Comparing the non-standard language of Slovene, Croatian and Serbian tweets

Darja Fišer[1], **Tomaž Erjavec**[2], **Nikola Ljubešić**[2,3], **Maja Miličević**[4]

[1] Department of Translation, Faculty of Arts, University of Ljubljana, Aškerčeva 2, Ljubljana
darja.fiser@ff.uni-lj.si
[2] Department of Knowledge Technologies, Institute »Jožef Stefan«, Jamova cesta 39, Ljubljana
tomaz.erjavec@ijs.si
[3] Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, Ivana Lučića 3, Zagreb
nljubesi@ffzg.hr
[4] Department of General Linguistics, Faculty of Philology, University of Belgrade, Studentski trg 3, Belgrade
m.milicevic@fil.bg.ac.rs

Abstract

In this paper we make a cross-lingual comparison of non-standard features in the language of social media for Slovene, Croatian and Serbian. The goal of the analysis is twofold: (1) we try to establish the extent to which the observed phenomena are universal versus language-specific, and (2) we propose an approach for automatic scoring of (non-)standardness levels of user-generated content, which can be used as a separate annotation layer in corpora. Quantitative and qualitative analyses of the results show that the majority of the language used on Twitter is in fact fairly standard, especially in Slovene and Croatian. The prevalent characteristic of non-standard Slovene tweets is non-standard orthography, while non-standard lexis is more typical of Serbian tweets, possibly due to a younger user profile.

Primerjava nestandardnih prvin v slovenskih, hrvaških in srbskih tvitih

V prispevku predstavimo večjezično primerjavo nestandardnih jezikovnih prvin v družbenih medijih za slovenščino, hrvaščino in srbščino. Cilj analize je dvojen: (1) ugotoviti želimo, do katere mere so identificirani pojavi univerzalni za to zvrst komunikacije in katere so tiste prvine, ki so jezikovno specifične ter (2) predlagati pristop za avtomatsko ocenjevanje stopnje (ne)standardnosti spletnih uporabniških vsebin, ki ga lahko kot dodatno oznako s pridom uporabimo pri označevanju korpusov. Kvantitativna in kvalitativna analiza rezultatov kažeta, da je jezik, ki se uporablja na Twitterju, pravzaprav zelo standarden, še posebej v Sloveniji in na Hrvaškem. Prevladujoča značilnost nestandardnih slovenskih tvitov je nestandardna ortografija, medtem ko je za srbske tvite tipična nestandardna leksika, ki nakazuje na mlajši profil uporabnikov tega družbenega medija v Srbiji.

Keywords: user-generated content, non-standard language, web corpora, corpus annotation, South-Slavic languages
Ključne besede: uporabniške spletne vsebine, nestandardni jezik, spletni korpusi, označevanje korpusov, južnoslovanski jeziki

1. Introduction

User-generated content (UGC) is becoming an increasingly frequent and important source of human knowledge and people's opinions (Crystal 2011). Language use in such content, particularly social media, is characterized by special technical and social circumstances (Noblia 1998), often deviating from the norms of traditional text production. However, non-standard language use does not reflect poor communication ability (Baron 2010), but is rather a sign of the users making the best possible use of a medium to meet their communicative needs (Tagg 2012), and reflect their identity and speech style in writing (Herring 2001). Studying the language of social media is thus of great value for linguists, but it is also beneficial for improving automatic processing of UGC, which has proven quite difficult, as consistent decreases in performance on UGC have been recorded in the entire text processing chain, from part-of-speech tagging (Gimpel et al. 2011) to sentence parsing (Petrov and McDonald 2012).

Non-standard linguistic features of UGC have been analyzed both qualitatively and quantitatively (Eisenstein 2013, Hu et al. 2013), and they have been taken into account in automatic text processing applications which either strive to normalize non-standard features (Liu et al. 2011), adapt standard tools to work on non-standard data (Gimpel et al. 2011), or use pre-processing steps to tackle UGC-specific phenomena (Foster et al. 2011). However, to the best of our knowledge, the level of (non-)standardness of UGC has not been compared across languages and the extent to which the observed phenomena are universal (versus language-specific) in this type of communication has not been established. A promising avenue of research appears to be the development of an automatic measure of the level of text (non-)standardness, which, added to corpora as a separate annotation layer, could be of great help in identifying non-standard texts. In this paper we present a related experiment in which we manually annotated and analyzed the (non-)standardness level of tweets in Slovene, Croatian and Serbian, and then used the manual annotation to train a regression model which automatically predicts the level of standardness of texts in the corpus; we believe this information to hold high promise for linguistic analyses as well as all stages of text processing.

2. Corpus construction and sampling

The corpus used in the experiment we report on comprises Slovene, Croatian and Serbian tweets harvested with TweetCat (Ljubešić et al. 2014), a custom-built tool for collecting tweets written in lesser-used languages. The collection of tweets took place from 2013 to 2015, resulting in a corpus of about 61 million tokens in Slovene, 25 million tokens in Croatian and 205 million tokens in Serbian, after deduplication and filtering of foreign-language tweets and tweets without linguistically relevant content (i.e. those containing only photos, links, or emoticons). The corpus is linguistically annotated; for Slovene, tokenizing, MSD tagging and lemmatization were performed with ToTaLe (Erjavec et al. 2005), while for Croatian and Serbian we used the tagger/lemmatizer constructed by Agić et al. (2013).

It is interesting to note the differences in size between the three sub-corpora. While the amount of data for Slovene and Serbian is roughly proportional to the number of their speakers (3.5 times more for Serbian), there are twice as many speakers of Croatian as of Slovene, but they seem to be tweeting over two times less. In addition, a first examination of the collected tweets showed that the corpus is heavily skewed towards standard language, especially in Slovene and Croatian, where Twitter is frequently used for dissemination of information by news agencies and other official accounts. For this reason, for the purposes of manual annotation we prepared a more balanced sample by relying on a simple heuristic which measures the rate of out-of-vocabulary words (i.e. word forms not found in the lexica of the given languages) per tweet, with the threshold set to 20%. We included in the sample 50% of tweets below, and 50% of tweets above this threshold.

3. Manual annotation of tweets

3.1 Annotation guidelines and annotation procedure

The manual annotation of (non-)standardness was based on the findings of previous linguistic analyses of computer-mediated communication, as well as on the issues commonly reported as problematic for automatic processing of user-generated content, most of them focused on out-of-vocabulary items, syntactic deviations and UGC-specific communication conventions such as hashtags, emoticons, or multiplication of characters. Common annotation guidelines were developed to ensure consistency among annotators and across languages. (Non-)standardness was evaluated at two levels: *technical* and *linguistic*; the former takes into account non-standard capitalization (including proper names), non-standard punctuation (excluding the comma, whose misuse is not necessarily indicative of non-standard language use), and typos (excluding omissions of diacritics on *č, ć, đ, š* and *ž*, which tend to be device-motivated and can be normalized automatically), while the latter looks at (non-)standard spelling, morphology, lexis, and word order.

Each tweet was evaluated as a whole and assigned a separate standardness score at each level – either 1 (standard), 2 (moderately non-standard), or 3 (very non-standard). Two examples of annotated Slovene tweets are shown in Figure 1, each very standard on one level, but very non-standard on the other. Tweets that are (almost) completely written in a foreign language, automatically generated (e.g. news or advert lead-ins), or contain no linguistic material (but only URLs, hashtags, etc.) are not relevant for this experiment and were thus marked 0 and excluded from further processing.

T=1 / L=3: Vrjetn nobene, ker tko al tko neb ta dnar šu za malce.
T=3 / L=1: se pravi,da predvidevaš razveljavitev

Figure 1. Annotated examples for Slovene

The initial step in the annotation process consisted in annotating a small batch of tweets that were then discussed by all annotators to ensure a high level of consistency among them. About 500 tweets per language were subsequently scored and divided into development data (needed in order to train the automatic system) and testing data (for the final evaluation of the automatically assigned scores).

3.2 Analysis of identified non-standard features

To gain a better understanding of the most common non-standard phenomena in tweets, as well as to enable a cross-lingual comparison, for each of the three languages we performed a manual analysis of 50 random tweets marked 2 or 3 at the linguistic level (25+25). Each observed non-standard feature was classified into one of five categories (Orthography, Morphology, Lexis, Grammar, Speech), and assigned a label marking features such as vowel dropping, phonetic spelling, word order, short infinitive etc. If a single element exhibited more than one non-standard feature (e.g. non-standard tokenization + vowel dropping), it was classified into the category that dominated the tweet.

In Slovene, we observed a total of 186 instances of non-standard features in the analyzed sample: 26% in tweets that were assigned a score of 2, and 74% in those marked with score 3; both portions of the sample displayed features from all five categories. The most frequent feature was non-standard orthography, observed in 40% of the cases (19% in score 2, and 81% in score 3 tweets). This feature was mostly exhibited as mid- or final vowel dropping (*kupla* for *kupila*, *pozim* for *pozimi*), but there were also several cases of phonetic spelling (*kuhno* for *kuhinjo*), non-standard tokenization (*neb* for *ne bi*), and vowel multiplication (*taaako* for *tako*). With a 30% share, the second most common category was non-standard lexis (25% found in score 2, 75% in score 3 tweets), comprising colloquial expressions (*flajšter*), slang (*homič*), words from foreign languages (*merci*), and neologisms (*trol*). Non-standard grammatical features, such as missing auxiliary verbs, represented 16% of the identified features, spoken-language elements such as discourse markers and fillers 10%, and non-standard morphology (*šu, prenesu, mislu* for *šel, prenesel, mislil*) 4%.

In Croatian and Serbian tweet samples of the same size yielded substantially fewer instances of non-standard features: 144 in Croatian and 111 in Serbian; the reason behind such a difference appears to lie in the much less standard orthography of Slovene tweets, in many cases found in almost every word in a tweet. Also, while ¾ of the identified non-standard features in the Slovene sample came from score 3 tweets, such features were more evenly distributed between score 2 and score 3 tweets in Croatian and Serbian (⅔ belonged to score 3), suggesting fewer differences between moderately and very non-standard tweets in these two languages, which might make them harder to distinguish automatically.

Another discrepancy in the cross-lingual comparison concerns the most frequent non-standard category in Croatian and Serbian, which is distinctly lexical, representing 48% of all identified non-standard features for Croatian, and as much as 57% for Serbian. The non-standard forms

are predominantly colloquial (Cro: *klopa,* Ser: *smarati*) and slang expressions (Cro: *cajka,* Ser: *pičvajz*), words from foreign languages (Cro: *hangover*, Ser: *single*), and abbreviations (Cro: *nmg* for *ne mogu*). Non-standard orthography, observed in 33% of the cases in Croatian and 22% in Serbian, mostly had the form of vowel and consonant dropping in Croatian (*onak, mrš*), while in Serbian phonetic spelling of foreign words (*rilejšnšip, vac ap*) and the use of foreign spelling in Serbian words (*shkolitza-školica, yedwa-jedva*) were popular instead.

With the exception of some examples of the Ikavian variety (*pisma, tribati, uvik*), non-standard morphology is very rare in the Croatian sample (7%), and it is not found at all in Serbian, where non-standard grammatical features (13%), such as omissions of the auxiliary verb and other function words, are more typical. In Croatian, the most distinctive non-standard grammatical feature (6%) is the short infinitive. Spoken-language elements (7% in Croatian, 8% in Serbian) are very similar to Slovene (Cro: *njomnjom,* Ser: *alooo*).

4. Automatic prediction of standardness level

For the automatic prediction of the level of standardness we used the manually annotated tweets to build a regression model for each language (Slovene, Croatian and Serbian) and each dimension of standardness (technical and linguistic). We used the support-vector machine regressor with an RBF kernel, as implemented in the scikit-learn toolkit (Pedregosa et al. 2011) thereby enabling non-linear regression modelling, which improved our results significantly. We represented the content of each tweet through 29 independent variables. Most were string-based (punctuation, vowel-consonant ratio, the ratio of alphabet characters, etc.), and some token-based (e.g. the ratio of short words). A few of the variables were lexicon-based, i.e. they relied on an external information source such as a lexicon of standard language, which enabled us to determine the out-of-vocabulary ratio of all words, only short words, etc.

The results of automatic prediction of standardness level for the three sub-corpora are given in Table 1. They confirm our early intuition that Twitter data are actually quite standard, with 67-73% of the corpus classified as score 1. Slovene and Croatian tweets are particularly standard, in all likelihood because in these languages Twitter is predominantly used by official accounts for information dissemination. At the other end of the spectrum, Slovene and Croatian also have a larger share of very non-standard tweets than Serbian, consistent with the results of the manual analysis, and confirming that non-standard orthography prevails in Slovene (and to a lesser degree Croatian), whereas non-standard lexis is characteristic of Serbian, most likely reflecting the much younger profile of Serbian Twitter users.

Language	Score 1	Score 2	Score 3
Slovene	70%	23%	7%

Croatian	73%	21%	6%
Serbian	67%	30%	3%

Table 1. Distribution of standardness by language

We evaluated the results using mean absolute error, which showed that the automatic estimate of the linguistic standardness was on average 0.41 points incorrect with respect to manual annotation for Slovene, 0.44 for Serbian and 0.46 for Croatian. The best score was obtained on Slovene data due to the lexicon (Sloleks[1]) being significantly larger than those for Croatian (Apertium[2]) and Serbian (Wikipedia and news-corpora based lexicon). The results for the technical dimension were even better, with error rates ranging from 0.37 for Serbian to 0.39 for Croatian, showing that the level of technical standardness is easier to predict.

5. Conclusion

In this paper we made a cross-lingual comparison of non-standard elements in Slovene, Croatian and Serbian tweets. Using a sample of tweets that were manually annotated on a three-level scale of technical and linguistic standardness, we performed a quantitative and qualitative analysis of their non-standard features, and found that the language used on Twitter is largely standard. The prevalent characteristic of non-standard Slovene tweets is non-standard orthography, while non-standard lexis is more typical of Croatian and Serbian. We also developed a method to automatically score the (non-)standardness levels of texts for use as an annotation layer in corpora, and performed an evaluation of its accuracy.

In future work we plan to conduct an in-depth linguistic study to determine whether the language used on Twitter is becoming more or less standard with time, as its popularity and the number of users grow. We also plan to explore automatic methods for standardizing the non-standard features in corpora of the three languages, and apply high quality annotation methods on the standardized word tokens in the corpora.

Acknowledgments

The work described in this paper was funded by the Slovenian Research Agency, Serbian Ministry of Education, Science and Technological Development and Croatian Ministry of Science, Education and Sports within the national basic research project “Resources, Tools and Methods for the Research of Nonstandard Internet Slovene” (J6-6842, 2014-2017), national basic research project “Standard Serbian Language: Syntactic, Semantic and Pragmatic Explorations” (178004, 2011-2015), Slovenian-Serbian bilateral project “The Construction of

[1] http://www.slovenscina.eu/sloleks
[2] https://www.apertium.org

Corpora and Lexica of Non-standard Serbian and Slovenian" (BI-RS/14-15-068) and Slovenian-Croatian bilateral project "Bilingual Lexicon Construction for Closely Related Languages from Existing Language Resources" (BI-HR/14-15-047).

References

AGIĆ, Željko, LJUBEŠIĆ, Nikola, MERKLER, Danijela, 2013: Lemmatization and morphosyntactic tagging of Croatian and Serbian. *Proceedings of BSNLP 4*. 48-57.

BARON, Naomi, 2010: *Always On: Language in an Online and Mobile World*. Oxford: Oxford University Press.

CRYSTAL, David, 2011: *Internet Linguistics: A Student Guide*. New York: Routledge.

EISENSTEIN, Jacob, 2013. What to do about bad language on the Internet. *Proceedings of HLT-NAACL 2013*. 359-369.

ERJAVEC, Tomaž, IGNAT, Camelia, POULIQUEN, Bruno, STEINBERGER, Ralf, 2005: Massive multi-lingual corpus compilation: Acquis Communautaire and ToTaLe. *Archives of Control Sciences* 15, 529-540.

FOSTER, Jennifer, CETINOGLU, Ozlem, WAGNER, Joachim, LE ROUX, Joseph, NIVRE, Joakim, HOGAN, Deirdre, VAN GENABITH, Josef, 2011: From news to comment: Resources and benchmarks for parsing the language of web 2.0. *Proceedings of IJCNLP 5*. 893-901.

HERRING, Susan C., 2001: Computer-mediated discourse. Deborah Schiffrin, Deborah Tannen, Heidi Hamilton (eds): *The Handbook of Discourse Analysis*. Oxford: Blackwell. 612-634.

GIMPEL, Kevin, SCHNEIDER, Nathan, O'CONNOR, Brendan, DAS, Dipanjan, MILLS, Daniel, EISENSTEIN, Jacob, HEILMAN, Michael, YOGATAMA, Dani, FLANIGAN, Jeffrey, SMITH, Noah A., 2011: Part-of-speech tagging for Twitter: annotation, features, and experiments. *Proceedings of ACL 49*. 42-47.

HU, Yuheng, TALAMADUPULA, Kartik, KAMBHAMPATI, Subbarao, 2013: Dude, srsly?: The surprisingly formal nature of Twitter's language. *Proceedings of ICWSM 2013*.

LIU, Fei, WENG, Fuliang, WANG, Bingqing, LIU, Yang, 2011: Insertion, deletion, or substitution? Normalizing text messages without pre-categorization nor supervision. *Proceedings of ACL 49*. 71-76.

LJUBEŠIĆ, Nikola, FIŠER, Darja, ERJAVEC, Tomaž, 2014: TweetCaT: a tool for building Twitter corpora of smaller languages. *Proceedings of LREC 9*. 2279-2283.

NOBLIA, Maria Valentina, 1998: The computer-mediated communication: A new way of understanding the language. *Proceedings of IRISS'98*. 10-12.

PEDREGOSA, Fabian, et al., 2011: Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12. 2825-2830.

PETROV, Slav, MCDONALD, Ryan, 2012: Overview of the 2012 shared task on parsing the web. *Notes of the First Workshop on SANCL 2012*.

TAGG, Caroline, 2012: *Discourse of Text Messaging*. London: Continuum.