# Birds of a feather don’t quite tweet together: An analysis of spelling variation in Slovene, Croatian and Serbian Twitterese

***Maja Miličević,*** *University of Belgrade*

***Nikola Ljubešić,*** *Jožef Stefan Institute and University of Zagreb*

***Darja Fišer,*** *University of Ljubljana and Jožef Stefan Institute*

## Abstract

In this paper, we investigate the spelling conventions on the Twitter microblogging platform. In order to gain insight into the universalities and specificities of communication on social media, we perform a comparative analysis of three closely related languages: Slovene, Croatian and Serbian. The data collection and annotation protocols were developed jointly for all three languages, allowing for maximum interoperability and comparability of results. The analysis reveals differences in the amount of deviation from the norm in the three languages, with Slovene twitterese being the most inclined to using non-standard spelling, and Serbian the least. Overall, closed word classes, especially interjections and abbreviations, are found to be more non-standard than the open classes. In terms of types of standard > non-standard transformations, character deletions are more frequent than insertions or replacements, and transformations mostly occur in word-final positions. The discrepancies between languages are largely due to the pronounced tendency of Slovene and Croatian to use spoken-like, regional and dialectal forms characterised by vowel omissions, especially at the end of words. This analysis and the resulting datasets can be used to further study the properties of non-standard Slovene, Croatian and Serbian, as well as to develop language technologies for non-standard data in these languages.

**Keywords:** netspeak, Twitter, social media corpus, spelling variation, cross-lingual comparison

# 1 INTRODUCTION

Due to its increasing popularity and impact on society, computer-mediated communication (CMC) has been attracting a lot of attention in fields ranging from linguistics and communication studies to natural language processing and data analytics. CMC is seen as an important source of knowledge and opinions (Crystal 2011), as well as a prolific source of data on lexical and structural variation. CMC occurs under special technical and social circumstances (Noblia 1998), imposing specific communicative needs and practices (Tagg 2012). As a consequence, its language often deviates from the norms of traditional text production, instantiating numerous non-standard features at all levels, from unorthodox spelling to colloquial and other out-of-vocabulary lexis, as well as atypical syntax involving, for instance, frequent ellipsis and different uses, with and without syntactic value, of Twitter-specific elements such as @ mentions and hash tags (see, for example, Kaufmann and Kalita 2010, Arhar Holdt et al. 2016).

CMC has featured prominently in recent linguistic research, and of the three languages we focus on in this paper, Slovene CMC has been researched most extensively. An analysis of shortening strategies in tweets (Goli et al. 2016) showed a very strong tendency towards shortening among users, predominantly in the form of reductions at the orthographic level. Marko (2016), a study focused on neography, looked at letter/number homophones, showing that they occur equally frequently in foreign and Slovene words, and that the same symbol can have both a graphic (*g33k - geek*) and a phonetic use (*u3nek - utrinek / shooting star*). The influence of highly interactive and instantaneous communication platforms has been shown to blur the boundary between spoken and written discourse, resulting in the frequent use of phoneticised spelling, interaction words, deixis and non-standard lexis (Zwitter Vitez 2015).

When it comes to Croatian and Serbian, most attention in this field has centred on CMC in terms of SMS (Filipan-Žignić et al. 2012, Vrsaljko and Ljubomir 2013), Facebook (Vlajković 2010, Stamenković and Vlajković 2012), and chatroom messages (Radić-Bojanić 2007). The focus of such works has mostly been on the use of non-standard lexis (especially slang and Anglicisms) and deviations from orthographic rules, such as those concerning the use of capital letters and punctuation, as well as on non-standard spellings such as the use of *w* instead of *v*, or *sh* instead of *š*. Another prominent strand of research is the influence of new media language in the contexts of both education and literacy (Filipan-Žignić et al. 2015, Filipan-Žignić and Turk Sakač 2016), with the results showing that while pupils frequently use all the elements characteristic of new media in the texts written in their spare time, this does not interfere with their school assignments. Overall, even though some quantitative data have been reported, qualitative analysis and survey questionnaires prevail in these studies.

The two studies that are most directly related to the work presented in this paper are Fišer et al. (2015) and Miličević and Ljubešić (2016). The first compares tweets published in Slovene, Croatian and Serbian. It finds that, contrary to popular belief, most of the language used in tweets is fairly standard, especially in Slovene and Croatian. Another interesting finding was that the key characteristic of non-standard Slovene tweets is non-standard orthography, while non-standard lexis is more typical of Croatian, and especially Serbian. The second study looked only at Croatian and Serbian, detecting both similarities and differences between them. While some of the discrepancies were interpreted as being due to linguistic differences between the two languages (e.g. Croatian tends to drop final vowels to a higher extent than Serbian), others appear to be better explained by looking at extra-linguistic factors, such as user age, which seems to be lower in the case of Serbian, leading to a more chat-like format of messages. Both studies shared the finding that diacritics on letters such as *č*, *ć*, *š*, *ž* and *đ* are omitted more often in Serbian than in Croatian and Slovene.

In the present paper, we focus on posts from the Twitter microblogging platform written in Slovene, Croatian and Serbian. As one of the most widely used CMC platforms, Twitter has already received a lot of attention in linguistics. The average number of tweets published per day amounts to about 500 million,[1] and the content ranges from news broadcasts and official announcements by companies and institutions, to personal thoughts and opinions the users share, making Twitter a rich and easily accessible source of data for a wide range of (socio)linguistic inquiries. An additional component influencing the structural properties of its language is that tweets are limited to only 140 characters.

The analysis we report on is based on manually normalised, lemmatised and part-of-speech tagged samples of tweets in Slovene, Croatian and Serbian, created with the goal of developing tools for automatic CMC normalisation and tagging. In the remainder of the paper we first describe the corpora the tweets were sampled from and the samples themselves, moving on to the procedure and guidelines used in the manual normalisation. We then present the results of the analysis of normalisation. Specifically, we look at the distribution of standard-to-non-standard transformations across parts of speech and lemmas, as well as at the distribution of transformation types (deletions, insertions, and replacements), and compare these phenomena across the three datasets. Since very little related previous work is available for Slovene, Croatian and Serbian, our main goals are to give an overview of the key trends, and to compare them across languages. On the one hand, we investigate the degree to which spelling

[1] http://www.internetlivestats.com/twitter-statistics/

variations in the language of social media are universal, and on the other try to identify phenomena that are language-specific. In doing so, we treat all orthography-related phenomena as relevant for spelling, including word shortening and the expression of emphasis through letter repetitions.

## 2 CORPUS CONSTRUCTION AND SAMPLING

The corpora we employ comprise 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 for all three languages took place from 2013 to 2015, resulting in corpora of about 107 million tokens in Slovene, 25 million tokens in Croatian, and 205 million tokens in Serbian, after deduplication and filtering of foreign-language tweets and those without linguistically relevant content (i.e. those containing only mentions, links, or emoticons).

The initial samples used for the analysis presented in this paper were subsets of 4,000 tweets per language, each containing at least 100 characters, that were manually normalised, tagged and lemmatised (see Erjavec et al. 2016). These datasets were created to facilitate the development of processing tools for non-standard language, and for this reason they were sampled to represent tweets with different levels of technical and linguistic (non-)standardness (see Ljubešić et al. 2015). However, since the focus of this paper is on non-standard spelling variants, we only take into account the linguistically non-standard portion of the dataset, resulting in 1,983 tweets (54,688 tokens) in the original Slovene sample, 1,904 tweets (45,582 tokens) in the original Croatian sample, and 1,856 tweets (45,134 tokens) in the original Serbian sample.[2] After normalisation, the samples contain 54,955 Slovene tokens, 45,930 Croatian tokens and 45,322 Serbian tokens.

Examples of tweets containing non-standard features in Slovene, Croatian and Serbian are shown in Table 1. These features include phenomena typical of CMC in general, such as phonetic spelling of foreign words (e.g., *lajk* for *like*), omission of diacritics (e.g., *razrednicarka* for *razredničarka* – *teacher*), or shortenings (e.g., *yt* for *YouTube*), Twitter-specific phenomena like hashtags, @ name mentions and emoticons/emoji, as well as phenomena common in informal communication settings, such as the use of colloquial and dialectal non-standard forms (e.g., the Ikavian dialectal form *san* for *sam* – *am* in Croatian).

[2] A previous analysis of Croatian and Serbian (Miličević and Ljubešić 2016) was performed on tweets of all standardness levels.

**Table 1: Sample tweets in Slovene, Croatian and Serbian (Original tweet [standard word form] // English translation).**

| **Slovene** |
|---|
| Original: @user99 vrjamm [Verjamem] ja :) nm [Nam] pa rece [reče] razrednicarka [razredničarka], da je naj do 6ihne [6-ih ne] budimo, in tko [tako] npr [npr.] smo bli [bili] ze [že] enkrt [enkrat] ob 4 zjutri [zjutraj] pred Louvrom :D<br>Translation: Yes, I believe you :) Our teacher told us not to wake her up before 6, so we were in front of the Louvre at about 4 a.m. already, for example. :D |
| **Croatian** |
| Original: Haha :-p nakon sta [što] san [sam] jucer [jučer] pricala [pričala] s iris [Iris] o supernaturalu, pocela [počela] sam sanjat [sanjati] one demone s creepy crnin [crnim] ocima [očima] ..... [...] brr<br>Translation: Haha :-p after talking to Iris about Supernatural yesterday, I started having dreams about those demons with creepy black eyes… Brr |
| **Serbian** |
| Original: Bad Copy i Sasa [Saša] Kovacevic [Kovačević] su skoro istovremeno objavili spotove veceras [večeras], a Bad Copy imaju vise [više] lajkova do sad na yt #geto #kvalitet<br>Translation: Bad Copy and Saša Kovačević published their videos almost simultaneously tonight, and up to now Bad Copy got more yt likes #ghetto #quality |

## 3 NORMALISATION PROCEDURE AND GUIDELINES

The annotation process for all three languages was carried out using the web-based annotation platform Webanno (Eckart de Castilho et al. 2014). The annotation guidelines were first developed for the Slovene Twitter data within the Janes project (see Čibej et al. 2016), and then adapted for Croatian and Serbian based on the differences between the orthography and grammar manuals of the languages concerned. This resulted in a unified set of guidelines for the three languages, which is a big advantage in data-driven linguistics, as it enables direct cross-lingual comparisons.

For each language, each tweet was annotated independently by two annotators. A curation procedure followed, in which disagreements in the annotators' decisions were resolved. Tweets were annotated on five levels: token (i.e., corrections of word boundaries), sentence (sentence segmentation corrections), normalisation (i.e., standardisation of non-standard language features), lemmatisation (i.e., assignment of the canonical form to each word form in the running text, e.g., *objavili > objaviti – publish*) and morphosyntactic description (assignment of a morphosyntactic tag to each word in the running text following the MULTEXT-East v5.0 standard,[3] e.g., *demone – demons* > *Ncmsay* for *noun, common, masculine, singular, accusative, animate*). The complete annotation guidelines are available in the CLARIN repository,[4,5] and these are also summarised in the following subsections.

## 3.1 Segmentation and tokenisation

The samples were pre-tokenised and split into sentences with standard tools, and then checked manually by the annotators. Corrections at the sentence segmentation level relied on punctuation, if present, and on other symbols (e.g., name mentions designated with @, emoticons/emoji, and hashtags), in cases when they occupied a position where punctuation would normally be found. As for tokenisation, guidelines were provided for cases known to be problematic: hyphenated inflectional endings for abbreviations (e.g., *BMWu* for *BMW-u – at BMW* [locative]), cases where a vowel omission is marked by an apostrophe (e.g., in Serbian *pos'o* for *posao – job*), and abbreviations ending with a dot (e.g., *dr.* for *drugi – other*), which often lead to incorrect automatic splitting of a single token into two or three separate ones. An opposite case was that of word combinations containing hyphens, which are sometimes not separated into multiple tokens when they should be (e.g., in Slovene *Nemčija-Grčija* for *Nemčija – Grčija*).

## 3.2 Linguistic normalisation

In this paper we are most interested in the level of linguistic normalisation. In our case, the main goal of manual normalisation was to provide training data for building tools for automatic normalisation of CMC data. However, normalisation is also important for the end users of CMC corpora, as it enables them to perform queries based on standard forms, much along the lines of dialectal or diachronic data.

Normalisation was restricted to the word level, while word order, syntax, punctuation, ellipses, usernames, hashtags, emoticons/emoji and lexical choice (e.g., colloquial *komp* for *kompjuter – computer*) were not normalised. Normalisation

[3] http://nl.ijs.si/ME/V5/msd/html/

[4] Janes-smernice-v1.0.pdf at: http://hdl.handle.net/11356/1084

[5] ReLDI-NormTag-Guidelines.pdf at: http://hdl.handle.net/11356/1121

included the standardisation of non-standard spelling variants (e.g., in Slovene *jst* > *jaz – I*), as well as spelling and typing errors (e.g., in Croatian *popodme* > *popodne – afternoon*) and diacritic restoration (e.g., in Serbian *veceras* > *večeras – tonight*). A minimal intervention approach was adopted (e.g., in Slovene the non-standard variant *pucajne – cleaning* is normalised into the canonical non-standard variant *pucanje*, not into its standard equivalent *čiščenje*). In other words, we focused on non-standard forms that can be seen as spelling deviations, and not on style, grammar, or Twitter-specific phenomena. Context was to be taken into account when resolving unclear and ambiguous cases; if an issue could not be resolved from the available context, no normalisations were made.

While in most cases each non-standard token was normalized to one standard token, on rare occasions one non-standard token had to be split into multiple standard tokens (1:n mapping, *nevem – ne vem*, *do not know* in Slovene), and vice versa (n:1 mapping, *ni jedno – nijedno*, *neither* in Croatian). The percentage of tokens with the 1:n mapping is 0.47% in Slovene, 0.7% in Croatian and 0.39% in Serbian, while the n:1 mapping is observed with 0.06% Slovene tokens, 0.14% Croatian tokens and 0.07% Serbian tokens.

The following normalisation rules were applied in all languages (with the examples below coming from all three):

- Insert missing diacritics: *noz* > *nož – knife*
- Normalise foreign letters or letter combinations: *kavizza* > *kavica – coffee*
- Normalise non-standard spellings (regardless of whether they are regional forms, phonetic adaptations, or forms containing an obvious typo): *maš* > *imaš – have*
- Normalise cases of vowel omission or merging: *al* > *ali – but*
- Normalise non-standard inflectional endings: *živin* > *živim – I live*
- Normalise cases of missing sound assimilations: *rijedkost* > *rijetkost – rarity*
- Normalise lexical words in which some letters or syllables are repeated for emphasis; the same rule was applied to foreign words: *kaakooo* > *kako – how*
- Normalise interjections in which some letters or syllables are repeated for emphasis to two repetitions; the same rule was applied to foreign interjections: *hahaha* > *haha*
- Normalise words containing numbers instead of letters: *je2* > *jedva – barely*

- Separate/merge words non-standardly written together/apart: *nebo* > *ne bo – will not*
- Add a hyphen before inflectional endings attached to abbreviations: *DS* > *DS-u – to DS*
- Add a dot to abbreviations missing one: *min* > *min. – minute*

Specific rules were applied to only one or two of the languages, due to linguistic differences, available reference resources or the need for upstream processing:

- Slovene: Do not normalise common deviations from prescriptive rules, such as incorrect preposition choice between *z/s – with*, or incorrect modal verb choice between *moči/morati – can/must*
- Croatian and Serbian: Spell out non-standard shortenings for words other than proper nouns: *msm* > *mislim* (*I think*) (in Slovene, this was not performed)
- Croatian and Serbian: Change *bi* (*would*) into standard inflectional forms *bih/bismo/biste* for the 1st person singular, 1st person plural and 2nd person plural respectively
- Slovene and Croatian: Normalise short infinitives into long infinitives (with the exception of future tense forms in Croatian): *vjerovat* > *vjerovati* (*believe*)
- Croatian: Normalise synthetic future forms into non-synthetic future forms: *biće* > *bit će* (*will be*)
- Croatian: Normalise long infinitives into short infinitives within future tense forms: *potpisivati ću* > *potpisivat ću* (*I will sign*)
- Croatian: Normalise dialectal interrogative pronoun forms *kaj* and *ća* to the standard form *što* (in Slovene, this was not performed)

Note that we distinguish between abbreviations, which tend to have a standard form (e.g. *min.* for *minute*), and shortenings, which are idiosyncratic. In the normalisation process, abbreviations were not expanded to their full form in either of the languages, while shortenings were kept in Slovene, and expanded in Croatian and Serbian. This is one of the very few differences in the guidelines, introduced due to the different needs related to the future use of the datasets in various different projects. In addition, abbreviations were assigned a dedicated PoS tag (see Section 4.2.1), while tags assigned to shortenings depended on what PoS classes they were normalised to (e.g. *msm* stands for *mislim – I think*, and was tagged as a verb).

# 4 DATA ANALYSIS

In this section we present the results of the analyses conducted on the normalised Slovene, Croatian, and Serbian Twitter datasets. Given that our normalisation guidelines were largely based on descriptive categories that are difficult to identify automatically (e.g., phonetic transcription or incorrect spelling), the analyses had to be adjusted to look at more readily identifiable criteria. We therefore decided to focus on transformations, i.e. character-level modifications that took place in non-standard language use compared to the standard. Note that this is the opposite from the normalisation process described in Section 3, where standard language forms were assigned to non-standard ones. For instance, in Section 3 we gave an example of the Croatian Ikavian verb form *živin*, which was normalised to the standard *živim* (*I live*); in the analyses presented in the remainder of the paper we treat this as a transformation of the standard *živim* into non-standard *živin* through character replacement.

We take into account the following: (1) original tokens, comparing them to (2) normalised tokens;[6] (3) morphosyntactic descriptions assigned to normalised tokens; and (4) lemmas assigned to normalised tokens. We study the frequency distribution of transformations by part of speech, and single out the most frequently transformed lemmas and surface forms. In addition, when looking at surface forms of normalised and original tokens, we classify the differences in terms of Levenshtein transformation types (deletions, insertions, replacements),[7] and we also look at the position of specific transformations within words.

Where appropriate, we use the log-likelihood (LL) statistical test to compare the frequencies of transformations between the three corpora. It has been argued that the LL test, similar to the chi-square test, is inappropriate as an inferential test for comparing corpus frequencies, given that word choice in corpora is not random, and words are not independent of one another (see Kilgarriff 1996). However, LL can be very useful as a measure for ranking differences between corpora, e.g. for finding words and/or tags that are distinctive of a corpus (Granger and Rayson 1998, Rayson 2002); we thus use the LL to identify those part-of-speech classes and transformation types on which non-standard Slovene, Croatian, and Serbian differ most, or look most alike.[8] To calculate the LL values, we use the pre-prepared Excel sheet created by Paul Rayson.[9]

[6] One original token could be normalised to up to four tokens, and multiple original tokens could be merged into a single normalised token (see Section 3.2).

[7] We do not include the transposition transformation from the Damerau-Levenshtein distance, as it has no linguistic grounding, but rather resolves non-intentional misspellings.

[8] Due to the shortness of individual tweets, alternatives such as the Mann-Whitney test, which takes individual texts rather than whole corpora as the unit of analysis, making sure that at least texts are independent of each other (Lijffijt et al. 2016), are not applicable in our case.

[9] http://ucrel.lancs.ac.uk/people/paul/SigEff.xlsx

Lastly, we should mention that in this study we do not control for sociolinguistic variables such as user age, education and location, or tweet topic; this is an additional reason for using the statistical tests for describing our samples rather than for drawing inferences. More specifically, while we are aware of the likely influence of at least some extra-linguistic variables, our initial goal was to provide a general overview of non-standard spelling in Slovene, Croatian and Serbian Twitter data. We leave a closer inspection of the contributions made by specific additional variables for future work.

## 4.1 Overall transformation frequency

The overall percentage of transformed tokens equals 17.39% (9,555 tokens) in Slovene, 13% (5,969 tokens) in Croatian, and 10.32% (4,679 tokens) in Serbian. However, many transformations are merely diacritic omissions (*č, ć, š, ž, đ > c, c, s, z, dj*), present for technical rather than linguistic reasons (possibly because typing on smartphones and international computer keyboards is faster without diacritics). After these are filtered out from the sample, we are left with 15.56% (8,552) transformed tokens in Slovene, 10.08% (4,628) transformed tokens in Croatian, and 3.96% (1,793) transformed tokens in Serbian. In line with the findings of previous works by Fišer et al. (2015) and Miličević and Ljubešić (2016), these numbers show that diacritics are most often omitted in Serbian, while Croatian and Slovene have a greater tendency towards non-standard forms beyond diacritic omission.[10]

## 4.2 Analysis by part of speech

The first analysis we focus on is based on the part-of-speech information assigned to each token in the normalised sample. We first compare the distributions of transformations by part of speech (i.e. among all transformations, how many belong to each PoS class) in Slovene, Croatian, and Serbian. We also look at the percentage of forms that have been transformed for each part of speech (i.e. out of all words that belong to a given PoS class, how many have undergone transformation) in each language. Both analyses are limited to the tokens that have undergone transformations other than diacritic omissions.

[10] The cross-lingual difference in the amount of diacritic omissions is most likely to be due to different rates of use of international keyboards on computers and the (non)availability of localized keyboards on smartphones. The reasons are unlikely to have a linguistic nature, so we do not look into this issue further, and focus on transformations that go beyond diacritic omission.

### *4.2.1 Distribution of transformations by part of speech*

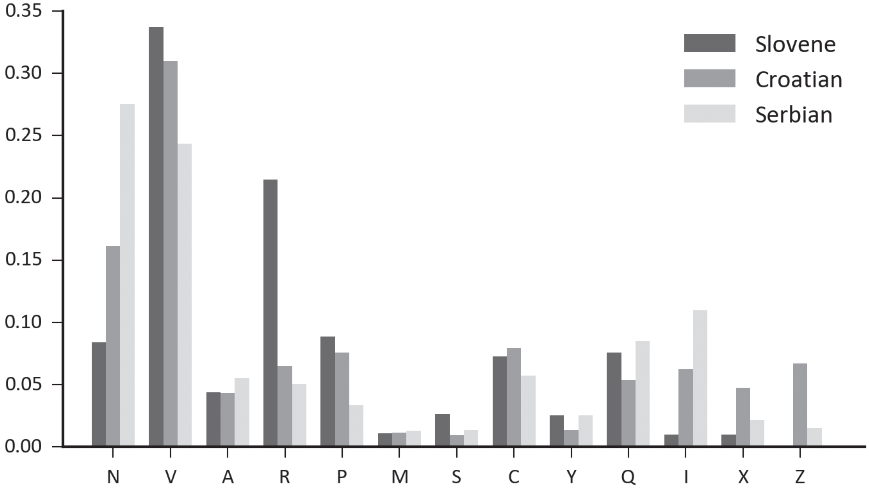

**Figure 1: Distribution of transformed forms by part of speech in the Slovene, Croatian, and Serbian Twitter datasets.**[11]

The relative frequencies of transformations by PoS are shown in Figure 1. It can be seen that despite the close relatedness of the three languages, some interesting differences emerge: while most transformations concern verbs in Slovene and Croatian, Serbian shows a more marked tendency towards noun transformation, with verbs coming second. Nouns occupy the second position in Croatian, but in Slovene they are preceded by adverbs (by a large margin) and pronouns (to a much lesser extent). It is also interesting to note that the rates of transformation in pronouns and prepositions are higher in Slovene than in the other two languages. Croatian takes the lead in the number of transformations of residuals, punctuation and conjunctions, whereas this is the case for adjectives, interjections and particles for Serbian.

The trends in Figure 1 are confirmed by log-likelihood values, which show that the difference between the three languages is most pronounced for adverbs (LL=649.66), with interjections coming second (LL=475.09), and nouns third (LL=412.03). On the opposite end of the spectrum, Slovene, Croatian and Serbian pattern together on numerals (LL=0.43), adjectives (LL=4.33), and conjunctions (LL=9.03). LL values for all parts of speech, as well as the raw frequencies they are based on, are reported in the Appendix (Table A1).

As will be shown in Section 4.3, verbal transformations in all three languages mostly belong to the auxiliary/copula *biti* (*be*), especially its 1st person singular form *sem*

[11] The tag values are as follows: N – noun, V – verb, A – adjective, R – adverb, P – pronoun, M – numeral, S – preposition, C – conjunction, Y – abbreviation, Q – particle, I – interjection, X – residual, Z – punctuation.

(often rendered as *sm*) and 3rd person singular past participle *bilo* (shortened to *blo*) in Slovene, and its 1st person singular preterite form *bih* (frequently realised as *bi*) in Croatian and Serbian. In addition, Slovene and Croatian are characterised by frequent transformations of other verbs through the shortening of the infinitive, e.g., *gledat* for *gledati – watch*, which is highly atypical of Serbian. Slovene adverbs are mostly shortened (e.g., *tako – so* frequently shortened to *tko*), but other kinds of transformations occur too. An interesting case is *zdaj – now*, which is transformed in three different ways in the dataset: *zdej*, *zdj* and *zj*. The transformations of interjections are mostly due to repeated vowels or syllables (e.g., *hahahaha*). Here, the differences across the languages are in all probability caused by minor differences in the application of the normalisation guidelines (e.g., despite the shared instructions, *ahaha* was normalised to *haha* in Croatian and Serbian, but left as *ahaha* in Slovene).

### 4.2.2 Shares of transformed forms within parts of speech

As for the percentages of forms that have been transformed within each part-of-speech class, Figure 2 shows that, overall, closed-class parts of speech tend to undergo more transformations than the open-class ones, with some differences between languages. The log-likelihood values indicate that Slovene, Croatian and Serbian differ the most on verbs (LL=1702.49), followed by adverbs (LL=1390.43) and pronouns (LL=734.56), while the classes that differ the least are numerals (LL=20.87), particles (LL=36.69), and abbreviations (LL=47.39). More detailed information is again provided in the Appendix (Table A2).

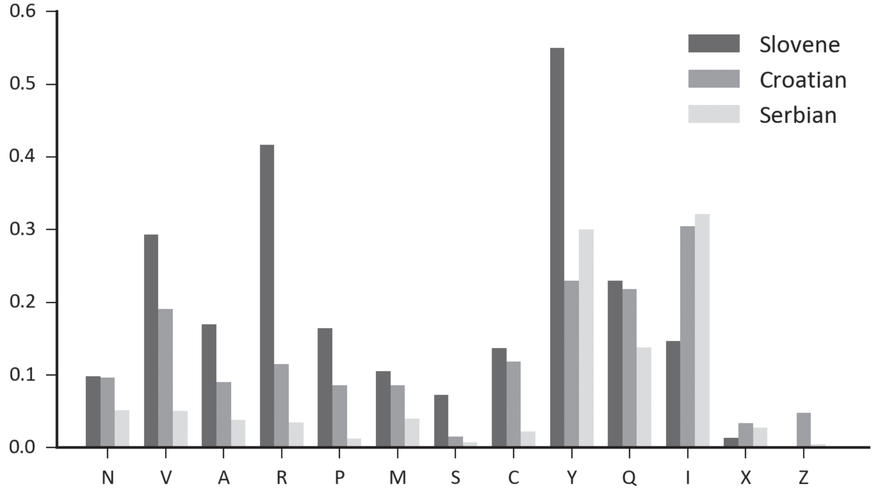

**Figure 2: Shares of transformed forms within part-of-speech classes in the Slovene, Croatian, and Serbian Twitter datasets.**

The highest percentage of transformed tokens in Slovene is found among abbreviations (mostly due to omissions of the final full stop, as in *slo*, used instead of *slo.* for *slovenski – Slovene*). In Croatian and Serbian it is the interjections that take the lead (mostly due to the aforementioned vowel or syllable repetitions, as in *hahahahaha*), followed by abbreviations (for the same reason as in Slovene), and particles (e.g., *neka – let it* is shortened to *nek*, and *je li – is it*, often merged and shortened to *jel*). Particles are transformed more in Croatian than in Serbian due to the more pronounced tendency of Croatian to omit final vowels in informal communication settings (cf. Sections 4.4 and 4.5). Conjunctions are another interesting case, as they have an overall low percentage of transformed tokens, but with about five times as many transformations in Slovene and Croatian as in Serbian. Similar to particles, most instances of transformed conjunctions are shortened versions with a (mostly final) vowel omitted. Some examples are *al* (from *ali – or* in Slovene / *but* in Croatian and Serbian), *il* (Croatian and Serbian *ili – or*), *kak* (in Slovene and Croatian, from *kako – how*), *ak* (Croatian, from *ako – if*). Pronouns are also transformed more often in Slovene and Croatian than in Serbian, but here the difference between Croatian and Serbian is mostly due to the frequent non-standard *ko* in place of the standard *tko – who*, and *šta* being used instead of *što* (*what*), while in Serbian *ko* and *šta* are the standard forms. In Slovene, the most frequent form is the 1st person singular personal pronoun *jaz - I*, commonly rendered as *jst*, *js, jest,* or *jz* instead.

Among the open part-of-speech classes, most transformations were detected for adverbs in Slovene, verbs in Croatian, and verbs and nouns in Serbian, which is consistent with the tendencies outlined for the distribution of transformations by PoS in Section 4.2.1. The trend of Slovene using more non-standard forms than Croatian, and especially Serbian, persists for adverbs, verbs, and adjectives. Interestingly, even though nouns prevail in the total percentage of transformations in Serbian, a look at within-PoS distributions reveals that more nouns actually undergo transformations in Slovene and Croatian, which can be traced back to the overall higher frequency of transformations in these two languages.

Overall, lexical word classes take up most transformations in the first comparison, while functional words take the lead in the second. In other words, despite the fact that lexical words are more frequent, a lower percentage of these are transformed, and this is why they dominate in Figure 1 but not Figure 2. From a linguistic point of view, however, this conclusion should be interpreted with caution, as some of the closed classes included in our analysis (abbreviations, residuals and punctuation), are not typically treated as PoS classes in linguistic analyses. While they do constitute a traditional PoS class, interjections too are a special case, as in our samples they mostly instantiate transformations based on repetitions, which have to do with emphasis and emotion and are not phonetic in nature (and were in addition normalised slightly differently in the three languages).

Finally, the PoS-based analyses confirm the initial observation that more non-standard spelling variants are used in Slovene and Croatian than in Serbian CMC. Multiple examples of the transformed tokens indicate that this might at least in part be due to a marked tendency of Slovene and Croatian towards vowel dropping. Before looking at this issue through Levenshtein transformations, we next present the results of the lemma- and surface form-based analyses.

## 4.3 Analysis by lemma and surface form

The set of analyses presented in this section focuses on the most frequently transformed lemmas (4.3.1) and surface forms (4.3.2).

### *4.3.1 Lemma analysis*

The lemmas that underwent most transformations in each of the three datasets are shown in Table 2, where for each lemma we report the overall percentage of the transformed forms this lemma covers (% total), on which the lemma ranking is based, as well as the percentage of all forms of that lemma that were transformed (% lemma). We again disregard transformations due to diacritic omissions.

There is a high overlap among the lemmas on the lists of all three languages, with some variation in rank. The overall most frequently transformed forms come from the auxiliary verb *biti* (*be*), first-ranked in Slovene and Serbian, and second-ranked in Croatian. The full stop, ranked first in Croatian, does not make it to the Slovene list, and is ranked 17th in Serbian. Function words and interjections follow. The interrogative particle *li*, the conjunction *kao* (*as*), and the interjections *haha* and *hajde* (*let's*) are some examples of lemmas shared by Croatian and Serbian, while the conjunction *ali* (*or* in Slovene / *but* in Croatian/Serbian) appears in all three lists. Another interesting indirect match is between the Slovene and Croatian interrogative pronouns *kaj* and *što* (*what*), the former mostly appearing as *kej* or *kj*, and the latter as either *šta* (non-standard) or *kaj* (dialectal).[12]

As for the lexical words, adverbs dominate the Slovene lemma list, while verbs are equally present in all three lists. The verbs present in the Slovene and Croatian lists (other than *biti*) undergo most transformations in the infinitive form, where their final *i* is often omitted. The situation is more varied in Serbian, where the

[12] Recall from Section 3.2 that dialectal forms of the interrogative pronoun were normalised in Croatian (as an exception to the general ban on lexical intervention), but not in Slovene.

transformations of *hteti* (*want*) are mostly due to the drop of the initial *h*, as in *oću* (*hoću – I want*), while those of the slang verb *jebati* (*fuck*) are mostly caused by the high frequency of its non-standard past participle forms *jebo* and *jeb'o* (for *jebao*). Interestingly, another two forms of the same verb, functioning as interjections, also make it to the list (*jebote* and *jebiga, fuck* and *fuck it*), due to often being shortened to *jbt* and *jbg* respectively.[13] As for nouns and adjectives, none appear in any of the three lists.

**Table 2: The 20 most frequently transformed lemmas in the Slovene, Croatian, and Serbian Twitter datasets.**

| Slovene | | | Croatian | | | Serbian | | |
|---|---|---|---|---|---|---|---|---|
| Lemma | % total | % lemma | Lemma | % total | % lemma | Lemma | % total | % lemma |
| biti#V | 8.33% | 17.02% | .#Z | 6.59% | 15.16% | biti#V | 7.53% | 6.12% |
| jaz#P | 3.24% | 33.90% | biti#V | 5.56% | 12.21% | li#Q | 6.53% | 61.26% |
| tudi#Q | 3.13% | 82.21% | što#P | 3.35% | 62.50% | haha#I | 2.90% | 81.25% |
| imeti#V | 3.09% | 66.50% | haha#I | 2.87% | 77.78% | hajde#I | 2.84% | 92.73% |
| saj#C | 1.61% | 79.77% | ne#Q | 2.38% | 24.55% | hteti#V | 2.01% | 9.78% |
| potem#R | 1.49% | 73.41% | kao#C | 2.33% | 57.45% | ali#C | 1.73% | 19.38% |
| tako#R | 1.39% | 74.38% | li#Q | 2.01% | 61.18% | kao#C | 1.51% | 14.21% |
| zdaj#R | 1.34% | 76.16% | ali#C | 1.71% | 38.35% | jebati#V | 1.45% | 27.08% |
| malo#R | 1.30% | 82.22% | hajde#I | 1.19% | 93.22% | ne#Q | 1.34% | 4.86% |
| samo#Q | 1.29% | 61.45% | moći#V | 1.17% | 27.84% | jebote#I | 1.23% | 68.75% |
| lahko#R | 1.20% | 52.82% | htjeti#V | 1.10% | 12.78% | da#C | 0.84% | 1.07% |
| toliko#R | 1.09% | 91.18% | ako#C | 0.84% | 32.23% | jebiga#I | 0.84% | 83.33% |
| ne#Q | 1.06% | 11.15% | znati#V | 0.82% | 21.35% | moći#V | 0.78% | 8.19% |
| kaj#P | 1.05% | 36.29% | tko#P | 0.82% | 45.78% | min.#Y | 0.78% | 77.78% |
| kar#R | 1.04% | 70.08% | gdje#R | 0.73% | 87.18% | ja#P | 0.73% | 1.35% |
| ali#C | 1.03% | 63.77% | kako#C | 0.65% | 33.71% | u#S | 0.67% | 1.36% |
| videti#V | 0.83% | 76.34% | nešto#P | 0.63% | 34.12% | .#Z | 0.61% | 0.62% |
| misliti#V | 0.81% | 62.73% | ići#V | 0.61% | 30.43% | ?#Z | 0.61% | 3.30% |
| kot#C | 0.72% | 32.46% | ili#C | 0.58% | 21.09% | ili#C | 0.56% | 8.85% |
| danes#R | 0.70% | 61.86% | tako#R | 0.58% | 36.99% | odmah#R | 0.56% | 50.00% |

## *4.3.2 Surface form analysis*

Moving on to surface forms, the 20 most frequent pairs of standard forms and their transformations are given in Table 3, omitting once again those that

[13] Note that idiosyncratic shortenings were expanded in Croatian and Serbian but not in Slovene.

only lack diacritics. The specific transformations are given in brackets, and the percentages these forms account for in the total number of transformations are also shown.

**Table 3: The 20 most frequently transformed surface forms in the Slovene, Croatian, and Serbian Twitter datasets.**

| Slovene | | Croatian | | Serbian | |
|---|---|---|---|---|---|
| Form | % total | Form | % total | Form | % total |
| sem (sm) | 3.37% | ... (..) | 5.68% | je li (jel) | 3.99% |
| tudi (tud) | 2.29% | kao (ko) | 1.94% | li (l’) | 1.81% |
| samo (sam) | 1.93% | ali (al) | 1.71% | ali (al) | 1.56% |
| bilo (blo) | 1.68% | je li (jel) | 1.61% | hajde (aj) | 1.50% |
| potem (pol) | 1.39% | što (sta) | 1.47% | jebote (jbt) | 1.31% |
| saj (sej) | 1.30% | što (šta) | 1.40% | jebiga (jbg) | 0.87% |
| tako (tko) | 1.28% | bih (bi) | 1.10% | min. (min) | 0.87% |
| jaz (jst) | 1.21% | ... (....) | 0.96% | kao (k’o) | 0.81% |
| malo (mal) | 1.21% | ako (ak) | 0.89% | kao (ko) | 0.78% |
| kar (kr) | 1.10% | gdje (di) | 0.86% | hajde (ajde) | 0.75% |
| ali (al) | 1.07% | što (kaj) | 0.86% | bismo (bi) | 0.62% |
| jaz (js) | 1.03% | tko (ko) | 0.77% | hajde (ae) | 0.62% |
| zdaj (zdej) | 0.97% | kako (kak) | 0.72% | haha (hahaha) | 0.56% |
| tudi (tut) | 0.89% | haha (hahaha) | 0.63% | odmah (odma) | 0.50% |
| imam (mam) | 0.76% | tako (tak) | 0.61% | haha (hahah) | 0.44% |
| pri (pr) | 0.70% | hajde (ajde) | 0.58% | bih (bi) | 0.44% |
| ko (k) | 0.70% | sam (san) | 0.51% | ili (il) | 0.44% |
| kaj (kej) | 0.70% | ili (il) | 0.51% | jebao (jebo) | 0.44% |
| nekaj (neki) | 0.66% | biti (bit) | 0.49% | u stvari (ustvari) | 0.44% |
| toliko (tolk) | 0.66% | haha (hahah) | 0.40% | li (l) | 0.37% |

The conjunction *al* is the only form shared between all three lists. While Slovene – expectedly – does not have any other forms in common with the other two languages, multiple additional forms are present in both Croatian and Serbian lists – for instance *jel* (*je li – is it*), *bi* (*bih – would*), and *ko* (*kao – like*). In Slovene *js* and *jst* instead of *jaz* (*I*) are very frequent, while all other forms instantiate either vowel replacement (typically *a>e*) or vowel omission, in different positions within words. In terms of PoS classes, most of the listed forms are adverbs. Ikavian forms (e.g., *di* for *gdje – where* and *san* for *sam – am*), as well as some final vowel omissions (*kak* for *kako – how*, *tak* for *tako – like that*, *ak* for *ako – if, bit* for *biti – be*) are specific to Croatian, while abbreviations such as *min* (*min.* for minute), and shortenings such as *jbt* (*jebote – fuck*) and *jbg* (*jebiga – fuck it*) are frequent only in Serbian.

## 4.4 Analysis by transformation type

In this section we present the probability distribution of the three types of Levenshtein transformations – deletions, insertions and replacements (Levenshtein 1966) for each language, again going from the normalised forms to the forms actually found in tweets. The results are summarised in Figure 3. The left half of the figure captures all transformations, and shows that while deletions are more frequent in Slovene than in Croatian, and in particular Serbian, the exact opposite is true of replacements. Insertions are most often found in Croatian, followed by Serbian, while they are very rare in Slovene. The high replacement rate in Serbian can be explained by its already mentioned pronounced tendency towards diacritic omission. Indeed, the right half of the figure, obtained after we discarded the tokens in which the transformation(s) consisted solely in the omission of diacritics, shows partly reversed trends: deletions and insertions become more frequent in Serbian than in Croatian (with deletions still less frequent than in Slovene), while Croatian outranks Serbian in the frequency of replacements. Overall, the most frequent transformation type is character dropping, followed by replacements, while insertions are the least frequent manifestation of the non-standard language used on Twitter.

We also performed log-likelihood tests on the data relative to the distribution of transformation types (without diacritics), confirming that insertions are the type

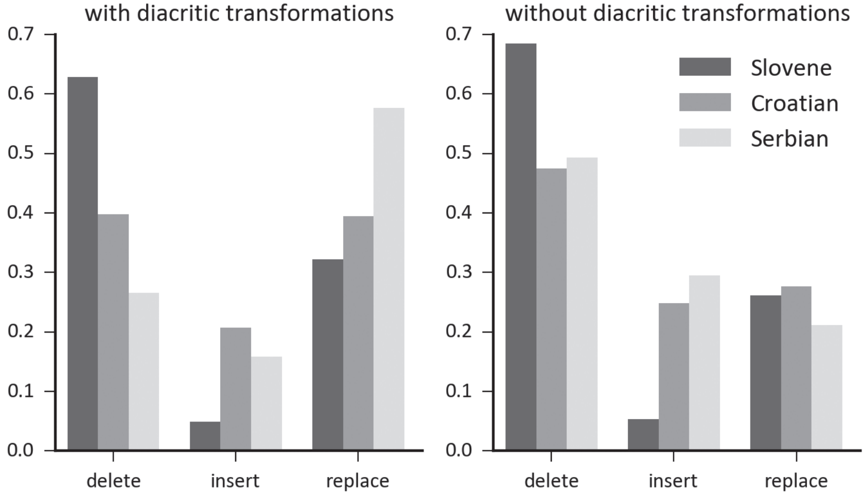

**Figure 3: Comparison of transformation distributions in the Slovene, Croatian and Serbian Twitter datasets, with (left) and without (right) diacritic transformations.**

that differs most between languages (LL=1723.79). Deletions occupy the second position (LL=400.71), while replacements reach the highest level of similarity in Slovene, Croatian and Serbian (LL=40.52). The raw frequencies that the LL values are based on are shown in Table A3 in the Appendix.

The next step in the analysis is to look at the most frequent specific transformations in each of the studied languages (again disregarding diacritic omissions). In Table 4 we show the top 10 transformations for each Levenshtein transformation type per language, together with a common example illustrating that particular transformation. The transformations are analysed at the level of single letters, so that digrams such as *lj* /lj/ are treated as two separate letters. However, special rules are added for treating 1:2 letter correspondences *đ* > *dj* and *ks* > *x* as single replacements rather than a replacement plus an insertion/deletion, as the latter approach would create a linguistically irrelevant bias in the frequency of *d* insertions and *k* deletions.[14] Moreover, an important and unavoidable consequence of the letter-by-letter approach is that many tokens contain multiple transformations defined on purely technical grounds (e.g. the definition of the Slovene transformation *potem* > *pol* is delete_t, delete_e, replace_m-l). Such transformations are not always linguistically relevant, and in some cases reflect technical decisions rather than linguistic regularities. The relative frequencies reported in Table 4 should thus be interpreted as primarily reflecting the technical side of the process, to which we add linguistic explanations in those cases where such explanations seem justified based on a qualitative analysis.

**Table 4: The 10 most frequent transformations by language and type (with examples).**

| Slovene | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Deletions | | | Insertions | | | Replacements | | |
| i | 35.04% | tudi > tud | a | 25.8% | pa > paa | l-u | 14.65% | mogel > mogu |
| e | 17.83% | sem > sm | h | 14.97% | haha > hahah | a-e | 13.32% | zdaj > zdej |
| o | 13.30% | lahko > lahk | e | 14.17% | ne > neee | j-i | 5.21% | zjutraj > zjutri |
| a | 11.23% | tako > tko | j | 9.24% | ne > nej | o-u | 4.37% | ono > uno |
| j | 3.88% | skoraj > skor | | 4.62% | odkar > od kar | a-s | 4.19% | jaz > jst |
| | 3.10% | ne bi > neb | o | 4.14% | zelo > zelooo | m-l | 4.09% | potem > pol |
| . | 2.79% | npr. > npr | s | 3.98% | imate > maste | a-o | 3.98% | danes > dons |
| t | 2.73% | potem > pol | i | 3.82% | vsak > saki | z-s | 3.95% | jaz > js |
| d | 1.77% | tudi > tut | u | 3.82% | super > suuuper | z-t | 3.88% | jaz > jst |
| u | 1.26% | tule > tle | m | 2.71% | bi > bim | i-t | 3.57% | tudi > tut |

[14] *Dj* is an alternative, non-standard spelling of the grapheme *đ*, while *x* is completely absent from the alphabets of the languages we study, which use *ks* instead (as in *maksimum* rather than *maximum*).

| Croatian | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Deletions | | | Insertions | | | Replacements | | |
| i | 24.08% | kupiti > kupit | a | 26.20% | na > naa | o-a | 10.89% | što > šta |
| | 9.51% | je li > jel | h | 15.85% | haha > haahhhaaa | e-i | 9.59% | treba > triba |
| . | 9.07% | 2013. > 2013 | o | 13.46% | to > tooo | m-n | 7.45% | sam > san |
| a | 8.49% | neka > nek | e | 10.73% | najviše > najvišeee | o-j | 3.27% | što > kaj |
| j | 8.14% | vridi > vrijedi | . | 6.40% | npr > npt. | a-e | 3.16% | pasje > pesje |
| o | 7.39% | kao > ka | i | 6.23% | ti > tii | t-a | 2.99% | što > kaj |
| e | 7.10% | čovik > čovjek | u | 3.39% | Au > Auuu | š-k | 2.93% | što > kaj |
| h | 5.84% | hajmo > ajmo | j | 2.56% | falio > falija | o-l | 1.86% | kupio > kupil |
| t | 3.90% | netko > neko | | 2.17% | A ha > Aha | ć-č | 1.64% | već > več |
| d | 2.50% | budeš > buš | s | 2.00% | sereš > seress | i-’ | 1.52% | velike > vel’ke |
| Serbian | | | | | | | | |
| Deletions | | | Insertions | | | Replacements | | |
| i | 13.62% | li > l | a | 22.51% | jao > jaao | i-’ | 7.49% | ali > al’ |
| e | 10.95% | hajde > aj | h | 12.63% | hehe > heheheh | a-’ | 5.05% | ostao > ost’o |
| a | 10.67% | kao > ko | e | 11.59% | umrla > umrela | ks-x | 3.06% | faks > fax |
| | 10.33% | je li > jel | . | 9.97% | … > ……… | i-e | 2.45% | zaspi > zaspe |
| h | 5.96% | hladan > ladan | o | 6.36% | Alo > Aloo | š-h | 2.29% | šiša > shisha |
| o | 5.90% | jebote > jbt | i | 5.03% | ima > iiima | h-’ | 2.14% | hoće > ’oće |
| d | 4.03% | hajdmo > hajmo | | 3.89% | trebaće > treba će | e-i | 2.14% | živce > živci |
| j | 3.97% | mi je > mie | ! | 3.61% | !!! > !!!! | a-e | 1.99% | nove > nova |
| u | 3.58% | ne mogu > nmg | u | 3.04% | juhu > juhuuuu | h-x | 1.83% | hehe > xexe |
| - | 3.46% | sms-a > smsa | ? | 2.85% | ?! > ??!! | r-v | 1.53% | smrde > smvde |

### *4.4.1 Analysis of deletions*

The most frequent deletions in all three languages are those of vowels and blank spaces. In Slovene, most deletions concern the vowel *i* (taking up over one third of all deletions), followed by *e*, *o*, and *a*. The vowels are omitted both word-finally (*tudi* > *tud – also*) and word-internally (*tako* > *tko – both*). They are followed by *j*, deletions of which are much less frequent, and similar in number to those of the blank space, full stop, *t*, *d*, and *u*. In Croatian, too, the most frequent cases, close to one quarter, are omissions of *i* (as in *al* for *ali – but*, and *kupit* for *kupiti – buy*). *I* is followed by the blank space (due to the merging of words such as *jel* for *je li – is it*), the dot (either within punctuation, or in abbreviations, as in *npr* for *npr. – e.g.*), *a* (e.g. in shortenings such as *ko* for *kao – like* and *nek* for *neka – let it*), and *j* (often due to the use of the Ikavian yat reflex *i* instead of the Ijekavian *(i)je*, as in *di* for *gdje – where*, or *uvik* for *uvijek – always*). In Serbian, the most frequent omissions are those of *i* (as in *jel* for *je li – is it*, *al* for *ali – but*), *e* (in shortenings like *aj* for *hajde – come on*, or *jbg* for *jebiga – fuck*), *a* (in shortened forms such as *ko* for *kao – like*, or *reko* for *rekao – said*), and the space (in merged words like *jel* for *je li – is it*, or *ustvari* for *u stvari – actually*). However, Serbian does not have a dominant deletion pattern similar to that of *i* in Slovene and Croatian.

### 4.4.2 Analysis of insertions

Insertions are mostly the result of expressive multiplication of syllables (e.g., *hahahahaha*) or vowels (e.g., in Slovene *zelooo – very*), in interjections and lexical words. The second most frequent category of insertions are strings of two words that were erroneously spelled as separate (e.g., *treba će* instead of *trebaće – will need* in Serbian). What follows are words that use foreign or idiosyncratic spelling for domestic words (e.g., in Croatian *bass* for *baš – very; right*), non-canonical abbreviation expansions (e.g., *esemes* for *sms* in Serbian), and dialectal forms that are longer than the standard ones (e.g., *falija* instead of *falio – lacked; missed* in Croatian).

### 4.4.3 Analysis of replacements

As for replacements, the most frequent case in Slovene is the *l* > *u* transformation in verbal past participles (*napisal* > *napisu – wrote*, *mogel* > *mogu – could*, *mislil* > *mislu – thought,* etc.); the second in frequency is *a* > *e* (*kaj* > *kej – what*, *zdaj* > *zdej – now*). In Serbian, replacements mostly cover the marking of character omissions with an apostrophe (as in *je l'* for *je li – is it*, or *ost'o* for *ostao – he stayed*), a phenomenon virtually non-existent in Croatian and Slovene. In Croatian, there are three frequent cases: *e-i* (due to the use of the Ikavian yat reflex, as in *triba* for *treba – needs*), *o-a* (in the substandard pronoun variant *šta* (*što – what*), and the southern dialectal endings of present participles like *falija* (*falio – lacked; missed*)), and *m-n* (transformation of the standard ending *m* in the southern dialect, as in *san* (*sam – I am*) or *van* (*vam – to you*)).

## 4.5 Analysis by position of transformation

In this section we focus on the position of transformations (deletions, insertions, and replacements) within words (with diacritic omissions once again excluded). In

Figure 4 we show the overall positional distributions of all transformations for Slovene, Croatian, and Serbian, while the following three panels (Figures 5, 6 and 7) show the results for the relative positions of deletions, insertions, and replacements.

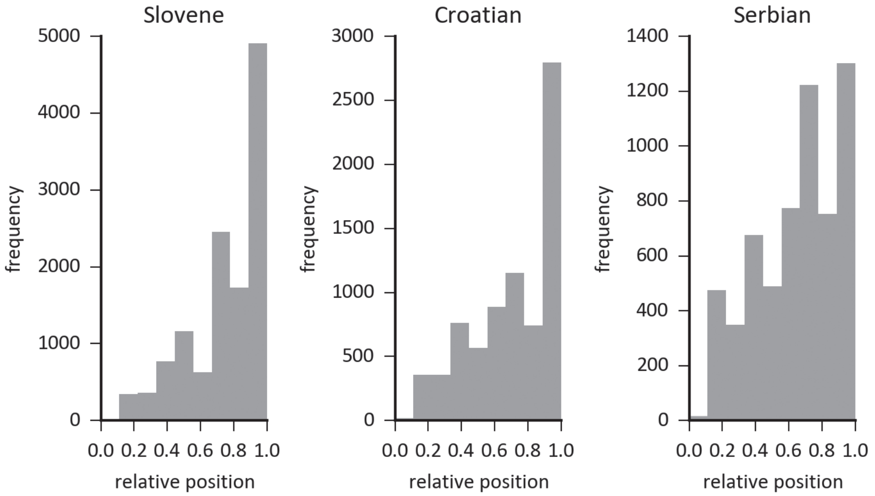

**Figure 4: Distributions of transformations by position, for Slovene, Croatian, and Serbian.**

The overall trend that emerges in the first set of histograms (Figure 4) is that transformations mostly occur at the word end, and only rarely at the beginning. The same trend is evident in all three languages, with Serbian standing out for its least marked bias towards word-final modifications in non-standard language.

Fairly similar trends are also found in all three languages for specific types of transformations. Deletions, as can be seen in Figure 5, are very biased towards the word end in Slovene, and even more so in Croatian, largely due to final vowel deletions (mostly in function words and infinitives, as outlined in Sections 4.2 and 4.3). Deletions are somewhat more evenly distributed across the word in Serbian, and not only because final vowel dropping is not as common in this language. Recall that in Serbian some of the most frequently transformed surface forms are rendered as shortenings, involving deletions at various positions within words, e.g., *jbg* < *jebiga*, *nzm* < *ne znam* (see Table 3 in Section 4.3). A tendency towards reducing words and entire phrases to shortenings is less present in Croatian, while in Slovene such phenomena were not normalised (see Section 3.2).

Insertions (Figure 6) and replacements (Figure 7) show similar distributions in all three languages, having overall an even stronger tendency towards the end of the

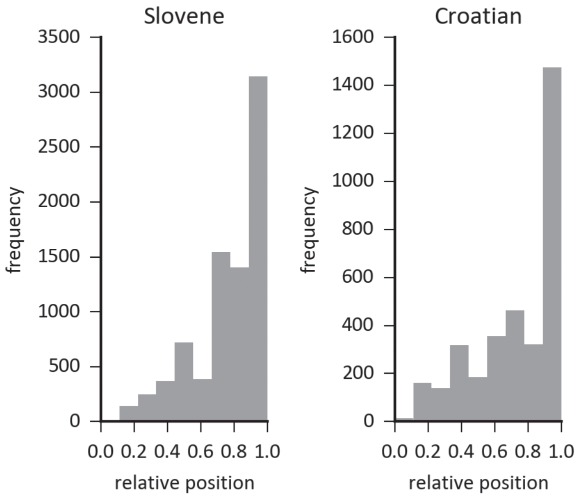

**Figure 5: Distributions of deletions by position, for Slovene, Croatian, and Serbian.**

word. For insertions, a closer inspection reveals that most cases are in fact expansions via repetitions of the final vowel. End-of-word replacements are largely accounted for by the *l* > *u* verb ending transformation in Slovene, the *o* > *a* in *što* > *šta* (*what*) and *m* > *n* in ending transformations on verbs in Croatian, and word-final vowel-to-apostrophe transformations in Serbian (e.g., *ali* > *al'* – *but*).

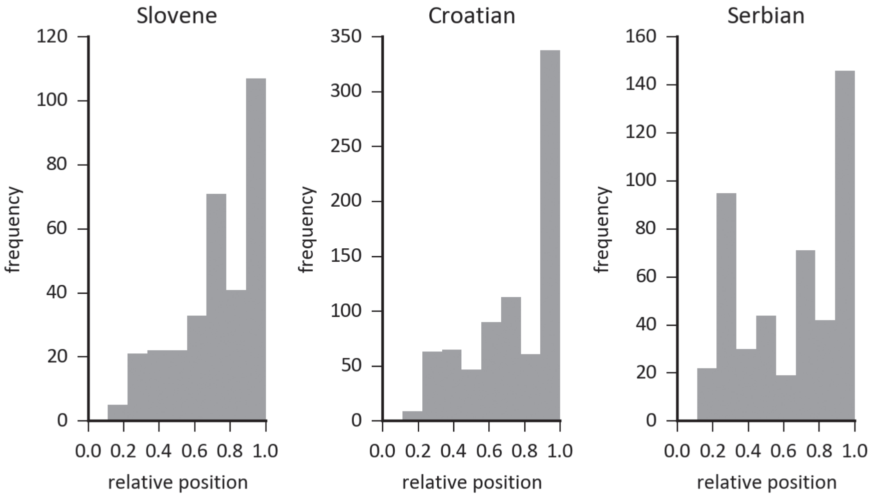

**Figure 6: Distributions of insertions by position, for Slovene, Croatian, and Serbian.**

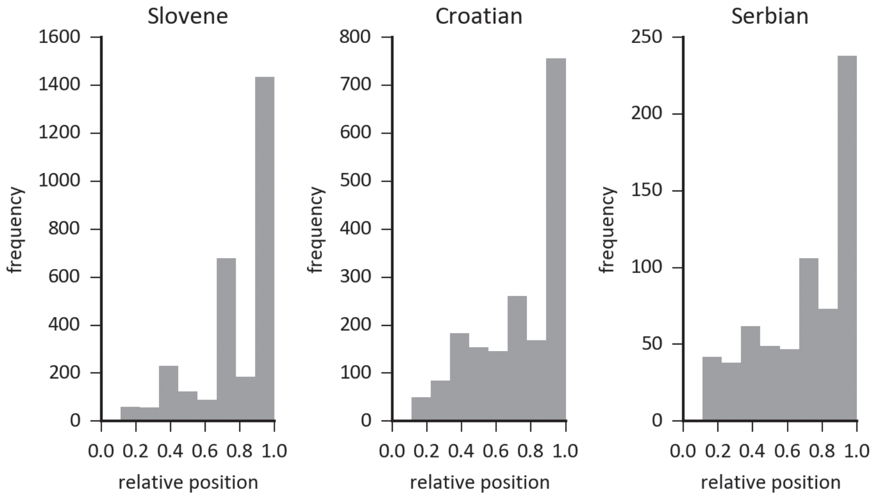

**Figure 7: Distributions of replacements by position, for Slovene, Croatian, and Serbian.**

## 5 CONCLUSION

In this paper we analysed a sample of Slovene, Croatian and Serbian tweets that were manually normalised by following unified annotation guidelines. Looking at the overall frequency of transformations, we established that the non-standard Serbian used on Twitter shows a greater tendency towards omitting diacritics, while its Slovene and Croatian equivalents are more prone to using other kinds of non-standard forms. The distribution of transformations by part of speech is such that the largest portion is occupied by open word classes (adverbs in Slovene, verbs in Croatian, and nouns in Serbian). However, looking within specific parts of speech, the most prominent transformations are those on closed classes, as confirmed by the lemma-based analysis, which revealed that the most frequently transformed lemmas belong to the classes of auxiliary verbs, interjections, and conjunctions.

By calculating the frequencies of Levenshtein transformations we observed that, leaving aside diacritic omissions, the most frequent transformations are deletions, as expected not only based on the general principle of language economy, but also due to the informal, highly interactive communication setting and frequent use of portable communication devices with suboptimal keyboards. Deletions are particularly present in Slovene, where insertions are less common than in

Croatian and Serbian. Across languages, deletions mostly consist of vowel droppings that resemble colloquial spoken language, while insertions are largely cases of expressive/emphatic vowel and syllable repetitions, especially in interjections. The picture is more varied for replacements, which also differ the most among the languages, and mostly include transformations into colloquial forms (especially in Serbian) and regional/dialectal variants (especially in Slovene and Croatian). Finally, we found that transformations are mostly word-final and very infrequently word-initial, especially in Slovene and Croatian, which is again characteristic of the colloquial spoken varieties.

While the goal of this paper was not to test specific linguistic hypotheses, we did identify some interesting spelling variation patterns. First of all, even though deletions were found to be the most typical transformation in all three languages, and vowels were consistently dropped the most in non-standard language, we also confirmed the tendency of Slovene and Croatian twitterese to omit these more often than their Serbian counterpart, especially in word-final positions. This tendency appears to be largely linguistic in nature, and mirrors the properties of the spoken varieties of the languages in question, and some historical dialectal differences (e.g. the wide presence of short infinitives in some dialects, see Stevanović 1986).

On a more sociolinguistic side, more shortenings seem to be used in non-standard Serbian than in non-standard Croatian (no data is available for Slovene, as its shortenings were not normalised). The exact reasons for this are yet to be established, given that the communicative and practical constraints are shared. One possible technical explanation is that shortenings are used in Serbian in order to gain the space that Croatian frees through single-vowel droppings. Another hypothesis is that Serbian twitterese is more “playful,” and that its users (who might belong to a different demographic than those in Croatia or Slovenia) use language in a particularly creative way. On the other hand, more regional and dialectal forms are used in Slovene and Croatian twitterese than the Serbian version, which could perhaps be traced back to differences in the official language policies of the three countries, and in how much different dialects are used and how they are viewed.

The overall picture thus seems to be one of a (socio-)linguistic non-standardness continuum going from Slovene to Serbian. What is particularly interesting is that Croatian patterns with Slovene in several respects when it comes to the non-standard language, despite the standard language of Croatian being overall much closer to Serbian, linguistically and historically. These conclusions should of course be tested in a more controlled manner in future work, and while some of the results that lead us to them might have been affected by minor discrepancies in the normalisation guidelines for the three languages, the tendencies seem robust enough to provide motivation for further studies.

In sum, given the relative scarcity of large-scale empirical data on Slovene, Croatian and Serbian CMC, the analyses reported in this work are intended to provide a valuable first insight into the nature of deviations from their norms, and to serve as a starting point for more focused studies of the linguistic phenomena at hand. In the future, our study could be complemented with an analysis of the impact of socio-demographic factors, such as user age or geographic location, on the observed transformations. Another topic that would be interesting to explore in future work would be a lexical analysis of CMC, i.e. a study of standard > non-standard lexical transformations. Such cases are not captured in our current normalisation guidelines, but previous work by Fišer et al. (2015) indicates that they are highly relevant for cross-linguistic comparisons, as Slovene was found to make less use of non-standard lexis than Croatian and Serbian.

## Acknowledgements

The research described in this paper was funded by the Slovenian Research Agency national basic research project J6-6842 *Resources, Tools and Methods for the Research of Nonstandard Internet Slovene,* the Swiss National Science Foundation (through *ReLDI – Regional Linguistic Data Initiative*, an institutional partnership among the Universities of Zurich, Belgrade and Zagreb, within the Scientific Co-operation between Eastern Europe and Switzerland programme; project No. 160501, 2015-2017), and the Ministry of Education of the Republic of Serbia (through the national research project *Standard Serbian Language: Syntactic, Semantic and Pragmatic Explorations*; project No. 178004, 2011-2016).

We are also grateful to our annotators, who made the analyses reported in this paper possible, and to the reviewers for their valuable comments.

## References

Arhar Holdt, Špela, Darja Fišer, Tomaž Erjavec and Simon Krek, 2016: Syntactic annotation of Slovene CMC: First steps. Fišer, Darja and Michael Beißwenger (eds.): *Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities*. Ljubljana: Academic Publishing Division of the Faculty of Arts. 3–6. http://nl.ijs.si/janes/wp-content/uploads/2016/09/CMC-2016_Arhar_et_al_Syntactic-Annotation-of-Slovene-CMC.pdf. (Last accessed 29 June 2017.)

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

Čibej, Jaka, Darja Fišer and Tomaž Erjavec, 2016: Normalisation, tokenisation and sentence segmentation of Slovene tweets. Andrius, Utka, Jurgita Vaičenonienė and Rita Butkienė (eds.): *Proceedings of Normalisation and Analysis of Social Media Texts (NormSoMe), LREC 2016*. 5–10. http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-NormSoMe_Proceedings.pdf. (Last accessed 29 June 2017.)

Eckart de Castilho, Richard, Chris Biemann, Iryna Gurevych and Seid Muhie Yimam, 2014: WebAnno: a flexible, web-based annotation tool for CLARIN. *Proceedings of the CLARIN Annual Conference (CAC) 2014*. Soesterberg, Netherlands. https://www.clarin.eu/sites/default/files/cac2014_submission_6_0.pdf. (Last accessed 29 June 2017.)

Erjavec, Tomaž, Jaka Čibej, Špela Arhar Holdt, Nikola Ljubešić and Darja Fišer, 2016: Gold-standard datasets for annotation of Slovene computer-mediated communication. *Proceedings of the Tenth Workshop on Recent Advances in Slavonic Natural Languages Processing (RASLAN 2016)*. Brno, Czech Republic. https://nlp.fi.muni.cz/raslan/2016/paper06-Erjavec_etal.pdf. (Last accessed 29 June 2017)

Filipan-Žignić, Blaženka, Katica Sobo and Damir Velički, 2012: SMS communication – Croatian SMS language features as compared with those in German and English speaking countries. *Revija za elementarno izobraževanje* 5. 5–22.

Filipan-Žignić, Blaženka, Vladimir Legac, Tea Pahić and Katica Sobo, 2015: New literacy of young people caused by the use of new media. *Procedia – Social and Behavioral Journal* 192. 172–179.

Filipan-Žignić, Blaženka and Marija Turk Sakač, 2016: Utjecaj novih medija na jezik mladih u pisanim radovima. *Slavistična revija* 4. 463–474.

Fišer, Darja, Tomaž Erjavec, Nikola Ljubešić and Maja Miličević, 2015: Comparing the nonstandard language of Slovene, Croatian and Serbian tweets. Smolej, Mojca (ed.): *Simpozij Obdobja 34. Slovnica in slovar - aktualni jezikovni opis (1. del)*. Ljubljana: Filozofska fakulteta. 225–231.

Goli, Teja, Eneja Osrajnik and Darja Fišer, 2016: Analiza krajšanja slovenskih sporočil na družbenem omrežju Twitter. Erjavec, Tomaž and Darja Fišer (eds.): *Proceedings of the Language Technologies and Digital Humanities Conference*. Ljubljana, Slovenia. 77–82. http://www.sdjt.si/wp/wp-content/uploads/2016/09/JTDH-2016_Goli-et-al_Analiza-krajsanja-slovenskih-sporocil.pdf. (Last accessed 29 June 2017.)

Granger, Sylviane and Paul Ryson, 1998: Automatic profiling of learner texts. Granger, Sylviane (ed.): *Learner English on Computer*. London: Longman. 119–131.

Kaufmann, Max and Jugal Kalita, 2010: Syntactic normalization of Twitter messages. *International Conference on Natural Language Processing (ICON 2010)*. Kharagpur, India. 149–158.

Kilgarriff, Adam, 1996: Which words are particularly characteristic of a text? A survey of statistical approaches. Evett, Lindsay J. and Tony G. Rose (eds.): *Proceedings of AISB Workshop on Language Engineering for Document Analysis and Recognition*, Sussex University. 33–40.

Levenshtein, Vladimir I., 1966: Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10/8. 707–710.

Lijffijt, Jefrey, Terttu Nevalainen, Tanja Säily, Panagiotis Papapetrou, Kai Puolamäki and Heikki Mannila, 2016: Significance testing of word frequencies in corpora. *Literary and Linguistic Computing* 31/2. 374–397.

Ljubešić, Nikola, Darja Fišer and Tomaž Erjavec, 2014: TweetCaT: a tool for building Twitter corpora of smaller languages. Calzolari, Nicoletta et al. (eds.): *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. 2279–2283. http://www.lrec-conf.org/proceedings/lrec2014/pdf/834_Paper.pdf. (Last accessed 29 June 2017.)

Ljubešić, Nikola, Darja Fišer, Tomaž Erjavec, Jaka Čibej, Dafne Marko, Senja Pollak and Iza Škrjanec, 2015: Predicting the level of text standardness in user-generated content. *Proceedings of Recent Advances in Natural Language Processing (RANLP 2015)*. 371–378. https://aclweb.org/anthology/R/R15/R15-1049.pdf. (Last accessed 29 June 2017.)

Marko, Dafne, 2016: The use of alphanumeric symbols in Slovene tweets. Fišer, Darja and Michael Beißwenger (eds.): *Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities*. Ljubljana: Ljubljana University Press (Faculty of Arts). 48–53. http://nl.ijs.si/janes/wp-content/uploads/2016/09/CMC-2016_Marko_Use-of-Alphanumeric-Symbols-in-Slovene-Tweets.pdf. (Last accessed 29 June 2017.)

Miličević, Maja and Nikola Ljubešić, 2016: Tviterasi, tviteraši or twitteraši? Producing and analysing a normalised dataset of Croatian and Serbian tweets. *Slovenščina 2.0* 4/2. 156–188. http://dx.doi.org/10.4312/slo2.0.2016.2.156-188. (Last accessed 29 June 2017.)

Noblia, Maria Valentina, 1998: The computer-mediated communication: A new way of understanding the language. *Proceedings of the 1st Conference on Internet Research and Information for Social Scientists (IRISS'98)*. 10–12.

Radić-Bojanić, Biljana, 2007: *neko za chat?! Diskurs elektronskih ćaskaonica na engleskom i srpskom jeziku*. Novi Sad: Filozofski fakultet.

Rayson, Paul, 2002: *Matrix: A statistical method and software tool for linguistic analysis through corpus comparison*. PhD dis., University of Lancaster.

Stamenković, Dušan and Ivana Vlajković, 2012: Jezički identitet u komunikaciji na društvenim mrežama u Srbiji. Mišić-Ilić, Biljana and Vesna Lopičić (eds.): *Jezik, književnost, komunikacija: zbornik radova. Jezička istraživanja*. Niš: Filozofski fakultet. 212–224.

Stevanović, Mihailo, 1986: *Savremeni srpskohrvatski jezik (gramatički sistemi i književnojezička norma. I Uvod, fonetika, morfologija* (5th ed.). Belgrade: Naučna knjiga.

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

Vlajković, Ivana, 2010: Uticaji engleskog jezika na srpski na planu pravopisa, leksike i gramatike u komunikaciji na Fejsbuku. *Komunikacija i kultura online* 1. 183–196.

Vrsaljko, Slavica and Tea Ljubomir, 2013: Narušavanje pravopisne norme u ranojezičnoj neformalnoj komunikaciji (na primjeru SMS poruka i internetske društvene mreže Facebook). *Magistra Iadertina* 8/1. 155–163.

Zwitter Vitez, Ana and Darja Fišer, 2015: From mouth to keyboard: the place of non-canonical written and spoken structures in lexicography. *Electronic lexicography in the 21st century: linking lexical data in the digital age: Proceedings of eLex 2015 Conference*. Ljubljana: Trojina, Institute for Applied Slovene Studies, Brighton: Lexical Computing. 250–267.

## APPENDIX

**Table A1: Raw frequencies and log-likelihood values for transformations by part of speech in the Slovene, Croatian, and Serbian Twitter datasets.**

| PoS | Slovene | Croatian | Serbian | LL |
|---|---:|---:|---:|---:|
| M | 94 | 53 | 23 | 0.43 |
| A | 376 | 201 | 99 | 4.33 |
| C | 623 | 368 | 103 | 9.03 |
| Y | 219 | 62 | 46 | 23.94 |
| Q | 647 | 248 | 153 | 28.82 |
| V | 2883 | 1435 | 437 | 44.41 |
| S | 227 | 43 | 24 | 54.12 |
| P | 760 | 351 | 60 | 70.82 |
| Z | 0 | 311 | 27 | 84.31 |
| X | 86 | 220 | 39 | 171.55 |
| N | 718 | 746 | 494 | 412.03 |
| I | 84 | 288 | 197 | 475.09 |
| R | 1835 | 302 | 91 | 649.66 |
| Total | **8552** | **4628** | **1793** | --- |

**Table A2: Raw frequencies and log-likelihood values for transformations within part-of-speech classes in the Slovene, Croatian, and Serbian Twitter datasets.**

| PoS | Number of transformations | | | Total number of tokens | | | LL |
|---|---|---|---|---|---|---|---|
| | Slovene | Croatian | Serbian | Slovene | Croatian | Serbian | |
| M | 94 | 53 | 23 | 891 | 619 | 575 | 20.87 |
| Q | 647 | 248 | 153 | 2814 | 1136 | 1110 | 36.69 |
| Y | 219 | 62 | 46 | 398 | 270 | 153 | 47.39 |
| I | 84 | 288 | 197 | 572 | 944 | 613 | 48.28 |
| X | 86 | 220 | 39 | 6415 | 6420 | 1416 | 61.31 |
| N | 718 | 746 | 494 | 7291 | 7745 | 9531 | 161.26 |
| A | 376 | 201 | 99 | 2215 | 2219 | 2611 | 221.98 |
| S | 227 | 43 | 24 | 3137 | 2739 | 3146 | 229.69 |
| Z | 0 | 311 | 27 | 7828 | 6526 | 5695 | 243.20 |
| C | 623 | 368 | 103 | 4553 | 3103 | 4508 | 444.18 |
| P | 760 | 351 | 60 | 4617 | 4065 | 4797 | 734.56 |
| R | 1835 | 302 | 91 | 4401 | 2623 | 2592 | 1390.43 |
| V | 2883 | 1435 | 437 | 9823 | 7521 | 8575 | 1702.49 |

**Table A3: Raw frequencies and log-likelihood values by transformation type in the Slovene, Croatian, and Serbian Twitter datasets.**

| Transformation type | Slovene | Croatian | Serbian | LL |
|---|---|---|---|---|
| Deletions | 7962 | 3439 | 1762 | 400.71 |
| Insertions | 628 | 1798 | 1053 | 1723.79 |
| Replacements | 3038 | 1998 | 758 | 40.52 |
| **Total** | **11628** | **7235** | **3573** | **---** |