Slovenščina 2.0, 2 (2016)

TVITERASI, TVITERAŠI OR TWITTERAŠI? PRODUCING AND ANALYSING A NORMALISED DATASET OF CROATIAN AND SERBIAN TWEETS

Maja MILIČEVIĆ
Filološka fakulteta Univerze v Beogradu

Nikola LJUBEŠIĆ
Filozofska fakulteta Univerze v Zagrebu, Institut »Jožef Stefan«

Miličević, M., Ljubešić, N. (2016): Tviterasi, tviteraši or twitteraši? Producing and analysing a normalised dataset of Croatian and Serbian tweets. Slovenščina 2.0, 4 (2): 156–188.
DOI: http://dx.doi.org/10.4312/slo2.0.2016.2.156-188.

In this paper we discuss the parallel manual normalisation of samples extracted from Croatian and Serbian Twitter corpora. We describe the datasets, outline the unified guidelines provided to annotators, and present a series of analyses of standard-to-non-standard transformations found in the Twitter data. The results show that closed part-of-speech classes are transformed more frequently than the open classes, that the most frequently transformed lemmas are auxiliary and modal verbs, interjections, particles and pronouns, that character deletions are more frequent than insertions and replacements, and that more transformations occur at the word end than in other positions. Croatian and Serbian are found to share many, but not all transformation patterns; while some of the discrepancies can be ascribed to the structural differences between the two languages, others appear to be better explained by looking at extralinguistic factors. The produced datasets and their initial analyses can be used for studying the properties of non-standard language, as well as for developing language technologies for non-standard data.

Key words: computer-mediated communication, CMC corpora, Twitter, normalisation

1 INTRODUCTION

Since the beginning of its wider use, computer-mediated communication (CMC) has been attracting a lot of attention in fields ranging from communication studies to natural language processing (NLP). On the one hand, CMC is seen as an important source of knowledge and opinions (Crystal 2011); on the other hand, its lexical and structural properties are a well-established research topic in linguistics and NLP. 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 (OOV) lexis, as well as simplified syntax (see e.g. Kaufmann, Kalita 2010).

The non-standard features of CMC are particularly important for NLP, as deviations from the norm make CMC difficult to process automatically, and tools developed for standard languages have a notoriously poor performance when applied to CMC data. This is evidenced by decreases in performance in the entire text processing chain, from tokenisation (Eisenstein 2013) and part-of-speech tagging (Gimpel et al. 2011) to sentence parsing (Petrov, McDonald 2012). The non-standard features of CMC have been analyzed both qualitatively and quantitatively (Eisenstein 2013; Hu et al. 2013), and different strategies have been proposed for dealing with non-standardness: adapting standard tools to work on non-standard data (Gimpel et al. 2011), using pre-processing steps to tackle CMC-specific phenomena (Foster et al. 2011), and normalising CMC corpora, i.e. using a dedicated annotation level in which standard forms are assigned to non-standard words (Kaufmann, Kalita 2010; Liu et al. 2011). Normalisation has most often been applied to English, with some work also available e.g. for German (Sidarenka et al. 2013), Spanish (Oliva et al. 2013), and Slovene (Ljubešić et al. 2014a; Ljubešić et al. 2016a).

In this paper we adopt the normalisation-based approach, focusing on Twitter messages (tweets) written in Croatian and Serbian. As one of the most widely used CMC platforms, Twitter has already received a lot of attention in NLP. The number of tweets published per day are counted in hundreds of millions (Benhardus, Kalita 2013), 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 source of data for NLP tasks related to text mining. To enable these tasks to be performed, automatic lower-level processing is a must, meaning in turn that the problem of non-standardness needs to be solved. In the specific case of Twitter, an additional component influencing the structural properties of its language is that messages are constrained by the length restriction of 140 characters. Given the recent availability of basic language tools for standard Croatian and Serbian, a normalisation-based approach was deemed more cost-efficient than an adaptation of standard language tools. Additionally, performing normalisation gives researchers easy access to deflections from standard language occurring in non-standard one.

Examples of tweets containing non-standard features in 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. *fešn* for *fashion*), abbreviations (e.g. *zg* for *Zagreb*), @ name mentions and emoticons, but also phenomena typical of Twitter like hashtags and some terms (e.g. *fave*), as well as some language-specific features, such as omission of diacritics (which occurs in both Croatian and Serbian, e.g. *kauc* for *kauč – couch*), and the use of fully language-specific dialectal and colloquial non-standard forms (e.g. the Ikavian dialectal form *isprid* for *ispred – in front of* in Croatian).

Croatian	**Serbian**
- ei [ej] karla trebam fejv [fave]. moš [možeš] ak [ako] nie [nije] bed pofotkat [pofotkati] ruke frendicama kojim [kojima] sam radila nokte jer se planiram bacit [baciti] u te vode :-P #karla_photography // Hey Karla, I need a fave. If that's ok, could you photograph the hands of your friends whose nails I did, I'm planning to embark on that career :-P #karla_photography - @arrrlo @bilicmaja ak [ako] neš [nešto] ne mogu smislit [smisliti] u životu, onda je to bome [bogme] med :). sluša opskurnu glazbu jopet [opet] -.- // @arrrlo @bilicmaja If there is one thing I absolutely hate, that's honey :). Listens to obscure music, again -.- - Kaj ima tak [tako] posebnog na tom rujanfestu? Zika [Muzika] je koma,lokacija jos [još] gora,cuga skupa ak [ako] se ne varam,al [ali] opet pol [pola] zg-a [Zagreba] ide... // What's so special about Rujanfest? The music is bad, the location even worse, the booze is expensive if I'm not mistaken, but half Zagreb attends anyway... - Danas stavin [stavim] sliku s mačkama i sad mi netko ostavio mače isprid [ispred] kuće. Neki tviteraš [twitteraš] zna di [gdje] živin [živim]. Ti ru ri ru. // Today I posted a photo with cats and now someone left a kitten in front of my house. Some Twitterer knows where I live. Tiru-riru.	- Nisi dobila brata da svaki dan kačiš njegove slike na fejz [fejsbuk] i da pišeš ŽIVOT MOJ jer je to fešn, aloe [alo e] seljanko // Hey, you didn't get a baby brother so that you can post his photos on Facebook every day and write MY LIFE because that's fashionable, you yokel - Cesto [Često] mi zalepe etiketu : Previse [Previše] izbirljiva. Jbg [Jebi ga] imam visoke kriterijume nije isto birati decka [dečka] i birati kauc [kauč]. // I am often labelled as : Too picky. Fuck it, I have high criteria, choosing a boyfriend is not the same thing as choosing a couch. - bas [baš] se radujem sto [što] cu [ću] od sutra imati veliki odmor od 10MINUTA A PEKARA JE 15MINUTA OD SKOLE [ŠKOLE] i posle zivot [život] je fer? mrs [mrš] // I am so happy that starting tomorrow the recess will last 10 minutes, with the bakery 15 minutes away from school, and that's supposed to be fair? Go to hell - Jebem te Bože.Ja cenim da sam toliki baksuz da ću poginuti tako što ću se okliznuti na puža koji kenja na trotoaru i polomiti vrat, jebo ma' [mater] // Fuck it God. I think I have such bad luck that I will die by slipping on a snail pooping on the pavement and breaking my neck, fuck it

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

With the future goal of developing tools for automatic CMC normalisation, we manually normalised a sample of 4000 tweets per language. In the remainder

of the paper we first describe the corpus the tweets were sampled from and the samples themselves, moving on to the procedure and the unified Croatian and Serbian guidelines used in the manual normalisation. We then present several initial analyses based on the normalisation outcomes; the analyses were performed starting from the normalised forms and looking towards forms found in the Twitter datasets. Specifically, we look at the distribution of standard -> non-standard transformations across parts of speech and lemmas, as well as the distribution of transformation subtypes (deletions vs. insertions vs. replacements), and we compare Croatian and Serbian. As very little related previous work exists for these languages, our main goals are to give an overview of the key trends, and to compare these trends in the two languages, facilitating the formulation of future specific linguistic hypotheses.

2 CORPUS CONSTRUCTION AND SAMPLING

The corpus we employ comprises Croatian and Serbian tweets harvested with TweetCat (Ljubešić et al. 2014b), a custom-built tool for collecting tweets written in lesser-used languages. The collection of tweets for both languages took place from 2013 to 2015, resulting in a corpus of about 25 million tokens in Croatian and 205 million tokens in Serbian, after deduplication and the filtering of foreign-language tweets and tweets without linguistically relevant content (i.e. those containing only photos, links, or emoticons).

The sample we used for the manual normalisation task contained a total of 4000 tweets per language, split into four categories with 1000 tweets each. The categories were based on automatically assigned levels of technical (T) and linguistic (L) standardness (Ljubešić et al. 2015), so that 1000 tweets belonged to each of the T1L1, T1L3, T3L1 and T3L3 combinations, with the marks being 1= standard and 3=very non-standard (for more detail about the annotation of standardness levels in Twitter corpora of Croatian, Serbian and Slovene see Fišer et al. 2015). These specific categories were included with the goal of sufficiently representing non-standard forms, given that it has been shown that

the language of tweets is mostly very standard in Serbian (67% of tweets being annotated with L1, and 30% with L2), and in particular Croatian (73% of tweets being annotated with L1, and 21% with L2), where Twitter is frequently used for dissemination of information by news agencies and other official accounts (Fišer et al. 2015). To ensure enough content was available, only tweets over 100 characters long were included in the sample.

Some tweets in the initial sample were deemed as irrelevant for the normalisation task and were excluded from further processing; these were messages that were unintelligible or automatically generated (e.g. news or advert lead-ins), as well as those that were (almost) completely written in a foreign language, and those that contained no linguistic material. After their removal, 3877 tweets (amounting to 89,215 tokens) remained in the Croatian sample, and 3750 tweets (91,877 tokens) in the Serbian one. Finally, due to non-one-to-one mappings (see section 3 for more detail), the token count changed during normalisation, so that the normalised sample comprises 89,542 tokens for Croatian, and 92,236 tokens for Serbian.

After manual normalisation, the normalised sample was automatically linguistically annotated; MSD (morphosyntactic description) tagging and lemmatisation were performed with the tagger and lemmatiser described in Ljubešić et al. (2016b). The accuracy of morphosyntactic tagging (773 different labels) is estimated at ~92% while the part-of-speech tagging (13 different labels) and lemmatisation reach ~98% accuracy.

3 NORMALISATION PROCEDURE AND GUIDELINES

The manual normalisation was performed using the web-based annotation platform Webanno, which allows users to define their own annotation levels. In our study, three levels were defined: corrections (tokenisation corrections), sentences (sentence segmentation corrections) and normalisation (linguistic normalisation). Guidelines were developed for each of the three levels, explaining both the technical (WebAnno-related) and the content-related side

of interventions. Up to four values could be entered per original token at each level.

Each tweet was normalised independently by two annotators. A curation procedure followed, in which the decisions of the different annotators were compared and cases of inter-annotator disagreement were resolved. For Croatian, the curation procedure was coordinated between the two annotators, while for Serbian the task was performed by an independent curator. The guidelines the annotators received are described in the following subsections.

3.1 General rules

The annotators were instructed to identify tweets deemed as irrelevant (e.g. due to being automatically generated, see section 2) and mark them for deletion. As for the relevant tweets, overall, a minimal intervention principle was adopted and it was decided not to make corrections that would be impossible, or extremely difficult for a machine learning algorithm to learn. Context was to be taken into account when resolving potentially problematic issues and ambiguous cases (e.g. in Croatian *ko -> kao – as, like*, in *sreću svu širimo ko zarazu – we spread happiness as if it were a contagious disease*, but *ko -> tko – who* in *Ko je ljep? – Who is beautiful?*); if an issue could not be resolved based on the context, no normalisations were to be made.

3.2 Segmentation and tokenisation

Defining tokens and sentences in CMC is less straightforward than in standard language corpora, and automatic procedures are more error-prone. For this reason, automatic tokenisation and segmentation were manually checked and corrected where needed.

Corrections at the sentence segmentation level relied on punctuation, if present, on other symbols (name mentions designated with @, emoticons/emojis, and hashtags) in case they occupied a position where punctuation would normally be found, and on the annotators' intuition if no

explicit symbols were used. Annotators were instructed to only insert a sentence boundary when they were fully confident one was needed, and to pay special attention to sentence-internal use of dots (...) and punctuation sequences such as *?!?!*, which can indicate pauses or surprise rather than being sentence boundary markers.

As for tokenisation, guidelines were provided for cases known to be problematic: hyphenated inflectional endings for abbreviations (e.g. *BMW-u – to BMW*), cases where vowel omission is marked by an apostophe (e.g. *pos'o*, from *posao – job*), and abbreviations ending with a dot (e.g. *dr.* from *drugi – other*), which often lead to incorrect automatic splitting of a single token into two or three separate ones. An opposite case that was mentioned was that of word combinations containing hyphens, which are sometimes not separated into multiple tokens when they should be.

3.3 Linguistic normalisation

The level we focus on in this paper is normalisation. The main goal of manual normalisation was to provide training data for building tools for automatic normalisation of CMC data, but normalisation in general 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.

In formulating the normalisation guidelines, we tried to strike a balance between the requirements of machine learning algorithms and those of linguistic analysis. The starting point of our work were the guidelines developed for Slovene Twitter data within the JANES project (see Čibej et al. 2016), which were adapted for Croatian and Serbian based on the authors' intuition, consultation with the annotators and other researchers, as well as orthography and grammar manuals of the languages concerned.

Normalisation was restricted to word level, and no word order or syntactic deviations from the standard were corrected. Additional kinds of corrections

that were explicitly excluded were those concerning lexical choice (e.g. colloquial words were not 'translated' into their standard equivalents; for instance, *komp* was not changed into *kompjuter – computer*), the use of punctuation, usernames and hashtags (regardless of what kind of linguistic material they contained), and ellipsis. In other words, we focused on non-standard forms that can be seen as spelling deviations, not intervening on OOV items that were not misspelt, on style, or on Twitter-specific phenomena. Finally, due to the complexity of the rules listed in orthography manuals, we decided not to intervene when it came to capitalisation, leaving everything as is, including lower case letters at sentence beginnings.

The following normalisation rules were applied:

- Insert missing diacritics: *juce -> juče* (*yesterday*), *najvece -> najveće* (*biggest*), *tviteras -> tviteraš* (Ser) */ twitteraš* (Cro) (*Twitterer*), *noz -> nož* (*knife*), *budzet -> budžet* (*budget*), *dovidjenja -> doviđenja* (*bye*), *iscasio -> iščašio* (*sprained*)

- Normalise Croatian/Serbian words making use of foreign letters or letter combinations: *shisha -> šiša* (*he/she cuts hair*), *chak -> čak* (*even*), *kavizzu -> kavicu* (*coffee*)

- Normalise non-standard spellings (regardless of whether they are regional forms, phonetic adaptations, or forms containing an obvious typo, and regardless of whether they are intended or non-intended): *isprid -> ispred* (*in front of*), *cili -> cijeli* (*whole*), *sumljiv -> sumnjiv* (*suspicious*), *moš -> možeš* (*you can*), *več -> već* (*already*), *gemeralno -> generalno* (*generally*), *oćeš -> hoćeš* (*you want*), *đevojci -> devojci* (*girl*), *zvjezda -> zvijezda* (*star*), *sjecass -> sjećaš* (*you remember*)

    - Normalise cases of vowel omission or merging: *UGRIZO -> UGRIZAO* (*bit*), *reko -> rekao* (*said*), *reka -> rekao* (*said*), *nek -> neka* (*let it*), *al -> ali* (*but*), *neg -> nego* (*but*), *pol -> pola* (*half*),

posudit -> posuditi (*borrow*), *ništ -> ništa* (*nothing*), *ko -> kao* (*as, like*)[1]

- Normalise non-standard inflectional endings: *živin -> živim* (*I live*), *iden -> idem* (*I go*), *njon -> njom* (*her*), *strgal -> strgao* (*broke*), *strunija -> strunio* (*rotted*)
- Normalise cases of missing sound assimilations: *iztetovirao -> istetovirao* (*tatooed*), *Predpostavljam -> Pretpostavljam* (*I assume*), *rijedkost -> rijetkost* (*rarity*)

- Normalise lexical words in which some letters or syllables are repeated for emphasis; the same rule was applied to foreign words: *poooozdraaaf -> pozdrav* (*hi/bye*), *Vrrrrh -> Vrh* (*peak*), *kaakooo -> kako* (*how*), *isssuusstti -> isus ti* (*Christ*), *etto -> eto* (*there you go*), *jbgggg -> jebi ga* (*fuck it*); *looool -> lol*
- Normalise interjections in which some letters or syllables are repeated for emphasis to two or three repetitions; the same rule was applied to foreign interjections: *hahaha -> haha*, *grrrr -> grrr* (*argh*), *MMMMMmm -> Mmm*; *faak -> fak* (*fuck*)
- Normalise words containing numbers instead of letters (e.g. *je2 -> jedva – barely*; no actual occurrences were found in the sample)
- Separate/merge words erroneously written together/apart: *neznaš -> ne znaš* (*you don't know*), *jel -> je li* (*is it*), *dal -> da li* (*is it*), *susereš -> se usereš* (*shit*), *jebemu -> jebem mu* (*fuck his*), *neželiš -> ne želiš* (*you don't want*), *nesmijem -> ne smijem* (*I must not*), *nebute -> ne budete* (*you aren't*)
- Spell out non-standard abbreviations and acronyms: *msm -> mislim* (*I think*), *Bgd -> Beograd* (*Belgrade*), *NG -> novogodišnjih* (*New*

[1] We treat more systematic phenomena applying to a larger number of forms as special subtypes of non-standard spellings.

Year's), *CZ -> Crvenu zvezdu* (*Red Star*), *pozz -> pozdrav* (*greetings, bye*), *aj -> ajde* (*come on*), *rt-ujete -> retvitujete* (*you retweet*), *Tw-ase -> Twitteraše* (*Twitterers*), *posl -> poslednjih* (*last*), *Jbt -> Jebo te* (*fuck*), *minh. -> minhensku* (*Munich*), *bgm -> boga mi* (*so help me God*), *Fkt -> Fakat* (*fact, really*)

- Change *bi* (*would*) into *bih/bismo/biste* for 1st person singular, 1st person plural and 2nd person plural respectively: *Postoje dve stvari od kojih bi da živim -> Postoje dve stvari od kojih bih da živim* (*There are two things I would like to live on*)
- (Croatian only) Normalise synthetic future forms into non-synthetic future forms: *biće -> bit će* (*will be*)
- (Croatian only) Normalise long infinitives into short infinitives within future tense forms: *potpisivati ću -> potpisivat ću* (*I will sign*)
- Add a hyphen before inflectional endings attached to abbrevations: *iz ldpa -> iz ldp-a* (*from LDP*), *DS -> DS-u* (*to DS*)
- Add a dot to abbreviations missing one: *min -> min.* (*minute*)

Rules applied to words from other languages differed for Croatian and Serbian. While phonetically transcribed words were normalised to their original spelling in Croatian (*fešn -> fashion, fejsbuk -> facebook, tviteraš -> twitteraš – Twitterer*), in Serbian foreign words (including personal names), whether phonetically transcribed or not, were not normalised, apart from obvious typos and/or errors, which were corrected: *dzenitals -> dženitals* (*genitals*); *bendz -> benč* (*bench*), *allready -> already, recomendation -> recommendation, shoppiiiiiing -> shopping*. The motivation for this decision came from the orthography rules regarding foreign proper names and non-adapted loan words in the two languages. Specifically, while the original spelling is always kept in Croatian (e.g. *Shakespeare*; *attachment*), phonetic transcriptions are the norm for proper names in Serbian (*Shakespeare -> Šekspir*), and also fairly

common for other words (*attachment* -> *atačment*).[2]

As can be seen from the examples, several of the above rules lead to non-one-to-one mappings between the original and normalised tokens, affecting the total token count discussed in section 2.

4 DATA ANALYSIS

In this section we present the results of a series of analyses performed on the manually normalised Croatian and Serbian Twitter datasets. In these analyses we look at (1) original tokens, (2) normalised tokens (up to four tokens per one original token), (3) morphosyntactic descriptions automatically assigned to normalised tokens, and (4) lemmata automatically assigned to normalised tokens.

As explained in section 3.3, the normalisation guidelines we used were formulated in terms of descriptive categories, some of which are difficult or impossible to identify automatically. In the analyses we thus look at the normalisation outcomes using more readily identifiable criteria: parts of speech, specific lemmas and surface forms, Levenshtein transformation types, and the position of transformations within words. While in section 3 we dealt with normalisation, i.e. the assignment of standard language forms to non-standard ones, in all analyses the focus is on the opposite direction (standard -> non-standard forms), as our the goal is to reconstruct the modifications that take place in non-standard language use compared to the standard; in this case we talk about transformations.

4.1 Analysis by part-of-speech

The analysis we dedicate most attention to is based on part-of-speech

[2] The tendency towards phonetic transcription in Serbian comes from its use of the Cyrillic script, in which transcription is compulsory. Serbian also uses the Roman script, in which the original spelling can be kept, but does not have to be (see Pešikan et al. 2010: 171). Note also that only tweets written in the Roman script were included in our corpus.

information assigned to each token in the normalised sample. We first look at part-of-speech distributions in Croatian vs. Serbian CMC, and in CMC vs. standard Croatian. In a second step, we further zoom in on CMC data and compare the distribution of transformations by part of speech in Croatian and Serbian.

The results of the comparison of part-of-speech distributions in the Twitter data are shown in Table 2. Both absolute and relative frequencies are shown; the LL column contains the values of the log likelihood statistic, which indicates the degree of significance of the difference between frequencies in Croatian and Serbian data; the +/- sign indicates over/under-use in Croatian compared to Serbian, and a log likelihood value between 3.8 and 6.5 is significant at $p<0.05$, while a value of 6.6 or more is significant at $p<0.01$ (Leech et al. 2000: 17; Mair et al. 2002).[3] We also compare the Twitter distributions to the part-of-speech distribution in a standard language dataset for Croatian – hr500k (Ljubešić et al. 2016b); given that a comparable standard dataset for Serbian was not available at the time of writing, here we only look at relative frequencies (%), without conducting statistical tests.

[3] The LL values were obtained using the online calculator by Paul Ryson, available at http://ucrel.lancs.ac.uk/llwizard.html.

	PoS distribution					
	hr500k	Croatian		Serbian		LL
		Freq	%	Freq	%	
Adjectives (A)	10.10%	5086	5.68%	5317	5.76%	-0.57
Conjunctions (C)	7.25%	6762	7.55%	9410	10.20%	-360.6
Interjections (I)	0.06%	465	0.52%	384	0.42%	10.33
Numbers (M)	2.52%	1487	1.66%	1222	1.32%	34.41
Nouns (N)	26.79%	16577	18.51%	18608	20.17%	-64.83
Pronouns (P)	7.95%	8032	8.97%	10233	11.09%	-204.64
Particles (Q)	1.65%	1934	2.16%	2300	2.49%	-21.76
Adverbs (R)	5.22%	5588	6.24%	5642	6.12%	1.13
Prepositions (S)	8.92%	5969	6.67%	6407	6.95%	-5.24
Verbs (V)	15.90%	15285	17.07%	17985	19.50%	-146.63
Residuals (X)	1.03%	10670	11.92%	3031	3.29%	4742.38
Abbreviations (Y)	0.33%	369	0.41%	215	0.23%	45.79
Punctuation (Z)	12.27%	11318	12.64%	11482	12.45%	1.33

Table 2: Comparison of part-of-speech distribution in the Croatian and Serbian Twitter datasets and the standard Croatian hr500k dataset.

The results show that the biggest difference in the distribution of parts of speech between Croatian and Serbian CMC data lies in the residuals, a part of speech that, in addition to the standard non-classifiable residuals, covers foreign words, emoticons/smileys, hashtags, @ name mentions and URLs. Looking at specific types of residuals, the biggest difference is observed for URLs, which are 5.3 times more frequent in Croatian than in Serbian (459 vs. 86 occurrences), followed by emoticons and foreign words, which are between 4.5 and 4.3 times more present in Croatian (1110 vs. 242 and 5025 vs. 1162 occurrences respectively). For hashtags the count is 3.8 times (1300 vs. 339), for name mentions 2.3 (2517 vs. 1072), and for general residuals 2 times (259

vs. 130) higher in Croatian than in Serbian. While the discrepancy in the number of foreign words is at least partly due to the difference in the normalisation rules for the two languages, given that phonetically transcribed words are often not tagged as foreign in Serbian, no straightforward explanation is available for the other categories. One possibility is that Twitter as a medium is used somewhat differently in the two languages: while its use in Croatian is prevalently that of a social network in which a lot of information exchange and discussion takes place, in Serbian it appears to be employed to a higher extent than in Croatian for publishing messages with personal content. This is, of course, a very tentative claim, whose further discussion we leave for future work, in which variables such as the users' age, education level and socio-economic status, as well as the private vs. corporate account status, need to be included.[4]

Among the remaining parts of speech, a substantial structurally motivated difference is observed on conjunctions, due mostly to *da* (*that*), whose relative frequency is twice as high in Serbian as in Croatian (see Table 4, section 4.2). *Da* is used in complex predicates in combination with the present tense in Serbian; in Croatian, verb infinitives are normally used instead of the *da* + present tense construction (Ser. *mogu da uradim* = Cro. *mogu uraditi – I can do*). As for the other PoS differences, they are mostly explained by the initial difference in the frequency of residuals.[5]

Moving on to the PoS distributions in the two CMC datasets vs. the hr500k

[4] We thank the two anonymous reviewers for undelining the relevance of these variables, of which age and account status (private vs. corporate) seem to be most promising in terms of data availability. Manual inspections of the corpus content so far indicate that more very young (secondary school age) Twitter users are found in Serbia than in Croatia, while more corporate accounts are present in the Croatian sample.

[5] To check this, we recalculated the relative frequencies and the LL values after removing residuals and interjections (another CMC-specific part of speech), obtaining the following LLs: adjectives 16.74, conjunctions 168.15, numerals 69.54, nouns 0.73, particles -2.49, pronouns -62.36, prepositions 8.97, adverbs 37.16, verbs -11.92, abbreviations 62.57, punctuation 69.32. While many of the differences remain significant, most values become smaller, indicating that no linguistic factors beyond those already mentioned are at play.

standard language dataset, this comparison reveals an expected ten times higher percentage of interjections and the already discussed residuals in CMC data. Furthermore, in CMC there are half as many adjectives as in the standard data, about one-third fewer nouns and one-fourth fewer prepositions, while verbs and pronouns are more present in CMC than in the standard data. Such findings are in line with CMC being a largely informal genre, where a high frequency of verbs compared to nouns is expected (see e.g. Biber et al. 1998: 68 for English).

Going back to the Twitter datasets, for each part of speech we also examined the percentages of forms that have been transformed; these results are given in Table 3. The overall percentage of tokens that were transformed is quite close in the two languages: 9.34% (8360) in Croatian and 8.57% (7910) in Serbian. However, after the transformations due to diacritic omissions are discarded, we are left with 6.87% (6156) transformed tokens in Croatian and 3.81% (3511) transformed tokens in Serbian, which shows that diacritics are omitted more often in Serbian, while Croatian has a greater tendency towards non-standard forms beyond diacritic omission. The frequencies of transformed tokens by PoS shown in Table 3 are limited to those tokens that have undergone transformations other than diacritic omissions. As above, the log likelihood statistic is reported alongside the frequencies.

	Transformations by PoS				
	Croatian		Serbian		LL
	Freq	%	Freq	%	
Adjectives (A)	257	5.05%	226	4.25%	3.61
Conjunctions (C)	499	7.38%	185	1.97%	272.04
Interjections (I)	170	36.56%	120	31.25%	1.75
Numbers (M)	64	4.30%	45	3.68%	0.65
Nouns (N)	1026	6.19%	1205	6.48%	-1.14
Pronouns (P)	360	4.48%	166	1.62%	127.95
Particles (Q)	285	14.74%	193	8.39%	37.35
Adverbs (R)	427	7.64%	206	3.65%	80.95
Prepositions (S)	60	1.01%	53	0.83%	1.07
Verbs (V)	1901	12.44%	773	4.30%	692.22
Residuals (X)	499	4.68%	213	7.03%	-23.33
Abbreviations (Y)	83	22.49%	73	33.95%	-6.48
Punctuation (Z)	525	4.64%	53	0.46%	453.85

Table 3: Transformed forms per part-of-speech in Croatian and Serbian Twitter dataset.

The highest percentage of transformed tokens is found among interjections (mostly due to vowel or syllable repetitions, as in *Hahahahaha*), abbreviations (mostly due to omissions of the final punctuation, as in *god* instead of *god.* for *godina – year*), and particles. The most frequently transformed particles with the corresponding absolute frequencies in Croatian and Serbian are *jel* (shortened from *je li – is it*, 82 vs. 73), *nebi* (shortened from *ne bi(h) – would not*, 16 vs. 7), *dal* (shortened form *da li – would it*, 12 vs. 4), *nek* (*neka – let it*, 10 vs. 9), *nezz* (*ne znam – don't know*, 9 vs. 0), and *nit'* (*niti – neither*, 8 vs. 1). Overall, particles are transformed almost twice as often in Croatian as in Serbian, most likely due to a more pronounced tendency of Croatian to omit final vowels (cf. sections 4.3 and 4.4).

Conjunctions are even more interesting for the comparison of the two

languages, as they have an overall low percentage of transformed tokens, but with four times as many transformations in Croatian as in Serbian, leading to a highly significant LL value. Most instances of transformed conjunctions are the shortened versions with a (mostly final) vowel omitted, such as *al* (from *ali – but*), *ko* (*kao – as, like*), *kak* (*kako – how*), *ak* (*ako – if*), *il* (*ili* – or). Roughly half of these shortened conjunctions occur in Serbian as well (*al*, *ko* and *il* among the previously mentioned ones), but less frequently. The situation is quite similar for pronouns, which are transformed less often still, but also to a significantly higher extent in Croatian. Here the difference between languages is mostly due to the non-standard *ko* often being used in Croatian instead of the standard *tko – who* (also in compunds such as *ne(t)ko – somebody*), and *šta* being used instead of *što* (*what*), where in Serbian *ko* and *šta* are the standard forms. The only two parts of speech that undergo significantly more transformations in Serbian are abbreviations and residuals, the latter possibly due to Croatian containing more URLs, hashtags and @ name mentions, which were not normalised.

Among the open part-of-speech classes most transformations happen among verbs (in particular the auxiliary/copula *biti – be*; see Table 5 in section 4.2) and adverbs, once again much more frequently in Croatian than in Serbian, as evidenced by very high LL values; one possible reason is the frequent shortening of infinitives in Croatian (e.g. *gledat* for *gledati – watch*), which is highly atypical for Serbian. Nouns come next, with a similar percentage of transformed forms in the two languages. Adjectives are placed last and are only slightly more frequently transformed in Croatian than in Serbian, with the difference not reaching significance.

Overall, the numbers suggest that, not counting diacritic omissions, more non-standard forms are used in Croatian than in Serbian CMC. Multiple examples of transformed tokens indicate that this might at least in part be due to a more marked tendency of Croatian towards final vowel dropping; before looking at this issue through Levenshtein transformations, we focus on a lemma-based analysis.

4.2 Analysis by lemma and surface form

The next set of analyses focuses on the most frequent lemmata in each of the resources, as well as their comparison to a standard-language resource. The most frequently normalised lemmas and surface forms are analysed as well.

The lists of the most frequent lemmata in the two Twitter datasets and the hr500k standard Croatian dataset are displayed in Table 4. The most obvious difference between the two languages, not traceable to the difference between CMC and standard language, is the higher frequency of the already discussed conjunction *da* in Serbian. The most obvious difference between the non-standard and standard registers is in the pronoun *ja* (*I, me*), which has more than 1% of occurrence in both CMC datasets, while it does not make it into the top 20 entries in standard Croatian. Most other lemmata are present in all three lists, with some slight differences in percentage and rank. The biggest difference in percentage can be observed on punctuation, with the full stop and comma being more frequent in standard Croatian than in non-standard Croatian and Serbian. On the other hand, the ellipsis, the exclamation mark and the question mark make it to either both or one of the lists of non-standard data, but not the standard data list. These divergences seem to point to punctuation not being underused in non-standard language, but rather being used somewhat differently, possibly due to its often expressive nature.

Croatian			Serbian			hr500k	
	Freq	%		Freq	%		%
biti#V	4545	5.08%	biti#V	4874	5.28%	biti#V	5.53%
,#Z	3224	3.60%	,#Z	4313	4.68%	,#Z	5.30%
.#Z	2455	2.74%	.#Z	3160	3.43%	.#Z	4.01%
i#C	1955	2.18%	da#C	3140	3.40%	u#S	2.62%
u#S	1545	1.73%	i#C	2230	2.42%	i#C	2.61%
da#C	1458	1.63%	ja#P	1843	2.00%	sebe#P	1.62%
sebe#P	1323	1.48%	u#S	1743	1.89%	na#S	1.40%
ja#P	1301	1.45%	sebe#P	1735	1.88%	koji#P	1.24%
na#S	1202	1.34%	na#S	1198	1.30%	da#C	1.21%
...#Z	1184	1.32%	ne#Q	1131	1.23%	za#S	1.00%
ne#Q	869	0.97%	taj#P	866	0.94%	taj#P	0.89%
"#Z	863	0.96%	a#C	731	0.79%	sa#S	0.75%
taj#P	821	0.92%	on#P	718	0.78%	"#Z	0.72%
htjeti#V	738	0.82%	koji#P	662	0.72%	htjeti#V	0.69%
za#S	689	0.77%	"#Z	658	0.71%	od#S	0.65%
sa#S	671	0.75%	za#S	657	0.71%	ne#Q	0.59%
!#Z	645	0.72%	sa#S	633	0.69%	a#C	0.56%
koji#P	586	0.65%	...#Z	614	0.67%	i#Q	0.53%
on#P	555	0.62%	što#P	604	0.65%	on#P	0.52%
a#C	554	0.62%	?#Z	560	0.61%	moći#V	0.44%

Table 4: The 20 most frequent lemmata in the Croatian and Serbian Twitter datasets and the standard hr500k Croatian dataset.

In Table 5 we show the lemmata that were most frequently transformed in each of the Twitter datasets. For each lemma we report the frequency, overall percentage of the transformed forms this lemma covers, as well as the percentage of all forms of that lemma that were transformed. We again disregard transformations due to diacritic omissions.

Croatian				Serbian			
	Freq	%	% trans.		Freq	%	% trans.
…#Z	512	8.32%	43.24%	biti#V	159	4.53%	3.26%
biti#V	355	5.77%	7.81%	li#Q	146	4.16%	45.77%
što#P	129	2.1%	33.86%	ali#C	57	1.62%	16.1%
kao#C	128	2.08%	39.38%	jebati#V	50	1.42%	39.06%
ne#Q	128	2.08%	14.73%	hteti#V	47	1.34%	9.13%
ali#C	118	1.92%	31.13%	kao#C	46	1.31%	11.98%
li#Q	115	1.87%	49.57%	ajde#I	40	1.14%	72.73%
haha#I	83	1.35%	80.58%	haha#I	39	1.11%	81.25%
htjeti#V	67	1.09%	9.08%	što#P	36	1.03%	5.96%
moći#V	63	1.02%	16.07%	twitter#N	28	0.8%	96.55%
hajde#I	56	0.91%	91.8%	ne#Q	27	0.77%	2.39%
hehe#I	53	0.86%	80.3%	on#P	26	0.74%	3.62%
tko#P	51	0.83%	33.12%	…#Z	23	0.66%	3.75%
kako#C	50	0.81%	22.42%	moći#V	22	0.63%	5.68%
znati#V	42	0.68%	14.14%	da#C	20	0.57%	0.64%
ako#C	40	0.65%	17.54%	zašto#R	19	0.54%	20.0%
da#C	38	0.62%	2.61%	Beograd#N	19	0.54%	44.19%
gdje#R	38	0.62%	50.67%	min.#Y	18	0.51%	81.82%
tweet#N	37	0.6%	74.0%	Vučić#N	18	0.51%	39.13%
tako#R	35	0.57%	22.01%	kazati#V	17	0.48%	6.75%

Table 5: The 20 most frequently transformed lemmata. The third numerical column describes the proportion of the lemma occurrences that were transformed.

Many lemmata are present in both lists, with some variation in rank. In Croatian the most frequently transformed lemma is the ellipsis punctuation (…), which occupies the 13th place in Serbian. The overall most frequently transformed forms come from the verb *biti* (*be*). In Croatian, *biti* is followed by a series of function words, while in Serbian two additional verbs make the top five as well: *jebati* (*fuck*), mostly due to the high frequency of abbreviations such as *jbg* (from *jebi ga – fuck it*), and *hteti* (*want*), mostly due to the drop of the initial *h*, as in *oću* (*hoću – I want*) or *oće* (*hoće – he/she wants*). The rest of the list mostly consists of function words and Twitter-specific nouns (*tweet* and

Twitter), as well as two proper nouns in Serbian: the name of the current prime minister Aleksandar Vučić (frequently mentioned and sometimes encoded using the initials *AV* or the form *AVučić*), and the Serbian capital Belgrade (mostly shortened to *Bg* or *Bgd*).

Finally, as for the 20 most frequently transformed surface forms, omitting those that only lack diacritics, they are given in Table 6.

Croatian			**Serbian**		
	Freq	%		Freq	%
..	410	7.26%	jel	78	3.18%
ko	141	2.50%	al	44	1.80%
al	110	1.95%	l'	36	1.47%
jel	85	1.51%	tw	32	1.31%
bi	74	1.31%	bi	31	1.27%
šta	73	1.29%	ko	29	1.18%
sta	69	1.22%	aj	28	1.14%
....	57	1.01%	jbt	22	0.90%
kak	44	0.78%	min	18	0.73%
di	41	0.73%	jbg	17	0.69%
ak	39	0.69%	k'o	16	0.65%
bit	38	0.67%	l	14	0.57%
tak	34	0.60%	hahaha	14	0.57%
hahaha	30	0.53%	al'	13	0.53%
san	29	0.51%	fb	13	0.53%
ajde	29	0.51%	reko	12	0.49%
il	25	0.44%	ae	10	0.41%
ajmo	21	0.37%	god	9	0.37%
hahah	21	0.37%	hahah	9	0.37%
uvik	18	0.32%	wtf	8	0.33%

Table 6: The 20 most frequently transformed surface forms in the Croatian and Serbian Twitter datasets.

While some forms are shared between the two lists – for instance *jel* (*je li – is it*), *al* (*ali – but*), *bi* (*bih – would*), *ko* (*kao – like,* also *tko – who* in Croatian) –

Ikavian forms (e.g. *uvik* for *uvijek – always*) and some final vowel omissions (*kak* for *kako – how*, *tak* for *tako – like that*, *ak* for *ako – if*) are specific to Croatian, while abbreviations such as *fb* (Facebook) and *tw* (Twitter), *min* (*min.* for minute) and *god* (*god.* for *godina – year*), or *jbt* (*jebo te – fuck*) and *jbg* (*jebi ga – fuck it*) are frequent only in Serbian.

4.3 Analysis by transformation type

We start the next analysis by calculating for each language the probability distribution of the three types of Levenshtein transformations – deletions, insertions and replacements (Levenshtein 1966), going from the normalised forms to the forms found in tweets.

The results are summarised in Table 7. The numbers in the first three rows capture all transformations, and show that while deletions and insertions are significantly more frequent in Croatian than in Serbian, the opposite is true for replacements. The fact that Serbian has over 10% more replacements than Croatian can be explained by its already mentioned more pronounced tendency towards diacritic omission. In fact, the numbers in the bottom rows, obtained after we discarded the tokens in which the transformations consisted solely in the omission of diacritics, show partly reversed trends: deletions become more frequent in Serbian, and replacements in Croatian. Overall, the most frequent transformation type is character dropping, followed by replacements, roughly half of which in Croatian, and four fifths in Serbian, are due to omission of diacritics.

	Transformation distribution				
	Croatian		Serbian		LL
	Freq	%	Freq	%	
Deletions	5056	40.01%	4459	34.64%	50.88
Insertions	2486	19.73%	1820	14.14%	117.69
Replacements	5065	40.18%	6592	51.22%	-170.21
Deletions (-d)	5055	50.41%	4459	62.59%	-110.29
Insertions (-d)	2417	24.10%	1436	20.16%	29.19
Replacements (-d)	2556	25.49%	1229	17.25%	131.64

Table 7: Comparison of transformation distributions in Croatian and Serbian, with and without (-d) diacritic omission.

We next analyse the most frequent specific transformations by language. In Table 8 we show the top 10 transformations per Levenshtein transformation type, separately for Croatian and Serbian.

Croatian						**Serbian**					
delete		insert		replace		delete		insert		replace	
i	23.32%	a	24.13%	š-s	19.94%	e	11.84%	d	22.69%	š-s	31.6%
.	10.11%	h	13.15%	ć-c	12.0%	a	10.65%	a	16.21%	č-c	17.13%
	8.23%	o	10.62%	č-c	11.85%	i	10.59%	h	9.07%	ć-c	16.7%
a	7.7%	i	9.81%	ž-z	8.75%		9.44%	e	7.31%	ž-z	12.11%
j	7.46%	e	9.13%	e-i	4.96%	o	6.8%	_	6.1%	đ-j	6.02%
e	7.3%	.	8.56%	o-a	4.24%	r	4.96%	.	5.66%	i-'	1.06%
o	6.9%	d	3.74%	m-n	2.92%	t	4.37%	o	4.18%	a-'	0.7%
h	4.81%	u	2.9%	a-e	1.64%	u	4.26%	i	3.96%	.-_	0.46%
t	4.02%	j	2.57%	đ-j	1.46%	n	3.92%	s	2.2%	a-.	0.33%
d	2.12%	s	2.05%	e-v	1.09%	d	3.41%	!	2.14%	š-h	0.33%

Table 8: The 10 most frequent transformations by language and type.

As expected, the most frequent deletions in both languages are those of vowels, but with some exceptions as well. In Croatian the most frequent cases are

deletions of *i* (as in *al* for *ali – but*, and *il* for *ili – or*), the dot (either within punctuation ..., or in abbreviations, as in *npr* for *npr. – e.g.*), the space (due to the merging of words such as *jel* for *je li – is it*, or *nezz* for *ne znam – I don't know*), *a* (in shortenings such as *ko* for *kao – like* and *nek* for *neka – let it*), *j* (due to the use of the ikavian yat reflex, as in *di* for *gdje – where*, or *uvik* for *uvijek – always*), and *e* (in shortenings such as *bu* for *bude – will be*, or *ajd* for *hajde – come on*). In Serbian, the most frequent deletions are those of *e* (in shortenings like *aj* for *ajde – come on*, or *jbg* for *jebi ga – fuck*), *a* (in shortened forms such as *ko* for *kao – like*, or *reko* for *rekao – said*), *i* (in *jel* for *je li – is it*, *al* for *ali – but*, or *msm* for *mislim – I think*), the space (in merged words like *jel* for *je li – is it*, or *ustvari* for *u stvari – actually*), and *o* (in shortenings like *jbt* for *jebote – fuck*, *fb* for *facebook* and *bi* for *bismo – we would*). This analysis indicates that in Croatian deletions are more frequent on high frequency words, while Serbian shows a tendency towards shortening frequently co-occurring terms or phrases.

Insertions in both languages are mostly due to interjections, and some lexical words, containing repeated syllables (e.g. *hahahahaha*), or repeated vowels (as in *vodiiiiiiii – leads*). As for replacements, while in Serbian they mostly cover the omission of diacritics and the marking of vowel 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), in Croatian there are three additional frequent cases: *e-i* (due to the use of the ikavian yat reflex, as in *vitar* for *vjetar – wind*), *o-a* (in the substandard pronoun variant *šta* (*što – what*), and the southern dialectal endings of present participles like *pogodia* (*pogodio – he hit*) and *falija* (*falio – lacked*)), and *m-n* (transformation of the standard ending *m* in the southern dialect, as in *san* (*sam – I am*) or *van* (vam – *to you*), both forms clashing with standard forms of different parts of speech (*san* meaning *dream*, and *van* meaning *out*).

4.4 Analysis by position of transformation

In the final part of the analysis we focus on the position of transformations (deletions, insertions, replacements) inside the word.

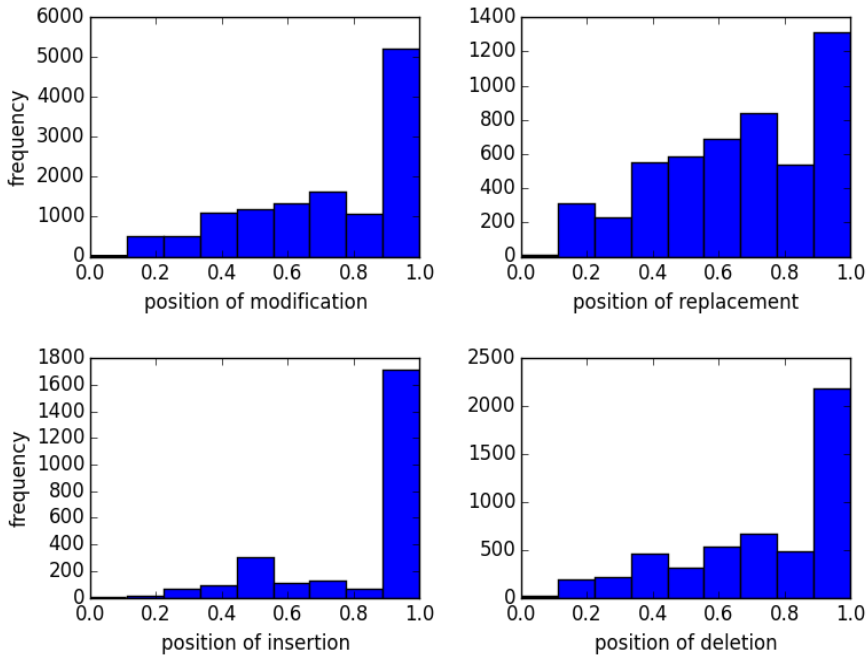

Figure 1: Transformations in Croatian by position.

Figure 1 shows the results for Croatian. The overall trend seen in the first histogram is that transformations mostly occur at the word end, and barely ever at word beginning. Replacements, typically being due to omissions of diacritics, as well as some dialectal transformations, occur inside the word as well, although still more frequently at word end. Insertions have the strongest tendency towards the end of the word; a closer inspection of all strings shows that most insertions are in fact expansions via repetitions of the final vowel. Compared to insertions, deletions are more frequently found inside the string, but there is again an emphasis on word end, largely due to final vowel deletions.

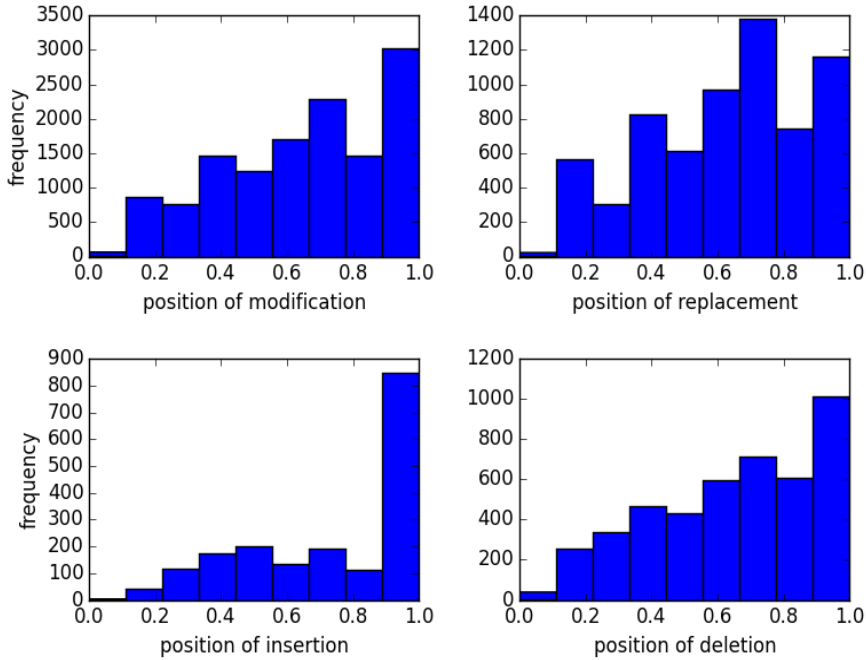

Figure 2: Transformations in Serbian by position.

The corresponding histograms for Serbian can be seen in Figure 2. These histograms show a much less pronounced trend of transformations predominantly being at the end of the string, primarily due to the more frequent omission of diacritics compared to Croatian. This is also reflected in the replacement histogram, where most transformations occur in the second half of the string, but not at its very end. Insertions again have the strongest tendency towards the end of the string, but both insertions and deletions are less biased towards the end than in Croatian.

5 CONCLUSION

In this paper we presented a sample of Croatian and Serbian tweets manually normalised by following unified annotation guidelines. The produced datasets will be highly useful both for studying the language of CMC and for developing

language technologies for CMC data, especially text normalisers that will enable standard language technologies to be used in downstream processing.

We also carried out a series of analyses on the described datasets. Inspecting the overall frequency of transformations, we concluded that Serbian shows a greater tendency towards omitting diacritics, while Croatian is more susceptible to other types of non-standard forms. The distribution of parts of speech in both languages, compared to a standard Croatian dataset, revealed a lower percentage of adjectives and nouns and a higher percentage of verbs in CMC. As for transformations of different parts of speech, most frequent transformations were those on closed part-of-speech classes. Lemma-based analyses showed the most frequently transformed lemmas to be auxiliary and modal verbs, interjections, particles and pronouns.

Focusing on Levenshtein transformations, we observed that, putting aside diacritic omissions, the most frequent transformations were deletions, the amount of insertions and replacements being similar. Deletions consisted mostly of vowel droppings, while insertions were mostly due to vowel repetitions and prolonged interjections; most replacemens were due to diacritic ommissions and regional variants. Finally, we found that transformations mostly occurred at word end, and very infrequently at word beginning, especially in Croatian. Insertions were found to have the most pronounced tendency towards the end, deletions coming second.

These initial analyses are intended to provide a starting point for studies of more specific linguistic phenomena, as well as extralinguistic factors such as user age. In future work we also plan to focus on a lexical analysis of CMC, not captured in our normalisation guidelines, but shown in previous work (Fišer et al. 2015) to be very relevant for Croatian in Serbian, as they both display a higher percentage of lexical than structural non-standard forms.

ACKNOWLEDGEMENTS

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

REFERENCES

Benhardus, J., and Kalita, J. (2013): Streaming trend detection in Twitter. *International Journal of Web Based Communities*, 9(1): 122–139.

Biber, D., Conrad, S., and R. Reppen (1998): *Corpus Linguistics. Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Crystal, D. (2011): *Internet Linguistics: A Student Guide*. New York: Routledge.

Čibej, J., Fišer, D., and Erjavec, T. (2016): Normalisation, tokenisation and sentence segmentation of Slovene tweets. *Proceedings of Normalisation and Analysis of Social Media Texts (NormSoMe) 2016, LREC 2016*: 5–10. http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-NormSoMe_Proceedings.pdf

Eisenstein, J. (2013): What to do about bad language on the Internet. *Proceedings of HLT-NAACL 2013*: 359–369. http://www.cc.gatech.edu/~jeisenst/papers/naacl2013-badlanguage.pdf

Fišer, D., Erjavec, T., Ljubešić, N., and Miličević, M. (2015): Comparing the nonstandard language of Slovene, Croatian and Serbian tweets. M.

Smolej (Ed.): *Simpozij Obdobja 34. Slovnica in slovar - aktualni jezikovni opis (1. del)*: 225–231. Ljubljana: Filozofska fakulteta.

Foster, J., Cetinoglu, O., Wagner, J., Le Roux, J., Nivre, J., Hogan, D., and van Genabith, J. (2011): From news to comment: Resources and benchmarks for parsing the language of web 2.0. *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP 2011):* 893–901. http://www.aclweb.org/anthology/I/I11/I11-1100.pdf

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yotogama, D., Flanigan, J., and Smith, Noah A. (2011): Part-of-speech tagging for Twitter: annotation, features, and experiments. *Proceedings of 49th Conference on Computational Linguistics (ACL 2011):* 42–47. http://www.aclweb.org/ anthology/P/P11/P11-2008.pdf

Hu, Y., Talamadupula, K., and Kambhampati, S. (2013): Dude, srsly?: The surprisingly formal nature of Twitter's language. *Proceedings of The 7th International AAAI Conference on Weblogs and Social Media (ICWSM 2013)*. http://www.public.asu.edu/~ktalamad/papers/icwsm13.pdf

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

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

Liu, F., Weng, F., Wang, B., and Liu, Y. (2011): Insertion, deletion, or substitution? Normalizing text messages without pre-categorization nor supervision. *Proceedings of 49th Conference on Computational Linguistics (ACL 2011)*: 71–76. http://www.aclweb.org/anthology/P/P11/P11-2013.pdf

Ljubešić, N., Erjavec, T., and Fišer, D. (2014a): Standardizing tweets with

character-level machine translation. A. Gelbukh (Ed.): *Proceedings of the 15th International Conference CICLing 2014*: 164–175. Lecture Notes in Computer Science. Berlin: Springer.

Ljubešić, N., Fišer, D., and Erjavec, T. (2014b): TweetCaT: a tool for building Twitter corpora of smaller languages. *Proceedings of LREC 9*: 2279–2283. http://www.lrec-conf.org/ proceedings/lrec2014/pdf/834_Paper.pdf

Ljubešić, N., Fišer, D., Erjavec, T., Čibej, J., Marko, D., Pollak, S. and Škrjanec I. (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

Ljubešić, N., Zupan, K., Fišer, D., Erjavec, T. Normalising Slovene data: historical texts vs. user-generated content. *Proceedings of KONVENS 2016:* in print.

Ljubešić, N., Klubička, F., Agić, Ž. and Jazbec I. (2016b): New inflectional lexicons and training corpora for improved morphosyntactic annotation of Croatian and Serbian. *Proceedings of LREC 10*: 4264–4270. http://www.lrec-conf.org/proceedings/lrec2016/pdf/340_Paper.pdf

Mair, C., Hundt, M., Leech, G., and Smith, N. (2002): Short term diachronic shifts in part-of-speech frequencies. A comparison of the tagged LOB and F-LOB corpora. *International Journal of Corpus Linguistics*, 7(2): 245–264.

Noblia, M. V. (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.

Oliva, J., Serrano, J. I., Del Castillo, M. D., and Igesias, A. (2013): A SMS normalization system integrating multiple grammatical resources.

Natural Language Engineering, 19: 121–141.

Pešikan, M., Jerković, J., and Pižurica, M. (2010): *Pravopis srpskoga jezika*. Novi Sad: Matica srpska.

Petrov, S., and McDonald, R. (2012): Overview of the 2012 shared task on parsing the web. *Notes of the First Workshop on SANCL 2012*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.261.2294&rep=rep1&type=pdf

Sidarenka, U., Scheffler, T., and Stede, M. (2013): Rule-based normalization of German Twitter messages. *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology*. https://gscl2013.ukp.informatik.tu-darmstadt.de/fileadmin/user_upload/Group_UKP/conferences/gscl2013/workshops/sidarenka_scheffler_stede.pdf

Tagg, C. (2012): *Discourse of Text Messaging*. London: Continuum.

TVITERASI, TVITERAŠI ALI TWITTERAŠI? IZDELAVA IN ANALIZA NORMALIZIRANEGA NABORA HRVAŠKIH IN SRBSKIH TVITOV

V prispevku predstavimo vzporedno ročno normalizacijo vzorcev, izluščenih iz korpusov hrvaških in srbskih tvitov. Najprej opišemo nabor podatkov, podamo poenotene smernice za anotatorje in predstavimo analizo pretvorb iz nestandardnega v standardni jezik, ki smo jih zajeli v gradivu. Rezultati kažejo, da se zaprte besedne vrste (tiste, ki redkeje sprejemajo nove besede ali pa jih sploh ne sprejemajo, torej predvsem slovnične besedne vrste) pretvarjajo pogosteje kot odprte (tiste, ki pogosteje sprejemajo nove elemente), da so najpogosteje pretvorjene leme pomožni in modalni glagoli, medmeti, členki in zaimki, da so izbrisi pogostejši kot vstavljanja ali zamenjave in da do pretvorb pogosteje prihaja na koncu besed kot na drugih mestih. Ugotovili smo, da si hrvaščina in srbščina delita številne pretvorbne vzorce, ne pa vseh. Medtem ko lahko nekatere razlike pripišemo strukturnim razlikam med jezikoma, se za druge zdi, da bi jih lahko lažje razložili z zunajjezikovnimi dejavniki. Izdelani nabori podatkov in začetne analize se lahko uporabljajo za proučevanje nestandardnega jezika kot tudi za razvoj jezikovnih tehnologij za nestandardne jezikovne podatke.

Keywords: računalniško posredovana komunikacija. korpusi CMC, Twitter, normalizacija