These included a number of CMC-specific elements (emoticons and emojis, hashtags or URLs), spelling mistakes that were perceived as obviously accidental, as well as the non-use of diacritics, which is often a consequence of technical limitations and rarely voluntary. Code-switching, although relatively common, was also omitted. If foreign language words or phrases were used as part of a Slovene sentence, they were annotated as non-standard vocabulary. Entire sentences or independent units in foreign language, however, were disregarded. The same was true of non-standard variants of proper nouns (e.g. phoneticised versions of Twitter and Facebook – *Tviter*, *Fejsbuk*).

6. Measures of Regional Specificity and Dispersion

In addition to statistical tests, we also propose a method to determine the level of regional specificity of a certain language element based on a number of criteria described in the following subsections. In addition, these measures should help to reduce the effect of potential noise in the dataset (e.g. users that are originally from a different region and have permanently moved to a different one, but continue to use language elements typical of their region of origin).

6.1. Relative Frequency

Relative frequency (f_R) is the ratio of the frequency of a language element and the total number of occurrences in its category. The greater the relative frequency, the more frequent the language element within the region.

6.2. User Ratio

The user ratio (u) is the ratio of the number of users using a language element and the number of all users from the region in question. The greater the number of users that use a language element, the greater the user ratio. This value thus measures how widespread the element is among the users of the region. It penalises idiosyncratic elements (especially with prolific users) or elements used by users that have been misclassified as pertaining to a specific region.

6.3. Type/Token Ratio

The type/token ratio (t) is the ratio of the number of types and the number of tokens used with a language element. The greater the t-ratio, the greater the number of words it occurs with, and the greater the likelihood that the element will arise in text. This value penalises frequent language elements that only occur in a limited number of words.

6.4. Annotation Ratio

Similar to the type/token ratio, the annotation ratio (a) is the ratio of the number of different tags the element occurs with and the number of all tags in its category. The greater the annotation ratio, the greater the number of tags it occurs with and the greater the likelihood it occurs.

6.5. Coefficient of Regional Dispersion

The coefficient of regional dispersion (δ_R) is meant as a simple summarisation of all other measures of regional specificity and dispersion. It is calculated as follows:

$$\delta_R = f_R \times u \times t \times a \times 100$$

The greater the coefficient of regional dispersion, the more widespread and frequent the element in question.

7. Annotation Results

In this section, we provide some of the preliminary results of the annotated dataset and demonstrate the use of the abovementioned f_R, u, t, a and δ_R values to measure regional specificity and dispersion for a particular language element.

Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities,
Ljubljana, Slovenia, 27–28 September 2016

19

The annotation results for the Gorenjska, Štajerska and Primorska regions are shown in Table 2. As all samples are of comparable size (they consist of 500 tweets each), the absolute frequencies are given.

Category	Primorska	Gorenjska	Štajerska
Non-standard vocabulary	394	347	371
Spelling variants of frequent standard words	233	322	183
Alternative graphemes	40	54	34
Reductions and ellipses	588	1122	648
Non-standard morphology	90	99	67
Frequent transformations	120	181	68
Miscellaneous	39	59	24
Total	1504	2184	1395

Table 2: Quantitative Analysis of Annotated Samples.

The regions do not differ to a great extent in terms of the frequency of non-standard vocabulary, although we expect that a detailed qualitative analysis will show differences in the type of non-standard words used (e.g. we expect to find more words originating from Italian in the Primorska region, which lies next to the border with Italy).

As far as the frequencies of other categories are concerned, the differences between the three regions are more pronounced. What is particularly interesting to note is that while reductions and ellipses are the most prolific category in all three regions, they are especially frequent in the Gorenjska region. The most frequent type of ellipsis in all three regions was the -*i* ellipsis. The frequencies of final and non-final -*i* ellipses are shown in Table 3, along with χ^2 p-values and Cramer's V effect sizes.

	Gorenjska vs. Primorska		Gorenjska vs. Štajerska		Primorska vs. Štajerska	
Final -*i* ellipsis	254	140	254	174	140	174
Non-final -*i* ellipsis	231	156	231	118	156	118
χ^2 p-value	>0.05		>0.05		0.037	
Cramer's V	0.05		0.07		0.12	

Table 3: χ^2 p-values and Cramer's V effect sizes for distributions of final vs. non-final -*i* ellipses.

The only statistically significant difference in the distribution of final vs. non-final -*i* ellipses is the one between Primorska and Štajerska, with a small, but not entirely negligible effect size. It would appear Štajerska slightly prefers final -*i* ellipsis to non-final -*i* ellipsis.
Table 4 shows the measures of regional specificity and dispersion for the final -*i* ellipsis for all three regions.

	Gorenjska	Štajerska	Primorska
f_R	0.52	0.60	0.47
u	0.75	0.42	0.54
t	0.61	0.53	0.67
a	0.43	0.41	0.24
δ_R	10.25	5.41	4.08

Table 4: Measures of regional specificity and dispersion for the final -*i* ellipsis.

As can be deduced from Table 4, the final -*i* ellipsis has a significantly greater user ratio in Gorenjska, as well as a significantly higher coefficient of regional dispersion, which would indicate that the language element is much more widespread in this region compared to Štajerska and Primorska.

8. Conclusion

In the paper, we described the creation of a dataset for the analysis of Slovene regional language variants on Twitter and presented a method for the analysis of regional language variants on Twitter.
In our future work, we will perfect the typology of non-standard language elements in Slovene CMC and make a comparison with phenomena presented in existing Slovene dialectological studies. We will also extend the annotated dataset to other Slovene regions and analyse all encountered language elements in terms of their regional specificity and dispersion, then compare the results with the results obtained through other statistical methods. In addition, elements that rarely occur in the samples (e.g. non-standard syntactic constructions) will be tested on larger text samples in the JANES corpus of Internet Slovene. The results of the analysis will be used to design features to be used in the development of a model for the automatic recognition of Slovene regional language variants on Twitter.

9. Acknowledgments

10. References

Fišer, D. Ljubešić, N. & Erjavec, T. (2015a). The JANES corpus of Slovene user generated content: construction and annotation. *International Research Days: Social Media and CMC Corpora for the eHumanities: Book of Abstracts, 23–24 October 2015*. Rennes, France, p. 11.

Ljubešić, N., Fišer, D., Erjavec, T., Čibej, J., Marko, D., Pollak, S. & Škrjanec, I. (2015). Predicting the level of text standardness in user-generated content. *10th International Conference on Recent Advances in Natural Language Processing: Proceedings of RANLP 2015 Conference, 7–9 September 2015*. Hissar, Bulgaria, pp. 371–378.

Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities, Ljubljana, Slovenia, 27–28 September 2016

20

Jørgensen, A. K., Hovy, D. & Søgaard, A. (2015). Challenges of studying and processing dialects in social media. *Proceedings of the ACL 2015 Workshop on Noisy User-generated Text*. Beijing, China, July 31, 2015, pp. 9–18.

Harrat, S., Abbas, M., Meftouh, K. & Smaili, K. (2013). Diacritics restoration for Arabic dialect texts. *Proceedings of the 14th Annual Conference of the International Speech Communication Association (INTERSPEECH 2013)*. France.

Bernhard, D. & Ligozat, A.-L. (2013). Hassle-free POS-Tagging for the Alsatian Dialects. Zampieri, M. & Diwersy, S. (eds.), *Non-standard Data Sources in Corpus-based Research*. Aachen: Shaker Verlag, pp. 85–92.

Čibej, J. & Ljubešić, N. (2015). "S kje pa si?" – Metapodatki o regionalni pripadnosti uporabnikov družbenega omrežja Twitter. Fišer, D. (ed.), *Proceedings of Konferenca Slovenščina na spletu in v novih medijih*. Ljubljana, ZIFF, pp. 10–14.

Kenda Jež, K. (2002). *Cerkljansko narečje: teroetični model dialektološkega raziskovanja na zgledu besedišča in glasoslovja*. PhD dissertation. Ljubljana: Faculty of Arts.

Eisenstein, J. (2015). Written dialect variation in online social media. Boberg, C., Nerbonne, J. & Watt, D. (eds.): *Handbook of Dialectology*. Wiley.

Ramovš, F. (1931). Dialektološka karta slovenskega jezika. Ljubljana: Rektorat univerze kralja Aleksandra I. in J. Blaznika nasl. – Univerzitetna tiskarna.

Fišer, D., Erjavec, T., Čibej, J. & Ljubešić, N. (2015). Gradnja in analiza korpusa spletne slovenščine JANES. Smolej, M. (ed.): OBDOBJA 34: Slovnica in slovar – aktualni jezikovni opis. Ljubljana: ZIFF, pp. 217–223.

Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities,
Ljubljana, Slovenia, 27–28 September 2016

21