MLO Modern Languages Open

From Fringe to Infrastructure: A Researcher's Journey through South Slavic Language Attitudes on Social Media

COLLECTION:
DIGITAL MODERN LANGUAGES SECTION LAUNCH ISSUE

ARTICLES – DIGITAL MODERN LANGUAGES

DARJA FIŠER
NIKOLA LJUBEŠIĆ
DAMJAN POPIČ

*Author affiliations can be found in the back matter of this article

LIVERPOOL UNIVERSITY PRESS

ABSTRACT

This paper presents a bottom-up approach to building a comprehensive infrastructure for the analysis of user-generated content for several South Slavic languages (Slovene, Croatian, Serbian). The goal of this collaboration was to leverage the available knowhow and language similarity in order to provide language resources and tools for the study of netspeak for all three languages in parallel and with minimal resources. We demonstrate the usefulness of the developed infrastructure for a corpus-based, comparative sociolinguistic investigation of language attitudes by Slovenian, Croatian, and Serbian Twitter users, who have witnessed a rapid codification divergence and reinforcement of national languages after the dissolution of Yugoslavia in the early 1990s.

CORRESPONDING AUTHOR:
Darja Fišer
University of Ljubljana, SI;
Jožef Stefan Institute, SI
darja.fiser@ff.uni-lj.si

TO CITE THIS ARTICLE:
Fišer, D, Ljubešić, N and Popič, D 2021 From Fringe to Infrastructure: A Researcher's Journey through South Slavic Language Attitudes on Social Media. *Modern Languages Open*, 2021(1): 24 pp. 1–13. DOI: *https://doi.org/10.3828/mlo.v0i0.385*

1 INTRODUCTION

The increasing popularity of Web 2.0 has resulted in an unprecedented surge of user-generated and social media content which is rapidly becoming a major source of knowledge and opinion, and is considered a catalyst of bottom-up communication practices that contribute towards the democratization of language. As a consequence, we are seeing a growing need for a thorough understanding of this type of communication, which is shaped significantly by the specific social and technical circumstances in which it is produced. However, researchers attempting to analyse such communication are faced with a number of technical, methodological and ethical challenges; these are only exacerbated for researchers working with nonworld languages, and can be detrimental if not prohibitive for them engaging in impactful international scholarly exchange.

In this paper we showcase our approach to building a comprehensive infrastructure for the analysis of user-generated content (UGC) for several South Slavic languages (Slovene, Croatian, Serbian) in parallel, as initiated by the JANES[1] (Fišer, Ljubešić, and Erjavec) and ReLDI[2] (Samardžić, Ljubešić, and Miličević) projects. The work was primarily motivated by the close relations among the languages in question, but also by the uneven socioeconomic circumstances over the past three decades in the countries where they are spoken and consequently the unequal development of research infrastructure for computational and corpus linguistics, which is most mature in Slovenia and least so in Serbia. The goal of this bottom-up collaboration was to leverage the available knowhow and language similarity in order to provide new language resources and tools for the study of netspeak for the three languages in parallel, with minimal investment of researchers' time, effort, and finances.

By using the same text harvesting criteria, annotation protocols and machine learning approaches, we were able to develop richly annotated (at text- and user level) comparable corpora and a complete chain of robust text-processing tools of (noncanonical) UGC for all three languages with minimal resources. These are available under open licences, the first of their kind for the community of researchers in South Slavic digital languages, culture, and media while also being highly relevant for any other communities that still lack well-documented, widely available and accessible as well as comparable resources and tools for the analysis of UGC. The successful initiative has since matured into a full-fledged CLARIN Knowledge centre CLASSLA (*http://www.clarin.si/info/k-centre/*) which offers expertise on general and specialized language resources and technologies for South Slavic languages (Slovene, Croatian, Serbian, Macedonian, Bulgarian).

The paper is structured as follows. In the next section we give basic background information on the languages involved. The third section presents the development of resources and tools for the analysis of Slovenian, Croatian and Serbian UGC using the transfer approach as well as follow-up activities in the context of the CLASSLA knowledge centre. In the fourth section we showcase the potential of the infrastructural support for cross-lingual and cross-cultural comparative research of UGC. We wrap up the paper with a short discussion on the importance of coordinated development of language infrastructure, especially for closely related and less resourced languages.

2 SOUTH SLAVIC LANGUAGES

Slovenian, Croatian and Serbian belong to the western branch of the South Slavic language group (Friedman).[3] While Slovenian and Croatian use the Latin alphabet exclusively, Serbian is a digraphic language in which both the Latin and the Cyrillic scripts are used (Ivković). In terms of mutual intelligibility of the standard varieties of these three languages, Slovenian is the most distinctive while Croatian and Serbian are considered mutually intelligible, with some phonetic, lexical, and morphosyntactic differences (Golubović and Gooskens). Dialectal language variation blurs the boundaries between the three languages. The most prominent example is probably that of the Kajkavian dialect from north-western Croatia, which is closer to

1 *http://nl.ijs.si/janes/english/*.

2 *https://reldi.spur.uzh.ch*.

3 While this language group also includes Bosnian and Montenegrin, they were not examined in the present study. For more information on Bosnian, see Katičić and for Montenegrin, see Nakazawa.

the Slovenian standard than to the Croatian standard (Kapović). Mutual intelligibility gets even more blurry in UGC, which regularly contains dialectal features but is also rich in noncanonical spelling variation, colloquialisms, slang, and foreign-language elements (Cenni).

3 DEVELOPMENT OF RESOURCES AND TOOLS FOR SOUTH SLAVIC UGC

The main motivation for building language resources and technologies for the South Slavic netspeak was corpus-linguistic research into UGC, which is full of noncanonical language features, such as spelling variation, abbreviations, dialectal elements, colloquialisms, code-switching, etc. This requires large collections of user-generated text, annotated with basic linguistic information, i.e., canonical (normalized) variant of the word form, morphosyntactic description, and lemma. Previous research has demonstrated that language technologies developed for a standard language underperform on nonstandard variants of the same language (Gimpel et al.) and that the most reasonable way forward is to construct manually annotated datasets for nonstandard variants that will enable the adaptation of language technologies to these variants. Additionally, given that most of the UGC does in fact follow the linguistic norm, the corpus also needs to be divided into texts that follow the norm relatively closely, and those deviating more from the norm. Having this information at their disposal, the researchers circumvent the "needle in the haystack" problem and focus directly on the desired noncanonical language features.

3.1 DATA COLLECTION

Due to the widespread usage of all three languages, low technical barriers to harvesting the data as well as the metadata, and a relatively researcher-friendly content redistribution policy (*https://developer.twitter.com/en/developer-terms/policy*), we focused on the microblogging and social networking service Twitter. Even though Twitter already offers its own API for harvesting data from the platform (*https://developer.twitter.com/en/docs/twitter-api*), it is less suited to collecting tweets written in a low-density language, such as Slovene (2.5 million speakers), Croatian (5.6 million speakers), or Serbian (12 million speakers). This is because most of the content published on Twitter is written in high-density languages, which makes it difficult to harvest the much less frequent content in the desired low-density language but also because language identification on Twitter tends to be unreliable, especially for low-density languages (Lui and Baldwin), which introduces a lot of noise in the harvested material with content in languages other than the desired language. This is why we built a simple tool called TweetCat (Ljubešić, Fišer, and Erjavec) that, using the Twitter search API and a set of seed terms, i.e., very frequent words specific to the language of interest, identifies users tweeting in that language, harvests their timeline, as well as their network of friends and followers who are then submitted to the same procedure. The seed terms were manually selected from high-frequency content words that are, however, not in the vocabulary of the related languages. The number of the seed terms was quite small: 40 apiece for Croatian and Serbian and 20 for Slovene. The tool also performs a post-processing step on the collected data, which filters out users that tweet predominantly in a foreign language and thus further cleaning up the collected corpora.

3.2 MANUAL DATA ANNOTATION

In order to build machine learning-based language technologies that are able to process UGC, it was of paramount importance to create manually annotated datasets that would serve for the training of such technologies. Manual annotation was designed to support the following processing steps:

1. text standardness prediction, which returns the level of standardness of a text, e.g. for the English "Because I want that by tomorrow" should be predicted as standard, while "bcz I wanna dat by 2morrow" is nonstandard;
2. text normalization, which transforms noncanonical words into their canonical variants (e.g. "precednikom" to "predsednikom");

3. morphosyntactic tagging, which assigns the morphosyntactic features of each word in running text (so, "predsednikom" gets the Ncmsi tag which means it is a noun, common, masculine, singular, in the instrumental case);
4. lemmatization, which assigns each word in running text its base, or dictionary form (e.g. "predsednikom" to "predsednik");
5. named-entity recognition, which flags personal names, locations, organisations, and other named entities in running text (e.g. Donald Trump gets the [pers] tag which means it is a person name).

The annotation of text standardness (Ljubešić et al., "Predicting") was performed at text level (in our case tweets) and it encoded to what degree a piece of text follows both surface-level standardness (use of punctuation, use of capitalization, character flooding, etc.) and deeper linguistic-level standardness (noncanonical spelling, use of dialectal and colloquial lexical features, foreign-language elements, noncanonical grammatical constructions, etc.). Being able to identify texts that do not follow the linguistic norm was important for two reasons: (1) manually annotated datasets had to contain an overrepresentation of nonstandard content, as this content cannot be processed reliably with standard technologies; (2) having this information available in the corpus enables researchers to focus on the parts of the corpora that contain noncanonical features.

The next three annotation steps — normalization, morphosyntactic tagging, and lemmatization — were key to making the data more accessible to researchers investigating UGC. Normalization was necessary so that querying for specific words was possible regardless of spelling variants used by different users (e.g. instances of the Slovene first-person singular personal pronoun *I* spelled as *js*, *jssst*, *jz*, *jasss* were all normalized to *jaz*), while morphosyntactic tagging and lemmatization were required to enable regular corpus linguistics techniques, such as collocation extraction or complex querying (e.g. looking for occurrences of a specific lemma followed by nouns in the dative case). This was enabled with two manually annotated datasets: Janes-Norm (Erjavec et al., "CMC Training Corpus Janes-Norm 1.2") and Janes-Tag (Erjavec et al., "CMC training corpus Janes-Tag 2.1"). While the process of normalizing words might seem straightforward, it proved to be the most challenging of all the manual annotations performed in the project. This is because normalization is often ambiguous (e.g. in the same context for Slovene, noncanonical *k* could be interpreted as *ki*, Eng. *which*, or as *kot*, Eng. *as*) but can also be performed at different granularity levels (e.g. *žmigauc* can be normalized to *žmigavec*, which is the canonical spelling of an otherwise nonstandard expression, the standard equivalent of which is *smernik*, Eng. *turn signal*). In order to produce a homogeneous and robust dataset, we had do develop precise annotation guidelines, provide good annotator training, and implement synchronization procedures throughout the annotation process (Čibej, Fišer, and Erjavec). The Janes-Tag dataset, a subset of Janes-Norm, has been additionally annotated with morphosyntactic information, lemmas, and named entities. The named-entity recognition task was mostly required to ensure data anonymization so that the JANES corpus could finally be published.

Given the great complexity of the manual annotation and the very high organizational overheads of the first annotation campaign, knowledge transfer and parallel development of comparable datasets for Croatian and Serbian seemed like a very good strategy. By teaming up with researchers from Croatia and Serbia, the process of creating annotation guidelines and providing annotator training was applied to both these languages and comparable manually annotated datasets were produced for both Croatian and Serbian as part of ReLDI project activities, resulting in the Twitter training corpus for Croatian ReLDI-NormTagNER-hr (Ljubešić et al., "Croatian") and its Serbian counterpart ReLDI-NormTagNer (Ljubešić et al., "Serbian"). Based on comparable datasets, a series of cross-lingual comparative studies were performed on the question of standardness of UGC (Fišer et al.; Miličević and Ljubešić; Miličević, Ljubešić, and Fišer). Our rough estimate is that the energy required to set up annotation guidelines for such a complex project for the two additional languages was lowered to a tenth of the resources that were required for Slovenian. Not only were the annotation guidelines for further languages obtained using just a fraction of the resources, but comparability of the annotation schemata was ensured too; this makes for added value that will last for the entire lifespan of the two manually annotated datasets. The costs of the manual annotation itself on the

two datasets were moderately lower than was the case for the Slovenian dataset. During the Slovenian annotation campaign, pilot annotation campaigns were necessary to test and improve the annotation process, which also informed wider improvements of the annotation guidelines and annotator training.

3.3 TOOLCHAIN DEVELOPMENT

For the most part, the development of text-processing tools corresponds to the levels of annotation described in the previous subsection. The tools are described in detail in Fišer et al., so in this section we focus on presentation of the UGC-specific tools that were developed for all three languages in the context of collaboration between the JANES and ReLDI projects.

The first tool to be developed was the text standardness predictor which, given a text, returns two continuous values—one encoding surface-level standardness and the other general linguistic standardness. This tool was used to annotate three levels of text standardness in the UGC corpora (standard, a little nonstandard, very nonstandard) but also for building the remaining manually annotated datasets where oversampling of less canonical texts was paramount. If this sampling had not been performed, the manually annotated datasets would have contained a low number of nonstandard features and the adaptation of tools to these features would have been very limited.

The ReLDI-tagger, which is based on conditional random fields (CRF) and achieves state-of-the-art results on part-of-speech tagging and lemmatization for Croatian and Serbian (Ljubešić et al.) as well as for Slovenian (Ljubešić and Erjavec), was primarily built for the processing of standard language but was adapted to work on nonstandard language too (Ljubešić, Erjavec, and Fišer). The main modification was the use of Brown clusters—a predecessor of the now omnipresent word embeddings, thereby encoding the distributional fingerprint of a word, such as all the variants of the personal pronoun *I* mentioned previously (*jaz*, *jst*, *jssss*, *js*, etc.) which have the same fingerprint. This demonstrates that machine learning tools, once built, simply need comparable data in other languages, making the machine learning-based language tools highly transferable onto other languages. However, the effort required for manual annotation in these other languages should be given due consideration if good results are desired.

3.4 BEYOND UGC

Following the successful collaboration between the JANES and the ReLDI projects on building the language technology infrastructure for the three South Slavic languages, the CLARIN Knowledge Centre for South Slavic Languages, CLASSLA,[4] was established in March 2019 under the coordination of the Slovenian CLARIN.SI and the Bulgarian CLADA-BG research infrastructures. Language coverage has now been extended to Bulgarian and Macedonian, primarily for standard language. The main components of the knowledge centre are an email-based helpdesk, FAQs, relevant guidance documents for all the mentioned languages, the CLARIN.SI concordancer that offers various South Slavic corpora, and the CLARIN.SI repository which contains many resources and tools for various South Slavic languages. The main planned activities for the CLASSLA knowledge centre are — similar to both the JANES and the ReLDI projects — to coordinate development of further language technologies, but also to build and develop a user base for the developing infrastructure. CLASSLA's future activities will involve development of support for the processing UGC for Macedonian and Bulgarian. Another very timely development is the improvements in speech technology, as CLASSLA would have a significant impact if it produced spoken corpora for all South Slavic languages which could be used to train speech-to-text systems.

4 USE CASE

In this section we demonstrate the potential of the developed resources and tools for comparative cross-lingual and cross-cultural analyses of UGC in the ex-Yugoslavia region. We examine attitudes of Twitter users towards their national languages from the sociolinguistic concept of prestige of standard language ideology and the social status attached to a speaker's

4 *https://www.clarin.si/info/k-centre/*.

language skills. The investigation is motivated by the specific sociolinguistic and political contexts of the post-Yugoslav era, in which we saw rapid and highly ideological divergence of the language standards and an emphasis on the idea of a national language (Požgaj Hadži and Balažic Bulc). Using comparable sets of language-use seed words, we analyse Slovenian, Croatian, and Serbian tweets that comment on (in-)correct language use on Twitter or in broader public discourse, in order to determine the prevailing attitudes towards language and language use in the respective communities.

4.1 SOCIOPOLITICAL CONTEXT

4.1.1 Slovenia

The Slovenian language environment is highly directive (Škiljan) and standard language competence plays a major role in a speaker's social prestige. This, characterized by a history inundated by fears of more powerful languages such as German and Serbo-Croatian prevailing over Slovene, has become the symbol of Slovenian national identity, and linguistic competence has become a prestige argument of power (Thomas). While language use is inherent to societal stratification (Labov), the Slovenian language environment is special in that it has developed a particular dislike of outside elements as a defence against foreign dominance. This is also one of the main reasons for the pronounced importance in Slovene of the linguistically 'correct' and 'good', especially in terms of orthography, which is very complex and difficult as the rules are often based on historical usage and a highly structured grammar. In contemporary society, even though Slovene has recognized status as a European language, purist tendencies remain in a substantial part of the Slovenian normative linguistic community as well as the general public, which is also reflected in the fear that is embedded in the national consciousness—i.e., that it is next to impossible to write in accordance with all the rules and by-laws that (supposedly) exist in Slovene and the normative tradition behind its codification (Popič and Logar). In our previous study (Popič and Fišer), we have found that the attitude of Slovene Twitter users is distinctly normativist and that the argument of linguistic expertise (or lack thereof) is often used as an instrument of power in public debate. The goal of this analysis is to compare these attitudes with those identified in the Croatian and Serbian Twittersphere.

4.1.2 Croatia

After the dissolution of Yugoslavia in the early 1990s, the Croatian standard language saw the most changes of all the languages spoken in former Yugoslavia, especially at the lexical level. This had a big symbolic charge and it helped to reinforce the Croatian national identity (Požgaj Hadži and Balažic Bulc). As a consequence, speakers who used appropriate lexis were patriots, while speakers who did not suddenly no longer counted as speakers of the 'pure' Croatian language and were assigned various political labels (Kordić). It was therefore not uncommon that speakers felt uncomfortable using their own language and struggled to find the 'right' words (Badurina). This movement, which included removal of words of foreign—especially Serbian—origin was supported by the political elite and implemented by the top-ranking academic institutions (Požgaj Hadži and Balažic Bulc), creating fertile ground for the publication of dictionaries and style guides that distinguished between Croatian and Serbian, and listed words that were superfluous or even forbidden in Croatian.

4.1.3 Serbia

Unlike the radical changes in Croatian, the Serbian standard language did not change in the politically tumultuous 1990s. In line with this position, Serbian language and national identity were in the hands of informal groups. The reason for this was that there was no real need for the Serbian language to reaffirm its identity by distancing itself from Croatian and changing its lexis, orthography, or grammar (Bugarski, "Jezička"). This does not mean that the nationalization of the Serbian language was not taking place, only that the nationalism observed in Serbian was of a more reductionist nature, the kind that is primarily interested in fencing off and protecting territory (Bugarski, "*Portret*"). In comparison to Slovenia and Croatia, a unique dimension of Serbian language identity is the Latin versus Cyrillic script dichotomy. As the Latin script prevailed in the second half of the 20th century, many saw this as an attack on Serbian autonomy, as Cyrillic script is seen as one of the hallmarks of the Serbian culture and identity. In response, Article 10 of the Serbian constitution was changed in 2006, abolishing

the equal status of the Latin script, despite its widespread usage. Hence, the choice of script here has never merely been a question of digraphia; it can also be a question of language ideology.

4.2 STUDY DESIGN

4.2.1 Dataset

The dataset used in the analysis was extracted from the corpora described in Section 3. Although distinct aspects of language and linguistics are relevant for the specific respective cultural environments in Slovenia, Croatia, and Serbia (such as the *comma* for Slovene, *foreign words* for Croatian, and *Cyrillic* for Serbian), we have taken into account five of the most important headwords for the topics that span all three languages in order to maximize the comparative potential of the analysis (see ***Table 1***). Since the corpora are normalized, morphosyntactically annotated, and lemmatized, we queried lemmas, thereby including all the spelling variations and word forms of the selected keywords in our dataset. This is important because all three languages included in this investigation are highly inflecting. Some of the headwords are polysemous and can occur in tweets that have nothing to do with the expression of language attitudes (e.g. *jezik* which in all three languages can also mean *tongue* or be used metaphorically, or *Croatian* and *Serbian* which can also be used as adjectives premodifying anything, anyone, or anything from Croatia/Serbia), which is why out-of-context frequency information is not directly useful but does serve as a good indicator of how important these concepts are for speakers in the three communities. For our analysis, only relevant tweets with the headword used in the appropriate sense were manually selected.

SLOVENIAN		CROATIAN		SERBIAN	
HEADWORD	FREQUENCY	HEADWORD	FREQUENCY	HEADWORD	FREQUENCY
jezik	18,102	jezik	8,138	jezik	46,079
language	119.52 / mio	*language*	687.00 / mio	*language*	515.89 / mio
pravopis	667	pravopis	510	pravopis	4,959
orthography	4.40 / mio	*orthography*	43.05 / mio	*orthography*	55.52 / mio
slovnica	1,461	gramatika	405	gramatika	5,636
grammar	9.65 / mio	*grammar*	34.19 / mio	*grammar*	63.10 / mio
slovar	1,653	rječnik	540	rječnik	58
dictionary	10.91 / mio	*dictionary*	45.59 / mio	*dictionary*	0.65 / mio
slovenščina	7,347	hrvatski	24,261	srpski	130,644
Slovenian	48.51 / mio	*Croatian*	2,048.08 / mio	*Serbian*	1,462,67 / mio
∑ tokens	151,457,091	∑ tokens	11,845,710	∑ tokens	89,318,700

Table 1 Slovene, Croatian, and Serbian headwords with their absolute and normalized frequencies.

The total size of the corpora is partially related to our harvesting approach, which started first for Slovenian but is also a reflection of the relative popularity of Twitter in the three communities: while it is broadly used by the general population in Slovenia, it is especially popular with younger users for ephemeral communication in Serbia, and in Croatia mostly by politicians and journalists for dissemination of information. Overall, language is a prominent topic in all three communities but is much more frequently discussed on Serbian and Croatian Twitter compared to Slovene. This is probably not surprising as, due to Serbian and Croatian's close proximity because of politically motivated blending into Serbo-Croatian in Yugoslavia, after the dissolution of Yugoslavia language ideology and reinforcement of distinctive language identities were used as major elements of national identities. Unlike *slovenščina/Slovenian*, which is a noun, *hrvatski/Croatian* and *srpski/Serbian* only exist in adjectival form, which is another reason why they are used a lot more broadly, and not only in reference to the language itself, as is reflected in usage frequency. A similar situation holds for *jezik/language*, which is very general and often also employed in senses not relevant to our study (e.g. the body part, part of a boot, metaphorical, etc.). Nevertheless, it is interesting to observe that, as per normalized frequency, in Croatian

and Serbian the level of usage is comparable and is five times more frequent than in Slovenian. *Grammar* as well as *orthography* are most prevalent in Serbian, where they are, relatively speaking, used twice as much as in Croatian, and an order of magnitude more often than in Slovenian. Due to the polysemy of the word *rječnik*/*dictionary* in Croatian and Serbian, in which it also means *vocabulary*, usage frequency cannot be directly compared to their Slovenian equivalent *slovar* which does not carry this additional meaning. But when Croatian and Serbian are compared, the results, which are two orders of magnitude higher for the former compared to the latter, do reflect the effects of the recent lexical interventions in Croatian to differentiate it from Serbo-Croatian, which was heavily based on Serbian and the resulting sensitivity towards new lexical items by Croatian speakers.

4.2.2 Typology

The typology was first developed in our previous study (Popič and Fišer) which helped us annotate attitude and tenor of Slovenian tweets pertaining to language and linguistics. In this paper, we apply the same typology to Croatian and Serbian data and, based on the observed attitudes in the sample, have extended it with three additional categories. The modified typology comprises fourteen categories (see ***Table 2***). It should be pointed out that not all the categories pertain to *attitudes* towards language but also include other the types of discourse in which language-related matters are featured, in order to cover all sentiments about language use in the community:

- attitudes towards the standard language and its status (e.g. *nationalist*);
- attitudes towards the rules, deviations, and errors (e.g. *lamenting*);
- attitudes towards people who use language in a certain way (mostly in an unsuitable way, as considered by the person expressing a specific attitude) (e.g. *dismissive*); and
- the types of discourse (or their function) in which language-related matters are featured (e.g. *idiomatic, informative*).

Table 2 Categories used for annotation with illustrative examples for the headword 'orthography' in Slovenian.

ATTITUDE	SLOVENE	CROATIAN	SERBIAN
inquisitive	Veste, da se po Slovenskem pravopisu imena praznikov datumov, razen tistih, ki so izpeljana iz priimkov, pišejo z malo začetnico? *Do you know that according to Slovene orthography, holiday names, except those that are derived from person names, are not capitalized?*	Pazite li na pravopis i gramatiku na društvenim mrežama? *Do you pay attention to orthography and grammar on social media?*	Da li ima neko knjigu pravopisa, priručnik neki? *Does anybody have the orthography guide or some other language manual?*
informative	Nekaj napotkov je menda v slovenskem pravopisu. *Some suggestions are surely provided in the Slovene orthography guide.*	Relevantni pravopis je online i jednostavan je za upotrebu http://t.co/oFvpzNyDci pa nema više izlike za nepismenost:) *The relevant orthography guide is online and easy to use http://t.co/oFvpzNyDci, so there is no excuse for illiteracy anymore :)*	Đ se po novom pravopisu, u zvaničnoj upotrebi, piše Dj. *According to the new orthography guide, Đ is spelled as Dj in formal language.*
lamenting	to tudi meni živec potegne, ker ne poznajo niti osnov pravopisa, pa nastanejo taka skropucala, da je groza *i also find it irritating that they don't know even the basics of the orthography, and produce such terrible hodgepodge*	Meni hrvatski pravopis je sve gori. Užas *I think the Croatian orthography is getting worse and worse. Horror*	O jebem ti državu kad se pravopis promeni minimum 4 puta samo za mog školovanja *To hell with a country that changes its orthography 4 times just during my schooling*
jocular	Pravopis je zarota levosučnih akademikov. *The orthography is a conspiracy of left-leaning academics.*	Inace ne volim viceve, ali novom pravopisu u Hrvata, kao laik, povremeno moras priznati komicni momenat. *I normally don't like jokes but, to a layman, the new Croatian orthography is hilarious in certain places.*	Ana, mislim da će da me ubace u pravopis neki ako nastavim ovako *Ana, I think they will include me in the orthography guide if I continue like this.*

(*Contd.*)

ATTITUDE	SLOVENE	CROATIAN	SERBIAN
dismissive	Siromak si. V pameti in znanju pravopisa. *You're a poor man. In terms of your intelligence and knowledge of orthography.*	Ne znam prema kojim to kriterijima se danas zapošljavaju novinari koji fejlaju već na pravopisu. *I don't know which criteria are used to employ journalists who fail already at orthography.*	ŠUPČINOO Pre svega, ne znaš pravopis. *Idiot. First of all, you don't know orthography.*
defensive	Aja, a zdej smo pa pri pravopisu. A je argumentov zmanjkal. *Oh, so now we're at orthography. Have you run out of arguments?*	Ne vjeruj ženi s lošim pravopisom. *Don't trust a woman with poor orthography.*	Kad baba pocne da ti kenja o srpskom pravopisu i nacinu izrazavanja danasnjih tinejdzera... *When the old lady starts to go on and on about Serbian orthography and the expression of today's teenagers...*
apologetic	Sram me je za "moj" pravopis ... *I'm ashamed of 'mine' orthography.*	nemojte mi o pravopisu na rano jutro, nisam nepismena samo lijena *spare me the orthography first thing in the morning, I'm not illiterate, just lazy*	Zbog tvitera cu iz pravopisa da imam keca *Because of twitter I'll fail orthography.*
idiomatic	Kdor se še nikoli ni zatipkal, naj vrže pravopis vame *Let him who has not misspelled cast the orthography guide in me.*	»Pravopis sa zvijezdama« bih gledao. *I'd watch "Orthography with the stars".*	Godine prolaze, greške u pravopisu ostaji. *Years go by, orthographic errors remain.*
nationalist	Če se imate za velikega Slovenca, se najprej naučite pravopisa. *If you consider yourself a patriot, learn orthography first.*	Spasimo spomenik Ljudevitu Gaju jer je bio ustaša i kao ustaša stvorio je ustaški pravopis 1830 *Let's save the monument to Ljudevid Gaj because he was a Ustashe terrorist and as such created the Ustashe orthography in 1830*	Ako si hejter i pljuvač nauči barem svoj maternji jezik i pravopis. *If you're a hater and a criticizer at least learn your mother tongue and the orthography.*
purist	Pravopis deklica, pravopis :) ;) *Orthography my girl, orthography :) ;)*	Zašto nitko ne provjeri pravopis prije tiska? Sramočenje *Why doesn't somebody check the Orthography guide before sending it to print? Disgrace.*	Kažu da je pravopis na tviteru nebitan. Ja baš izbegavam pravopisne greške. Mora biti savršeno napisan svaki tvit. *They say orthography doesn't matter on Twitter. But I consciously avoid orthographic errors. Each tweet has to be written perfectly.*
neutral	@user Razumem, prav res govoris o pravopisu :). Sem mislila, da te je opredelitev *Oh, I see, you really are talking about orthography, I thought that*	/	/
praising	/	Meni je pravopis bio najdrazi predmet u skoli. *Orthography was my favourite subject at school.*	/
anti-purist	/	Danas ne postoji jedinstven pravopis – ima ih čak četiri. *A single orthography doesn't exist this day and age – there's as many as four.*	Jebeš pravopis, daj nešto da jedemo *To hell with orthography, let's have something to eat.*
anti-nationalist	/	/	Sramota me što 80% internet-nacionalista ne zna pravopis (nećemo računati reč „srbski") *I'm embarrassed that 80% of internet-nationalists don't know their orthography (disregarding the word »srbski« [author's note: the correct spelling is »srpski«]).*

While the *inquisitive* attitude refers to actual questions regarding language (use), the *informative* attitude provides explanation(s) on language use and/or answers to questions on the use of a particular linguistic element. The *lamenting* attitude covers instances of exasperation about the perceived difficulty of the use of a language or one of its features. The *jocular* attitude comprises examples that apply irony to a linguistic situation or one of its features, whereas *dismissive* tweets use (alleged) misuse of a language feature to portray a person or institution as

incompetent. The *apologetic* and *defensive* attitudes, on the other hand, aim to explain or justify one's 'transgressions'. While the *apologetic* approach involves giving reasons for a particular example of nonstandard language use (either not knowing the rules or more pragmatic excuses such as haste, typos, etc.), the *defensive* attitude conveys a stronger reaction, either against the user(s) exposing the misuse or against the relevance of correctly using the linguistic element in question (normally in contrast to its meaning, which is emphasized over its form). Messages with a positive attitude towards fellow users, decision-makers, rules, reference books, etc. are annotated as *praising*. The final categories are *idiomatic* references to language-related matters (in our example by using 'comma' as a metaphorical (metonymic) expression for 'language' or 'text') as well as *nationalist* and *purist* tweets. We differentiate between the latter on the basis of their reference: we placed tweets stressing the necessity for 'proper' use of the standard language on the basis of nationality in the nationalist category, while tweets purporting to the correctness of someone's language use based on correctness alone went in the purist category. The opposite attitudes are classified as *anti-nationalist* and *anti-purist*. Finally, we placed all tweets without a discernible attitude, tenor, or discursive type into the *neutral* category.

4.2.3 Annotation procedure

For each language, we annotated fifty random tweets per headword, or 750 tweets in total. Annotation was performed by a single annotator. In order to ensure consistent and objective annotation, the annotator was asked to rely exclusively on the available explicit (linguistic as well as extralinguistic) attitudinal markers. For instance, if a tweet shows no sign of being dismissive (even though it could be interpreted as such), it should not be tagged as being dismissive. The annotator was allowed to assign more than one label to a single tweet. If a tweet admitted of more than one interpretation, the annotator was asked to assign all the relevant labels. In cases where a headword was part of a proper name, the tweet was classified as neutral.

4.3 RESULTS

As can be seen from ***Table 3***, the headwords *language* and *orthography* have been annotated with the highest number of attitudinal labels, while the headword *dictionary* has received the fewest. The most frequently observed attitudes are *jocular* and *dismissive*, while *anti-national* is the rarest. The highest number of attitudes have been consistently assigned to Serbian tweets while Slovenian has received the fewest.

Table 3 Overview of the identified attitudes per headword per language.

HEADWORDS	LANGUAGE				ORTHOGRAPHY				GRAMMAR				DICTIONARY				SLOVENE/ CROATIAN/SERBIAN				∑			
ATTITUDES	SI	HR	SR	∑	SI	HR	SR	∑	SI	HR	SR	∑	SI	HR	SR	∑	SI	HR	SR	∑	SI	HR	SR	GRAND
apologetic	2	1	9	12	2	2	5	9	2	4	7	13	0	4	2	6	0	1	6	7	6	12	29	47
defensive	1	1	2	4	1	4	6	11	4	12	6	22	0	2	2	4	2	1	2	5	8	20	18	46
dismissive	7	2	8	17	13	9	15	37	9	5	14	28	6	6	5	17	11	4	17	32	46	26	59	131
idiomatic	7	3	12	22	1	0	2	3	0	0	0	0	1	0	4	5	0	0	0	0	9	3	18	30
informative	4	1	0	5	17	5	1	23	14	5	1	20	17	20	11	48	15	15	2	32	67	46	15	128
inquisitive	6	1	0	7	4	4	0	8	5	0	1	6	7	1	1	9	4	7	0	11	26	13	2	41
jocular	0	11	13	24	7	7	26	40	7	15	15	37	6	8	15	29	5	5	9	19	25	46	78	149
lamenting	5	6	9	20	4	2	7	13	11	7	3	21	4	6	9	19	10	5	7	22	34	26	35	95
nationalist	7	7	1	15	0	2	1	3	1	0	1	2	4	6	0	10	3	21	15	39	15	36	18	69
neutral	7	17	0	24	4	9	0	13	3	2	0	5	6	9	1	16	10	1	0	11	30	38	1	69
praising	3	5	13	21	0	1	0	1	0	0	0	0	1	1	0	2	0	2	0	2	4	9	13	26
purist	7	2	8	17	0	9	17	26	1	3	19	23	1	3	3	7	0	0	7	7	9	17	54	80
anti-nationalist	2	6	1	9	0	0	6	6	0	0	0	0	0	1	0	1	0	4	3	7	2	11	10	23
anti-purist	1	1	2	4	0	2	5	7	0	7	7	14	0	0	0	0	0	0	0	0	1	10	14	25
grand total	59	64	78	201	53	56	91	200	57	60	74	191	53	67	53	173	60	66	68	194				

There are significant differences among the attitudes observed in the datasets across the languages: while the highest-ranking attitude in Serbian is *jocular*, and *informative* in Slovenian, Croatian stands in the middle of the two with a tie between exactly these two attitudes. The biggest cross-lingual discrepancies are observed for the *jocular* attitude, which is very common in the Serbian dataset and much less so in the Slovenian; and the *informative* which is very frequent in Slovenian and rare in Serbian. On the other hand, the *lamenting* attitude has the most even distribution across all three languages. It is interesting to note that both *purist* and *anti-purist* are most frequently observed in Serbian, and the same is true for *nationalist* and *anti-nationalist* in Croatian. All these four attitudes are very rare in Slovenian. Both *dismissive* and *praising* are most often observed in Serbian, as are *lamenting* and *apologetic*. However, only *lamenting* is also frequent in Slovenian—where all the others are much rarer. *Defensive* is most common in Croatian, while *inquisitive* is most characteristic of Slovenian, and rare in both of the other two languages.

The distribution of the observed attitudes is very uneven across the headwords. Even though the two concepts are very similar, and in principle the expressions can be used interchangeably as near synonyms in most contexts, in the Twitter dataset *neutral* and *jocular* attitudes prevail for *language*, while *informative*, *dismissive*, and *nationalist* are the most frequent for *Slovene/Croatian/Serbian* respectively. *Orthography* and *grammar*, on the other hand, seem to express similar attitudes, that is, *dismissive* and *jocular*. Interestingly, tweets with an indiscernible attitude (neutral) are relatively rare in all three languages, especially in Serbian where it was observed only once.

These results on the one hand reflect the differences in by whom, why, and how Twitter is used in these three communities. In Serbia, it seems largely to be young people who use Twitter for ephemeral communication (lots of joking, ironic, sarcastic, and self-effacing tweets). In Slovenia, a broad spectrum of the active population uses Twitter both as a pastime as well as for professional purposes (frequent information dissemination or seeking, but also frequent lamentation or complaints). In Croatia, the main users seem to be journalists and representatives of various institutions, employing the platform as a dissemination channel (lots of information dissemination, but also a good deal of national identity reinforcement).

However, the results also very clearly reflect the wider sociopolitical context (e.g. Croatia joining the EU and gaining official status as an EU language is a big success; Croatians display a particularly strong need to differentiate between Croatian, Serbian, and Bosnian), the changes in and effects of language policy (e.g. orthography reform, new reference manuals, and lexical changes are confusing for Croatian speakers; Slovenian users do not find reference manuals exhaustive/up-to-date/user-friendly; Slovenian/Croatian users have different opinions on new legislation on the foreign/domestic names of private companies), the social status and stereotypes of the national language, foreign languages, and dialects (e.g. Slovenian/Croatian/Latin are hard to learn; French/English/Spanish are attractive languages, German is not an attractive language; north-eastern Slovenian dialects have low social status; speaking (many) foreign languages is respected; Slovene/Serbian are boring school subjects), the social status and stereotypes of language professionals (e.g. language teachers must always be perfect; language editors are an annoyance; journalists are illiterate), language skills are an important factor in a person's social status (e.g. Slovenian users cannot respect politicians with poor language skills; Serbian users are not attracted to nor can be in a relationship with someone of poor language skills; Serbians make lots of self-deprecating remarks and jokes to save face apropos to their language skills; users from all three communities differentiate between communicative situations where it is important to use standard language and others where it is acceptable/desirable to deviate from the norm).

For all three languages, we have observed a similar dismissive (and in many cases also nationalist) tone that requires good language skills from anyone in public position (e.g. politicians) and discredits all those who do not follow the norm in their public communication (on Twitter and beyond). For each language we have also observed a common apologetic tone for any linguistic errors in tweets. Especially characteristic for the Slovenian dataset were frequent questions on how correctly to follow the orthographic rules, something not witnessed in the Croatian and Serbian datasets. In those datasets we observe frequent news on the new orthography guides that have been published, and the related dissatisfaction and frustrations with them in the community. In the Croatian dataset, comments or uncertainties about lexis are commonly expressed, as well as messages that highlight the distinctions between Croatian and other similar languages (mostly Serbian and Bosnian). The Slovenian and Croatian datasets contain comments on various regional dialects, while this was not observed in the Serbian dataset. The Serbian dataset contains frequent praising (or dismissive) comments about people who can (or cannot)

speak (several) foreign languages, as well as positive attitudes on multilingualism. Remarks about users' followers possibly not understanding Serbian are frequent as well, again something not seen in the Slovenian and Croatian datasets and indicating a more outward-looking orientation among Serbian Twitter users, which is most likely related to their younger age profile but also to their more emigrant culture due to there being a worse socioeconomic situation in Serbia.

5 CONCLUSIONS

In this paper, we have presented a bottom-up approach to building a comprehensive infrastructure for the analysis of UGC for several South Slavic languages (Slovene, Croatian, Serbian) that has since been formalized and extended into a knowledge centre for South Slavic languages, offering resources and tools for the analysis of their standard and nonstandard written and spoken forms. In the second part, we showcased the usefulness of the developed infrastructure on a cross-lingual and cross-cultural analysis of tweets pertaining to language use and attitudes by user communities from former Yugoslavia.

We hope that our experience will motivate further research in infrastructure development methodology in the community at large, but especially more coordination of infrastructure development for related languages, particularly in the case of languages that lack the socioeconomic support necessary for the development of top-down language technology infrastructure. Such coordination is their best chance to kickstart infrastructure development and enhance the functioning of language in the digital age. It would be interesting also to compare our results with the attitudes expressed by speakers of major world languages that have similarly directive linguistic traditions, such as German and French, as well as those with a more liberal orientation, such as English.

ACKNOWLEDGEMENTS

The work described in this paper was funded by the Slovenian Research Agency within the Slovenian-Flemish bilateral basic research project "Linguistic landscape of hate speech on social media" (N06-0099 and FWO-G070619N, 2019–2023) and the research programmes "Slovene language – basic, contrastive, and applied studies" (P6-0215) and "Language resources and technologies for Slovene" (P6-0411).

AUTHOR AFFILIATIONS

Darja Fišer *orcid.org/0000-0002-9956-1689*
University of Ljubljana, SI; Jožef Stefan Institute, SI
Nikola Ljubešić *orcid.org/0000-0001-7169-9152*
Jožef Stefan Institute, SI
Damjan Popič *orcid.org/0000-0003-0229-5343*
University of Ljubljana, SI

REFERENCES

Badurina, Lara, Boris Pritchard, and Diana Stolac. *Jezična norma i varijeteti*. Zagreb: Hrvatsko Društvo Za Primijenjenu Lingvistiku, 1999.

Bugarski, Ranko. "Jezička politika i jezička stvarnost u Srbiji posle 1991. godine." *Jezik između lingvistike i politike*, edited by Vesna Požgaj Hadž. Beograd: Bibioteka XX Vek, 2013, pp. 91–111.

Bugarski, Ranko. *Portret Jednog Jezika*. Beograd: Biblioteka XX Vek, 2012.

Cenni, Irene. "Multilingualism 2.0: Language Policies and the Use of Online Translation Tools on Global Platforms." *Argentinian Journal of Applied Linguistics* 7.1 (2019): 79–92.

Čibej, Jaka, Darja Fišer, and Tomaž Erjavec. "Normalisation, tokenisation and sentence segmentation of Slovene tweets." *Proceedings of Normalisation and Analysis of Social Media Texts (NormSoMe)*. Portorož, 2016, pp. 5–10.

Erjavec, Tomaž, et al. "CMC Training Corpus Janes-Norm 1.2." *Slovenian Language Resource Repository CLARIN.SI*. 2016. *http://hdl.handle.net/11356/1084*

Erjavec, Tomaž, et al. "CMC training corpus Janes-Tag 2.1." *Slovenian language resource repository CLARIN.SI*. 2019. *http://hdl.handle.net/11356/1238*

Fišer, Darja, et al. "Comparing the nonstandard language of Slovene, Croatian and Serbian tweets." *Simpozij Obdobja* 34 (2015): 225–31.

Fišer, Darja, Nikola Ljubešić, and Tomaž Erjavec. "The Janes Project: Language Resources and Tools for Slovene User Generated Content. " *Language Resources and Evaluation* 54.1 (2020): 223–46. DOI: https://doi.org/10.1007/s10579-018-9425-z

Friedman, Victor. "Balkans as a Linguistic Area." *Encyclopedia of Language & Linguistics*, edited by Keith Brown. 2nd edn, vol. 1. Oxford: Elsevier, 2006, pp. 657–72. DOI: https://doi.org/10.1016/B0-08-044854-2/00178-4

Gimpel, Kevin, et al. "Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments." *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*. Oregon: Association for Computational Linguistics, 2011, pp. 19–24.

Golubović, Jelena, and Charlotte Gooskens. "Mutual intelligibility between West and South Slavic languages." *Russian Linguistics* 39.3 (2015): 351–73. DOI: https://doi.org/10.1007/s11185-015-9150-9

Hadži, Vesna Požgaj, and Tatjana Balažic Bulc. "(Re) standardizacija v primežu nacionalne identitete: primer hrvaškega, srbskega, bosanskega in črnogorskega jezika." *Slovenščina 2.0: Empirical, Applied and Interdisciplinary Research* 3.2 (2015): 67–94. DOI: https://doi.org/10.4312/slo2.0.2015.2.67-94

Ivković, Dejan. "Pragmatics meets ideology: Digraphia and non-standard orthographic practices in Serbian online news forums." *Journal of Language and Politics* 12.3 (2013): 335–56. DOI: https://doi.org/10.1075/jlp.12.3.02ivk

Kapović, Mate. "The position of Kajkavian in the South Slavic dialect continuum in light of old accentual isoglosses." *Zeitschrift fur slawistik* 62.4 (2017): 606–20. DOI: https://doi.org/10.1515/slaw-2017-0038

Katičić, Radoslav. "Undoing a "unified Language": Bosnian, Croatian, Serbian." *Undoing and Redoing Corpus Planning*, edited by Michael Clyne. Berlin: De Gruyter Mouton, 2016, pp. 165–92.

Kordić, Snježana. *Jezik I Nacionalizam*. Zagreb: Durieux, 2010.

Labov, William. *The Social Stratification of English in New York City*. Cambridge University Press, 2006. DOI: https://doi.org/10.1017/CBO9780511618208

Ljubešić, Nikola, et al. "Croatian twitter training corpus ReLDI-NormTagNER-hr 2.1." *Slovenian language resource repository CLARIN.SI*. 2019. http://hdl.handle.net/11356/1241

Ljubešić, Nikola, et al. "Predicting the level of text standardness in user-generated content." *Proceedings of the International Conference Recent Advances in Natural Language Processing*. Hissar: Incoma Ltd. Shoumen, 2015, pp. 371–8.

Ljubešić, Nikola, et al. "Serbian twitter training corpus ReLDI-NormTagNER-sr 2.1." *Slovenian language resource repository CLARIN.SI*. 2019. http://hdl.handle.net/11356/1240

Ljubešić, Nikola, Darja Fišer, and Tomaz Erjavec. "Tweet-cat: a tool for building twitter corpora of smaller languages." *Proceedings of LREC*. Reykjavik: European Language Resources Association, 2014, pp. 2279–83.

Ljubešić, Nikola, and Tomaž Erjavec. "Corpus vs. lexicon supervision in morphosyntactic tagging: the case of Slovene." *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož: European Language Resources Association, 2016, pp. 1527–31.

Ljubešić, Nikola, Tomaž Erjavec, and Darja Fišer. "Adapting a state-of-the-art tagger for south slavic languages to non-standard text." *Proceedings of the 6th Workshop on Balto-Slavic natural language processing*. Valencia: Association for Computational Linguistics, 2017, pp. 60–8. DOI: https://doi.org/10.18653/v1/W17-1410

Lui, Marco, and Timothy Baldwin. "Accurate language identification of twitter messages." *Proceedings of the 5th workshop on language analysis for social media*. Gothenburg: Association for Computational Linguistics, 2014, pp. 17–25. DOI: https://doi.org/10.3115/v1/W14-1303

Miličević, Maja, and Nikola Ljubešić. "Tviterasi, tviteraši or twitteraši? Producing and analysing a normalised dataset of Croatian and Serbian tweets." *Slovenščina 2.0: empirical, applied and interdisciplinary research* 4.2 (2016): 156–88. DOI: https://doi.org/10.4312/slo2.0.2016.2.156-188

Miličević, Maja, Ljubešić, Nikola, and Fišer, Darja. "Birds of a Feather Don't Quite Tweet Together: An Analysis of Spelling Variation in Slovene, Croatian and Serbian Twitterese." *Investigating Computer-Mediated Communication: Corpus-based Approaches to Language in the Digital World*, edited by Darja Fišer and Michael Beißwenger. Ljubljana: University Press, Faculty of Arts, 2017, pp. 14–43.

Nakazawa, Takuya. "The Making of "Montenegrin Language." *Nationalism, Language Planning, and Language Ideology after the Collapse of Yugoslavia (1992–2011)*. Südosteuropäische Hefte 4.1 (2015): 127–41.

Popič, Damjan, and Darja Fišer. "Fear and Loathing on Twitter: Attitudes towards Language." *Media Corpora for the Humanities (cmccorpora17)*: 61.

Popič, Damjan, and Logar, Nataša. "Med dvema ognjema: kje stoji vejica v slovenskih gimnazijah." *Slovnica in slovar – aktualni jezikovni opis*, edited by Mojca Smolej. Ljubljana: University Press, Faculty of Arts, 2015, pp. 619–27.

Samardžić, Tanja, Ljubešić, Nikola, Miličević, Maja. "Regional Linguistic Data Initiative (ReLDI)." *Proceedings of the Fifth Workshop on Balto-Slavic Natural Language Processing*. Hissar: Incoma Ltd. Shoumen, 2015, pp. 40–2.

Škiljan, Dubravko. *Javni jezik: k lingvistiki javne komunikacije*. Ljubljana: Studia Humanitatis, 1999.

Thomas, George. "The Impact of Purism on the Development of the Slovene Standard Language." *Slovenski jezik – Slovene Linguistic Studies* 1 (1997): 133–52. DOI: https://doi.org/10.17161/SLS.1808.800

TO CITE THIS ARTICLE:
Fišer, D, Ljubešić, N and Popič, D 2021 From Fringe to Infrastructure: A Researcher's Journey through South Slavic Language Attitudes on Social Media. *Modern Languages Open*, 2021(1): 24 pp. 1–13. DOI: https://doi.org/10.3828/mlo.v0i0.385

Published: 17 December 2021

Modern Languages Open is a peer-reviewed open access journal published by Liverpool University Press.