Nikola Ljubešić*, Tomaž Erjavec, Maja Miličević Petrović, and Tanja Samardžić

Together We Are Stronger: Bootstrapping Language Technology Infrastructure for South Slavic Languages with CLARIN.SI

Abstract: In this chapter we describe the recent developments in language technology infrastructure building for three South Slavic languages – Slovenian, Croatian, and Serbian. These developments are primarily the result of intense coordination between different projects. Our experience shows that the infrastructure for language technologies can be significantly improved even in countries with a less favourable socio-economic situation, such as Croatia and Serbia, where insufficient organizational capacity and funding are available for a standard, top-down development. We suggest that such countries can adopt a bottom-up approach in which even minor, personal, or topically marginal projects are coordinated within the emerging community. Furthermore, such bottom-up environments can benefit from coordination with other similar environments, in our case in Croatia or Serbia. We further propose that bottom-up approaches can profit from coordination with top-down environments in neighbouring and/or culturally close countries, Slovenia in our case, with both sides experiencing a positive impact. We illustrate the synergistic effect of these different types of collaboration and coordination on the examples of textual data harvesting, manual data annotation, language tool development, and general infrastructure building. We wrap up with the most recent development – a CLARIN knowledge centre for South Slavic languages, where the collaborative methodology is expanded to all South Slavic languages. We close the chapter with a set of suggestions and good practices for researchers and language communities in a similar position to the ones discussed in this chapter.

Keywords: South-Slavic languages, collaborative LT infrastructure development, CLARIN Knowledge Centre

*__Corresponding author: Nikola Ljubešić,__ Dept. of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia, e-mail: nikola.ljubesic@ijs.si
Tomaž Erjavec, Dept. of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia, e-mail: tomaz.erjavec@ijs.si
Maja Miličević Petrović, Dept. of Interpreting and Translation, University of Bologna, Bologna, Italy, e-mail: maja.milicevic2@unibo.it
Tanja Samardžić, University of Zürich, Zürich, Switzerland, e-mail: tanja.samardzic@uzh.ch

https://doi.org/10.1515/9783110767377-017

1 Introduction

Slovenian, Croatian, and Serbian have a lot in common. They are not only linguistically closely related, but also share a complex history in and out of former Yugoslavia. However, the countries in which they are currently spoken differ substantially in the approach to infrastructure building for language technologies: while an open infrastructure is continuously being developed in Slovenia, such an infrastructure is for the most part still missing in Croatia[1] and Serbia. In this chapter, we describe the efforts of a group of researchers to start a collaboration on language technology infrastructure building for Croatian and Serbian from 2012 onward. We also recount the collaboration between these bootstrapping efforts and the well-developed Slovenian infrastructure CLARIN.SI (founded in 2013, part of CLARIN ERIC since 2015), which has yielded an added value for all three parties involved.

To capture the cross-country differences, we propose a distinction between *top-down* and *bottom-up* infrastructure building approaches. We consider any approach to scientific infrastructure building that is based on strategic national documents and well funded to be top-down. Where there is a lack of an overall strategy and of the necessary funding (which mostly go hand in hand), we refer to the efforts to bootstrap at least parts of an infrastructure as bottom-up. Given that infrastructure building is a complex and non-monolithic process, our position is that no single case can be strictly defined as top-down or bottom-up, but that most infrastructure building processes can be considered to predominantly belong to one or the other type. In the case of infrastructure building for Slovenian, Croatian, and Serbian, we consider Slovenian to mostly follow the top-down paradigm, while Croatian and Serbian predominantly rely on the bottom-up approach.

We also propose a preliminary explanation for why a country and a language take the top-down or the bottom-up approach, based on socio-economic factors such as GDP per capita[2] and R&D expenditure. While we do not claim that the same kind of explanation is appropriate for all contexts, we do believe that this is a suitable systematization of the course of infrastructure building taken in the three countries we are interested in.

1 Croatia became a member of CLARIN ERIC in 2018, but the infrastructure building process is still in an early phase.

2 There have already been attempts at explaining the level of technical maturity of a language through the GDP of its speakers, as was the case with the GLP (Gross Language Product) in (Hammarström 2009).

The remainder of this chapter is structured as follows. We first give a very brief introduction to what is the currently dominant paradigm in language technology development, namely machine learning. We continue with an outline of the linguistic, socio-economic, and technological context of the collaborations we discuss. We then move on to present two projects, dedicated respectively to Croatian and Serbian (ReLDI) and to Slovenian (JANES), whose separate and joint efforts led to major improvements in the quality and availability of language technologies for South Slavic languages. A CLARIN knowledge centre (CLASSLA) established as a follow-up initiative that also involves Bulgarian and Macedonian is subsequently described. We conclude the chapter with some practical remarks that can be taken as a set of guidelines for researchers working on resource-poor languages and/or in unsupportive environments.

A timeline visualization of the main projects described in this chapter is given in Figure 1.

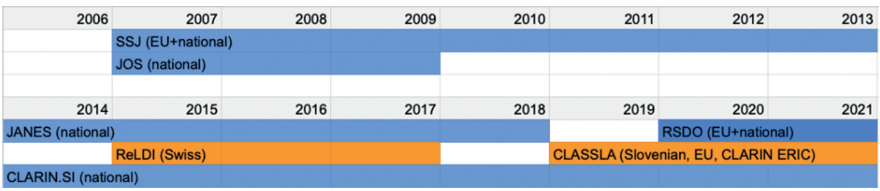

Figure 1: The timeline of the main projects described in this chapter. Blue indicates Slovenian (top-down) projects, while orange is used to mark bottom-up initiatives. The type of funding is given in parentheses.

The content of this chapter is related to (Hennelly et al. 2022), who discuss the development of digital language resources skills in South Africa, and also portray the historical development of language technologies in the area. The chapter by (Lindén et al. 2022) describes the collection of spoken data in Finland via an online platform, and is also related to this chapter in terms of identifying elegant technological solutions for collecting large quantities of language data, and taking into account the language variation present in an area.

2 Machine learning as the backbone of current language technologies

Language technologies can be simply defined as computer programs that can process language input into some desired (language) output (Tadić 2003). Some

examples of language technologies are machine translation systems (accepting input in one language and producing the output in another), speech-to-text systems (accepting recorded speech as input and producing textual output), text normalization (accepting user-generated textual content and producing standardized text on output), or hate speech identification (accepting a text as input and producing a label that indicates whether hate speech is present in the text).

Since the mid-1990s, the dominant paradigm in developing language technologies has been machine learning. This paradigm allows computers to solve language-related problems (machine translation, text normalization, hate speech detection, etc.) by learning from examples, i.e., from instances in which the task at hand has already been solved by humans. For text normalization such examples would be sentences of non-standard, user-generated text paired with a manually normalized version of the same sentences. For hate speech detection such data would be a set of texts complete with manually assigned labels that show if hate speech is present in the text or not. Such datasets are called "manually annotated" or "training" datasets, as they are used for training computer programs called language tools that automate the task initially performed manually by humans. Manually annotated datasets are one of the basic ingredients for developing modern language technologies, and, unlike the language tools themselves, they have to be developed separately for each language.

The production of manually annotated datasets is a costly and complex process if such data are not created as a side-product of a regular human activity, for example, translation of texts. For the most part, the process requires multiple steps: (1) a detailed definition of the problem in the form of annotation guidelines; (2) the training of human annotators; (3) the annotation itself; and (4) the resolution of annotation disagreements. To complicate matters further, this process is not linear, but rather iterative, for instance, disagreements between annotators mostly point to issues either in the annotator training or in the definition of the problem itself. Given the complex, labour-intensive, and costly set-up of annotation campaigns, the possibility of reducing the complexity and/or costs of annotation campaigns through joint efforts of multiple teams or projects is highly attractive, yet difficult to implement in practice.

An important feature of machine learning is the capacity to generalize from training data, which enables language tools to process previously unseen data. This feature is also very useful in settings where similar languages are to be processed. Specifically, language tools, trained on one language, are capable of processing another, similar language, where the quality of this processing depends, inter alia, on the language similarity, the processing task, and the amount and quality of the training data.

3 Setting the scene: The case of South Slavic

In order to provide some background on the synergistic potential of collaborative development of language technology infrastructures for South Slavic languages, we first briefly introduce the language group and the three languages for which the synergy has been exploited the most – Slovenian, Croatian, and Serbian. Next, we outline a basic socio-economic context and describe the language technology infrastructure developments in the last decade in the countries where the three languages are spoken.

3.1 The South Slavic language group

One of three branches of the Slavic languages (along with East and West Slavic), the South Slavic language group is itself divided in branches: a western branch that comprises Slovenian, Croatian, Bosnian, Serbian, and Montenegrin, and an eastern branch composed of Macedonian and Bulgarian (Stanojčić and Popović 2008). Both are rather unique in terms of linguistic and sociolinguistic properties. The eastern branch is somewhat of a linguistic outlier among Slavic languages in general, having a definite article but no nominal cases or infinitive verb forms (Ivić 1985). The western branch is particularly well-known for the complex sociolinguistic situation surrounding the languages that used to be part of Serbo-Croatian, which underwent gradual separation and have been developing as separate standards from the 1990's onward.

Despite now being independent standard languages, Croatian, Bosnian, Serbian, and Montenegrin remain highly mutually intelligible, reflecting the fact that standard Serbo-Croatian was based on a single dialect (called Shtokavian, from the question word *što* 'what'). Such a high level of mutual intelligibility does not exist among any other pairs of standard South Slavic languages. However, when dialectal variation is taken into account, it is easily observed that the South Slavic group forms a continuum spanning from Slovenia at the north-west to Bulgaria at the south-east (see e.g., Ivić 1985). In fact, the Kajkavian dialect (from another version of 'what', *kaj*), spoken in densely populated north-western Croatia, is closer to standard Slovenian than to standard Croatian (Kapović 2017), while Torlak vernaculars spoken in eastern Serbia are closer to Macedonian and Bulgarian than to Serbian (Ivić 1985). The continuum is also reflected in alphabet choices, with Slovenian and Croatian using only the Latin script, Bosnian, Serbian, and Montenegrin both Latin and Cyrillic, and Macedonian and Bulgarian only the Cyrillic script.

The synergistic efforts described in this chapter were in part made possible by the dialectal continuum and the ensuing similarities between South Slavic languages. Our focus is on language technology developments for three languages of the western branch, namely Slovenian, Croatian, and Serbian. As outlined above, standard Slovenian is the most distinct of the three, while Croatian and Serbian are fully mutually intelligible (albeit with some phonetic, lexical, and morphosyntactic differences). Croatian and Serbian thus provide particularly rich opportunities for joint language technology developments, with technologies developed for one language often being applicable to the other, but Slovenian is sufficiently close to also take part in the collaborative efforts.

3.2 The state of infrastructure for language technologies in Slovenia, Croatia and Serbia

From 1918 to 1991, Slovenia, Croatia and Serbia were parts of the same country (initially the Kingdom of Serbs, Croats, and Slovenes, and then Yugoslavia) and their scientific development was to some extent coordinated, although differences in the socio-economic status were present throughout the whole period. For instance, in 1988 the GDP index (which averaged to 100 for Yugoslavia as a whole) was 198 for Slovenia, 125 for Croatia and 89 for Serbia and Montenegro (Stiperski and Lončar 2008). After the break-up of Yugoslavia, Croatia and Serbia were heavily affected by the Yugoslav wars, while for Slovenia this was the case to a much lesser extent. The conflicts of the 1990s deepened the economic divide even more. In 2005, years after the conflicts, the previously introduced GDP index in Slovenia amounted to 313, in Croatia it was 152, while in Serbia and Montenegro it was only 59 (Stiperski and Lončar 2008). A similar divide is still visible today, with the 2019 GDP per capita (in euros) being 21,260 in Slovenia, 13,480 in Croatia, and 5,890 in Serbia.[3] A similar divide is visible in the expenditure on research and development in 2018, with Slovenia spending 1.94% of its GDP on R&D, while the figure for Croatia is 0.97% and for Serbia 0.92%.

The differences in socio-economic factors also follow the level of Euro-Atlantic integration, with Slovenia being a member of the European Union since 2007, Croatia joining in 2013, and Serbia being at present a candidate state. This kind of integration has been particularly important in terms of funding for research infrastructure developments.

3 https://ec.europa.eu/eurostat/databrowser/view/sdg_08_10/default/table?lang=en, data for 2021.

And indeed, the development of language technology infrastructure in the three countries roughly matches their overall socio-economic and political situation depicted above. In Slovenia, there have been continuous developments since the national project "Linguistic annotation of Slovene language: methods and resources"[4] (2007–2010) and the EU structural funds project "Communication in Slovene" (2007–2013), followed by Slovenia setting up its CLARIN.SI infrastructure in 2013, with a repository of language resources and tools. Currently ongoing is the project "Development of Slovenian in a digital environment" (2020–2022),[5] which is funded with EUR 4 million through the Slovenian Ministry of Culture and the European Regional Development Fund. These projects were both supported and followed by development of strategic documents and bodies on the national level, the most prominent being the Resolution on the National Programme for Language Policy (2013), the Action Plan for Language Infrastructure (2015), and the Council for Monitoring the Development of Language Resources and Technologies (2017).[6]

In Croatia and Serbia there have been very few top-down efforts and no wide-reaching national projects aimed at building a language technology infrastructure. Academic institutions and societies for language technologies (established in both countries) did participate in some relevant projects and language technology developments, but not comparable in magnitude to the ones in Slovenia. Croatia has in addition become a CLARIN ERIC member in 2018, but the infrastructure building is still in its early days. Moreover, the transfer potential between Croatian and Serbian, enabled by their great linguistic similarity, was not at all exploited and no joint projects were realized, with the exception of the MULTEXT-East project (Erjavec 2012) (1995–1997), which produced, inter alia, unrelated morphosyntactic specifications and resources for Croatian and Serbian. In fact, the lack of joint efforts in developing language technologies is a consequence of a complicated language history, with opposing and intertwined tendencies towards unification and diversification (Ljubešić, Miličević Petrović, and Samardžić 2018).

This is why a largely bottom-up approach had to be taken for both languages, with researchers personally dedicating themselves to develop basic language technologies, frequently within projects that were in fact focused on different, more specific topics. A good example is the development of the largest training dataset for basic processing of Croatian, which started as a personal side-pro-

4 http://nl.ijs.si/jos/index-en.html
5 https://slovenscina.eu/
6 http://www.efnil.org/projects/lle/slovenia/slovenia

ject, was improved through projects on unrelated topics of machine translation and text input assistant development, and finally received some focused attention in the ReLDI project that is described in more detail in Section 4.1. Such a development, although strenuous for the researchers involved, ensured that both Croatian and Serbian are today present in the Universal Dependencies project,[7] an open community effort with nearly 200 treebanks in over 100 languages with consistent syntactic annotation, and can be processed through the many annotation pipelines developed on the basis of these treebanks, such as Stanza (Qi et al. 2020), UDPipe (Straka and Straková 2017) or SpaCy.[8]

4 ReLDI, JANES, and CLARIN.SI: Moving forward together

In this section we present examples of bottom-up infrastructure development (the ReLDI project), examples of top-down developments (the JANES project), as well as the collaboration of bottom-up and top-down activities through collaboration of the ReLDI and the JANES project, with the support of the CLARIN.SI infrastructure.

4.1 Bottom-up infrastructures for Croatian and Serbian: The ReLDI project

The Swiss-funded institutional partnership Regional Linguistic Data Initiative – ReLDI[9] – had as one of its primary objectives the coordination of bottom-up infrastructure developments for Serbian and Croatian, two mutually intelligible languages with shared linguistic history, but with little prior history of joint language technology development.

We showcase the ReLDI project as a good example of bottom-up infrastructure development via international funding in a situation in which socio-economic reasons do not allow for top-down developments. We reiterate here why we consider the ReLDI project to be a bottom-up initiative in building language technology infrastructure for Croatian and Serbian: due to a lack of strategic

7 https://universaldependencies.org/
8 https://spacy.io
9 https://reldi.spur.uzh.ch

documents and national funding for infrastructure building in both countries, younger generation researchers aware of the need for a language technology infrastructure had to apply for international funding to start a collaborative cross-border process of infrastructure building for both languages. The partners in the project were the University of Zürich, the University of Belgrade, and the University of Zagreb.

4.1.1 How it all started

An initiative for a collaboration between younger generation researchers from Croatia and Serbia on joint development of language technologies for the two languages first occurred at the outskirts of the LREC 2012 conference in Istanbul, where they decided to apply for a bilateral Croatian-Serbian project that would provide them with some basic funding for meetings, and a formal framework for joint work. However, following major organizational issues, the call for projects was cancelled and the proposal was not even evaluated. The same researchers then applied to a call for the Swiss-funded SCOPES programme, aimed at strengthening scientific cooperation between Eastern Europe and Switzerland. The submitted ReLDI project proposal was positively evaluated, enabling researchers to start coordinating the development of language technologies, with substantial financial support for activities other than travelling and networking.

4.1.2 Early efforts in Croatia

Prior to these coordination efforts, bottom-up data collection projects were already underway in Croatia, in the form of building large web corpora. Since there was full awareness of the lack of open language technologies for Serbian as well, and given the simplicity of extending the collection process to highly similar languages, while building the second version of the Croatian web corpus, a web corpus of Serbian and Bosnian was also built, with minimal additional efforts (Ljubešić and Klubička 2014). Similarly, while crawling parallel data from the Southeast European Times website, which used to publish news in languages of South-Eastern Europe, parallel data in Serbian, Croatian, and Bosnian were collected. The Southeast European Times (SETimes) parallel corpus kick-started research on discriminating between similar languages (Tiedemann and Ljubešić 2012), as well as the VarDial evaluation campaigns on natural language processing for similar languages, dialects and varieties (Zampieri et al. 2014; Chakravarthi et al. 2021).

In parallel with these data collection efforts, basic open language technologies for the Croatian language, based on manually annotated data and machine learning algorithms, also started to emerge (Agić, Ljubešić, and Merkler 2013; Agić and Ljubešić 2015). This provided additional motivation for setting up a Croatian–Serbian collaboration and for transferring to Serbian the resource and tool development methodology, as well as the data themselves, given the relatedness of the two languages. The key dataset behind these first open language technologies for Croatian was based on a portion of the SETimes Croatian corpus, which was manually annotated for part-of-speech information, lemmas, syntactic dependencies, and named entities, resulting in the SETimes.HR dataset (Agić, Ljubešić, and Merkler 2013). The entire endeavour was a side-project with no dedicated funding, but it represented the turning point in the future development of language technologies for Croatian. The annotation of the dataset was performed by one annotator only and without quality assurance in the form of double annotations or annotation curation, primarily due to the very limited resources available. However, this set of limited activities did not just result in the first freely available tagger and lemmatizer for the Croatian language, but in similar tools for Serbian as well, as a Serbian test set, constructed along the SETimes.HR dataset, showed that Croatian models performed reasonably well on Serbian too (Agić, Ljubešić, and Merkler 2013).

4.1.3 Main activities and results

The ReLDI project focused primarily on two tasks: joint development of language technologies for Croatian and Serbian, and training sessions in using these technologies for linguistic research.

As part of the language technology building, the first freely available manually annotated dataset for Serbian, SETimes.SR, was constructed (Batanović, Ljubešić, and Samardžić 2018), an obvious result of know-how transfer from Croatian (the SETimes.HR dataset) to Serbian. In addition to transferring the know-how in manually annotated dataset development for basic linguistic processing, the already-developed language technologies for Croatian proved to be highly useful for pre-annotating Serbian data, which cut the production costs of the Serbian dataset significantly. Inside the ReLDI project, the SETimes.HR dataset was also expanded to the hr500k dataset (Ljubešić et al. 2016), more than five times the size of the original SETimes.HR dataset (taking its ssj500k Slovenian dataset equivalent (Krek et al. 2019) as motivation and an example of good practice). Both datasets were much more carefully annotated than their predecessor SETimes.HR, and improvements on these datasets have since been turned into an

ongoing process. Simultaneously, both languages were also added to the Universal Dependencies project[10] (Agić and Ljubešić 2015; Samardžić et al. 2017), which put Croatian and Serbian on the map of the modern language technology world.

Together with the development of manually annotated datasets for basic technologies, the recently finished inflectional lexicon of Croatian, hrLex (Ljubešić 2019a), built in a semi-automatic process (Ljubešić et al. 2015, 2016) inside the Abu-MaTran FP7 machine translation project, was used as a basis for building a comparable inflectional lexicon of Serbian, srLex (Ljubešić 2019b). With this coordinated effort, a 100,000-lexeme inflectional lexicon of Serbian was built for a fraction of the cost of building an inflectional lexicon of a highly-inflected language.

All the resources developed inside the ReLDI project were deposited in the Slovenian CLARIN.SI repository,[11] the nearest point that enabled high-quality long-term depositing of language resources for Croatian and Serbian.

4.2 Top-down infrastructure for Slovenian: The JANES project

The Slovenian national project JANES – *Jezikoslovna analiza nestandardne slovenščine* (Linguistic Analysis of Nonstandard Slovene) (Fišer, Ljubešić, and Erjavec 2020) had as one of its main goals the development of basic language technologies for Slovenian user-generated content. The project was run by the Faculty of Arts from the University of Ljubljana, and the Jožef Stefan Institute, also located in Ljubljana. This project was a logical continuation of top-down infrastructure building for the Slovenian language, given that basic language technologies for processing standard Slovenian had already been developed (Erjavec et al. 2010; Holdt, Kosem, and Berginc 2012), but were not fully suitable for user-generated online language. Previous research had shown that language technologies developed for standard language fail on non-standard variants, and that the most effective way forward is to build manually annotated datasets for non-standard variants that would enable an efficient adaptation of language technologies (Gimpel et al. 2011).

The three main outputs of the JANES projects were: (1) the JANES corpus, (2) the JANES manually annotated datasets, which were the basis for (3) the JANES toolchain, used for linguistically annotating the JANES corpus, the

10 https://universaldependencies.org
11 https://www.clarin.si/repository/xmlui/

most important resource to date for research into non-standard Slovenian. We describe these three components in the following subsections.

4.2.1 The JANES corpus

To produce the JANES corpus, three main sources were used: (1) Twitter (with a very good API for content harvesting); (2) web pages with a significant amount of user-generated content, i.e., newspapers with comments, blogs and fora; and (3) Wikipedia talk and discussion pages. The first two sources proved to be the richest in terms of non-standard features.

For the collection of data from Twitter, a simple dedicated tool was built, TweetCat (Ljubešić, Fišer, and Erjavec 2014), which enables continuous collection of tweets written in a low-density language. TweetCat requires only seed words (very frequent words specific of a language), to start the data collection process. Given the simplicity of extending the procedure to other languages, the decision was made to collect, in parallel with Slovenian tweets, Twitter posts in Croatian and Serbian. This was the starting point of a future collaboration and parallel infrastructure building for user-generated-content technologies for the two additional South Slavic languages described in Section 4.3.

As opposed to the Twitter collection procedure, scraping content from web pages proved to be highly site-dependent, as each web platform requires a specific tool to be built. What is more, the tool has a limited lifetime as any modifications in the web page layout break it. For that reason, harvesting of similar sources written in other languages was not even considered. Finally, while harvesting Wikipedia pages is simple, the analyses of the data showed them to be of limited informativeness for non-standard language features, so no harvesting of additional languages was performed.

4.2.2 The JANES manual data annotation

As discussed in Sections 2 and 4.2, to develop language technology tools that are able to process user-generated content, it was necessary to produce manually annotated datasets that would serve as their training data. The types of processing that were of most interest were (1) text standardness prediction, (2) text normalization, (3) part-of-speech and morphosyntactic tagging, (4) lemmatization, and (5) named entity recognition. A very basic example of a sentence with these annotation layers is given in Table 1.

Table 1: An example sentence of low orthographic and linguistical standardness, with manual token-level annotation of normalization, part-of-speech tagging, lemmatization, and named entity recognition.

Token	Normalized	Part-of-speech	Morphosyntax	Lemma	NER
ja	ja	PART	Q	ja	O
jst	jaz	PRON	Pp1-sn	jaz	O
sm	sem	AUX	Va-r1s-n	biti	O
poa	pa	CCONJ	Cc	pa	O
slisau	slišal	VERB	Vmbp-sm	slišati	O
da	da	SCONJ	Cs	da	O
je	je	AUX	Va-r3s-n	biti	O
CLARIN.SI	CLARIN.SI	PROPN	Npmsn	CLARIN.SI	B-ORG
top	top	ADJ	Agpmsnn	top	O
…	…	PUNCT	Z	…	O

The standardness level annotation was performed at the (short) text level (tweet, comment) and it indicated the degree of orthographic standardness (punctuation usage, character repetitions, etc.) and linguistic standardness (use of non-standard word forms). Identifying non-standard texts in an automatic manner was important for two reasons: (1) it was crucial that manually annotated datasets over-represent non-standard content, as this content is hard to process with standard technologies; and (2) having non-standardness information available in the whole JANES corpus enables researchers to focus on those parts of the corpora that contain non-standard features. Manually annotating and then automating the annotation of these two variables on the entire JANES corpus was crucial for the project given that, perhaps unexpectedly, most of user-generated content closely follows the norm.

The two main manually annotated datasets produced in the project were Janes-Norm (Erjavec et al. 2016) and Janes-Tag (Erjavec et al. 2019). In Janes-Norm (185,000 tokens in size), each word was manually assigned a standardized spelling. While the process of standardizing words might seem straightforward, it proved to be the most challenging of all the manual annotation campaigns in the project. This was mostly due to a large number of borderline cases (e.g., what is the normalized form of a word without a standard equivalent?), where problems had to be discovered first, a solution then agreed upon, and finally added to the annotation guidelines. Once the annotation guidelines were prepared, annotator training followed. The second dataset, Janes-Tag (Erjavec et al. 2019) (75,000 tokens), is a subset of Janes-Norm that was manually annotated at the levels of part-of-speech, lemma, and named entity.

Overall, these annotation campaigns were by far among the most complex to be performed by the research team, mostly due to a lack of standards for linguistic analysis of user-generated content. The opportunity to transfer the accumulated knowledge to other languages thus became very appealing.

4.2.3 The JANES toolchain

The tools developed inside the JANES project correspond for the most part to the levels of annotation described in the previous subsection. The first tool to be developed was the text standardness predictor which, given a text, returns two continuous values – one encoding orthographic standardness, the other linguistic standardness.

The remaining tools in the JANES toolchain consist of a text normalizer (Ljubešić et al. 2016),[12] part-of-speech tagger, lemmatizer, and named entity recognizer (Ljubešić, Erjavec, and Fišer 2017).[13] Given that all the developed tools were based on the machine learning paradigm, in order to adapt them for other languages, only manually annotated data in the specific languages were required, making the already considered possibility of constructing annotated datasets for other languages even more interesting.

All the three main deliverables of the JANES project were deposited and made available to the research community via the CLARIN.SI infrastructure.

The JANES project is a good example showing that almost any top-down infrastructure building activity carries a significant potential for extending the impact of that activity to other languages. While collecting data for the language of primary interest, data in related languages was collected as well, with minimal additional effort. During the manual annotation of a part of the collected data, to automate the annotation of the remaining data collection via machine learning, the significant potential for transfer of annotation guidelines and the annotation methodology to other languages was observed. Finally, a machine-learning-based toolchain was developed, which requires only the manually annotated data in the other languages to automate the annotation of these languages.

12 https://github.com/clarinsi/csmtiser
13 https://github.com/clarinsi/janes-tagger

4.3 Bottom-up *and* top-down: JANES + ReLDI = more than the sum

Thanks to the time overlap (as seen in Figure 1) and good personal relationships, ReLDI and JANES collaborated closely on extending the language technology infrastructure for user-generated-content processing from Slovenian to Croatian and Serbian. This is a great example of collaboration between a top-down language technology development environment (Slovenia), and two bottom-up environments (Croatia and Serbia), serving both sides involved. It is important to note that none of the developments described in this section would have been possible without the many preceding activities described in the previous sections.

4.3.1 How it all started

Unlike the unsuccessful application for a bilateral project between Croatia and Serbia, which produced the ReLDI partnership as a direct consequence, researchers from Slovenia and Serbia did receive a bilateral project grant with limited funding, primarily aimed at funding joint meetings. Given that this funding was obtained around the beginning of the JANES project, when the collection of Croatian and Serbian Twitter data via the TweetCat tool was already underway, and the initial manual annotation campaigns for text standardness in Slovenian have already been performed, the focus in the bilateral project was on producing Twitter datasets manually annotated with standardness level for Croatian and Serbian.

Aside from producing datasets manually annotated for standardness, developing training tools, and applying standardness labels over the full Twitter collections for Croatian and Serbian, this bilateral project also included a series of comparative studies on the three languages performed on the issue of standardness of user-generated content (Fišer et al. 2015; Miličević and Ljubešić 2016; Miličević, Ljubešić, and Fišer 2017). These studies were of great use in future activities on preparing training datasets for user-generated content processing in Croatian and Serbian. Specifically, they showed that, while the amount of non-standard elements in user-generated content was already low in Slovenian, in Croatian it was even lower, with non-standardness in Serbian being mostly encoded through lexical choices only, rather than through non-standard grammatical forms present in the two other languages.

4.3.2 JANES EXPRESS

The JANES project put a lot of emphasis on dissemination. One of the related activities in the project was the JANES EXPRESS series of lectures for students and fellow researchers in corpus and computational linguistics, which were organized in Ljubljana (Slovenia), Zagreb (Croatia) and Belgrade (Serbia). The lectures were organized in collaboration with the ReLDI project and they were meant to communicate the guidelines for the manual annotation of corpora for user-generated content processing, and to provide an introduction to the annotation platform WebAnno,[14] an offering of the CLARIN.SI infrastructure, used for annotating the Janes-Norm and Janes-Tag datasets. In addition to communicating with potential annotators and the interested public, the goal of the meetings was also to adapt the guidelines to the specificities of Croatian and Serbian, so more focused activities were performed with the annotators for both languages at the outskirts of the JANES EXPRESS events.

Once the JANES annotation procedure was communicated to Croatian and Serbian colleagues via JANES EXPRESS, and the annotation guidelines were (moderately) adapted, manual annotation on the Croatian and Serbian data excerpts, sampled in a manner comparable to Slovenian data for the Janes-Tag dataset, was performed as part of the ReLDI project activities. The CLARIN.SI infrastructure offered technological support for the WebAnno platform, and the JANES project offered advice on linguistic issues arising during the annotation process. The results of this collaboration are the ReLDI-NormTagNER-hr manually annotated dataset of non-standard Croatian (Ljubešić et al. 2019a), 89,000 tokens in size, and the ReLDI-NormTagNER-sr manually annotated dataset of non-standard Serbian (Ljubešić et al. 2019b), composed of 92,000 tokens.

Our rough estimate is that the time and energy invested in setting up annotation guidelines for the two additional languages was lowered to one fifth of the effort that was required for the original Slovenian dataset. In addition to the annotation guidelines being obtained for a minor fraction of the effort, the comparability of the annotation schemas was ensured, which is an important added value for the usage of the three datasets. The cost of the manual annotation itself were also moderately lower for Croatian and Serbian than was the case for the Slovenian dataset. In particular, during the Slovenian annotation campaign, pilot campaigns were necessary to test-run the annotation process and improve the annotation guidelines and the annotator training. No such pilots were necessary during the development of the Croatian and Serbian datasets.

14 https://webanno.github.io/webanno/

4.3.3 Joint technology development

The JANES project made a significant impact on the Croatian and Serbian language technology infrastructure by ensuring the production of manually annotated datasets of user-generated-content for a fraction of the overall price. In return, the ReLDI project developed a CRF-based tagger, named reldi-tagger,[15] which also included models for Slovenian (Ljubešić and Erjavec 2016). This tagger achieved not only new state-of-the-art results on part-of-speech tagging and lemmatization of Croatian and Serbian (Ljubešić et al. 2016), but also on Slovenian (Ljubešić and Erjavec 2016), regardless of the intensive language technology developments in Slovenia. The reldi-tagger tool was primarily built for processing of standard language, therefore an adaptation to the requirements of non-standard language was performed as part of the JANES project. The main modification was the usage of Brown clusters – a predecessor of the now omnipresent word embeddings. These activities resulted in the janes-tagger (Ljubešić, Erjavec, and Fišer 2017),[16] which was equipped not just with a model for tagging and lemmatizing non-standard Slovenian, but also non-standard Croatian and non-standard Serbian. This was possible primarily due to the comparable manually annotated datasets described above.

These developments show how the well-resourced, top-down infrastructure for Slovenian managed to profit from the two bottom-up infrastructures in the realm of technology development. Because of the collaboration between the JANES and the ReLDI teams, the top-down infrastructure obtained a new state-of-the-art tool for processing standard and non-standard language from the two bottom-up infrastructures. The two bottom-up infrastructures did not need to invest any additional resources to be of use to the Slovenian infrastructure because of (1) the relatedness of the three languages, and (2) the high capacity for technology reuse under the machine learning paradigm.

5 Scaling up and ensuring long-term impact: The CLASSLA knowledge centre

Given the successful collaboration between the JANES and the ReLDI projects and the CLARIN.SI infrastructure on building the language technology infrastructure

15 https://github.com/clarinsi/reldi-tagger
16 https://github.com/clarinsi/janes-tagger

for processing user-generated content for the three South Slavic languages, and also the success of the ReLDI project in coordinating the development of language technologies for the standard language, an idea emerged to institutionalize this paradigm for future collaborative language technology development.

Two possibilities were considered: (1) keeping the language focus on the Western South Slavic branch, i.e., continuing working on Slovenian, Croatian and Serbian (and, as much as resources allow, Bosnian and Montenegrin); or (2) expanding the collaboration to the Eastern South Slavic branch, namely Macedonian and Bulgarian. The decision was made to embrace the latter option, for the following reasons: (1) while Slovenian and Croatian use only the Latin alphabet, Serbian is a two-script language, being in that respect close to the eastern branch which uses the Cyrillic script only; (2) the Macedonian language has significant similarities to both Serbian and Bulgarian; (3) Macedonian is a heavily under-resourced language that would significantly benefit from such collaboration and, finally; (4) colleagues from the Bulgarian CLADA-BG infrastructure were enthusiastic about such a collaboration.

Following this idea, the CLARIN Knowledge Centre for South Slavic Languages (CLASSLA)[17] was born. The knowledge centre received official status in March 2019 and thereby became part of the CLARIN ERIC infrastructure. It is currently jointly led by the Slovenian CLARIN.SI and the Bulgarian CLADA-BG infrastructures. The main components of the knowledge centre are an e-mail-based helpdesk, frequently asked questions documents for all the mentioned languages, the CLARIN.SI concordancers (which are being expanded with various South Slavic corpora), and the CLARIN.SI repository, which already contains many resources and tools for various South Slavic languages. The main planned activities for the CLASSLA knowledge centre are – similar to the ReLDI project – to coordinate development of additional language technologies, but also to jointly build and serve a user base of the developing infrastructure.

As part of the CLARIN.SI infrastructure and the RSDO project (2020–2022), both aimed at enhancing the Slovenian language technology infrastructure, the CLASSLA linguistic processing pipeline[18] was produced. Its aim was to become the new state-of-the-art tool for basic linguistic processing, primarily of Slovenian, by exploiting the newer neural technologies (Ljubešić and Dobrovoljc 2019). Thanks to previous collaboration and the existence of comparable datasets for Croatian and Serbian, the CLASSLA pipeline covered both standard and non-standard Slovenian, Croatian, and Serbian from the very start.

17 https://www.clarin.si/info/k-centre/
18 https://pypi.org/project/classla/

As part of the collaboration inside the CLASSLA knowledge centre, the Bulgarian CLADA-BG infrastructure prepared the required data for training the CLASSLA pipeline for Bulgarian as well. The Bulgarian data enabled the training of a full stack of tools for the standard language. Manually annotated training data for user-generated content in Bulgarian are not yet available.

Another successful collaboration inside the CLASSLA knowledge centre was on the development of basic linguistic processing for the Macedonian language, namely tokenization, part-of-speech and morphosyntactic tagging, and lemmatization. For this to happen, a manually annotated dataset had to be produced, which was made possible through two developments: (1) a dataset of Macedonian suitable for training language technologies had been, starting with the MULTEXT-East project, continuously developed in a bottom-up approach for more than a decade, receiving recently a push from the CLASSLA knowledge centre to become usable for the CLASSLA pipeline; and (2) a large crawl of the Macedonian web was performed by the CLASSLA knowledge centre to enable the learning of good word embeddings (Ljubešić 2020), a crucial ingredient of any modern language technology. Thanks to these coordination efforts, the CLASSLA pipeline is now able to process Macedonian on a basic tokenizer-tagger-lemmatizer level, making it the go-to tool for the processing of Macedonian.[19]

Other successful collaborations of the CLASSLA knowledge centre concern the construction and publication of the first two corpora of the Montenegrin language, namely the English-Montenegrin parallel corpus (Božović et al. 2018)[20] and the Montenegrin web corpus (Ljubešić and Erjavec 2021),[21] as well as the preparation of Wikipedia corpora of all South Slavic languages, processed and presented in a uniform way, to be updated on a yearly basis (Ljubešić et al. 2021; Markoski et al. 2021).[22]

Many future collaborative activities are planned. One is the production of methodologically comparable monitor web corpora of all South Slavic languages, an activity planned inside the MaCoCu project,[23] focusing on enhancing machine translation for less-resourced languages. Another very timely development are open speech technologies and the significant impact CLASSLA would have if it managed to produce spoken corpora for South Slavic languages with available

19 The only tool previously freely downloadable for basic linguistic processing of Macedonian was BTagger (Aepli, von Waldenfels, and Samardžić 2014).

20 https://www.clarin.si/noske/run.cgi/corp_info?corpname=opusmonte_cnr&struct_attr_stats=1

21 https://www.clarin.si/noske/run.cgi/corp_info?corpname=mewac&struct_attr_stats=1

22 https://github.com/clarinsi/classla-wikipedia

23 https://macocu.eu

transcriptions. A strong source candidate for such a resource are parliamentary proceedings. Transcripts of parliamentary speeches in three South Slavic languages – Slovenian, Croatian and Bulgarian – have recently been processed with the CLASSLA pipeline with minimum overhead, inside the CLARIN ERIC-funded ParlaMint project (Erjavec et al. 2021).[24] A transcript-to-speech-aligned resource of just a few tens of hours could be all one needs to train basic speech-to-text systems[25] given the recent developments in pre-trained models for speech (Baevski et al. 2020). Having at least a basic speech-to-text system would start opening the ever-growing collection of spoken language recordings to researchers who nowadays still focus, mostly due to technical constraints and accessibility issues, primarily on written language.

6 An experiential "how-to" for other languages

In this section we share some insights and best practices for researchers and communities who find themselves in positions similar to those of the three languages we work on. The insights and advice focus on three topics: (1) general infrastructure building, (2) building language technology infrastructure, and (3) funding bottom-up infrastructure building.

6.1 General infrastructure building

Building an infrastructure top-down should not be considered "easy", as it requires the highest possible level of dedication by researchers, who need to push for the strategic documents to be drafted and accepted on the national level, for funding to be ensured, for projects to be successfully run, and so on. It is also crucial to consider in advance whether such top-down developments are feasible at all, and, depending on one's estimate, the choice between a top-down and a bottom-up road should be made as early as possible. For example, there are rather evident socio-economic and political reasons behind the lack of more top-down developments in open language technology infrastructures in Croatia and

24 https://www.clarin.eu/content/parlamint-towards-comparable-parliamentary-corpora

25 While finalizing this chapter, we have released such a system for Croatian (https://huggingface.co/classla/wav2vec2-xls-r-parlaspeech-hr), with the release of a two-thousand-hour ASR training dataset pending. The released dataset will be the first openly available ASR dataset for Croatian or Serbian.

Serbia, and waiting for a top-down approach to happen would not have made much sense in these countries.

Building an infrastructure in a bottom-up fashion is a very laborious endeavour, with fewer results than is the case with the top-down approach, but this is the only option available in most countries of a lower socio-economic standing. We also wish to stress here that it is not only the funding that is required for an infrastructure to be built top-down, but an overall high research and public administration capacity as well, which tends to go hand in hand with finances.

If one needs to rely on bottom-up infrastructure building, the best remedy is to ensure continuous coordination of efforts between researchers ready to take on specific tasks, even if this coordination is established through less official channels. Individual researchers investing all their efforts in their own solutions are highly unlikely to produce any tangible results.[26]

Regardless of whether two bottom-up infrastructures start to coordinate their efforts, or a bottom-up and a top-down infrastructure work together, benefits are to be expected for both sides.

6.2 Language technology infrastructure building

While performing language data collection, APIs and data dumps should be considered first, as their harvesting tends to be much simpler than any sort of web scraping. Moreover, data collection projects based on APIs and dumps can often be easily expanded to additional domains or languages with almost no extra effort.

Today's language technologies are based on machine learning algorithms that require manually annotated datasets. Building good quality datasets of this type is a very costly and complex process. Once an annotation campaign focused on a specific language problem starts, it is highly advisable to set the annotation goal as wide as possible, covering – if possible – additional domains or languages. This is because a comparable annotation result on another domain or language can be achieved with a fraction of resources that would be required for a full annotation campaign on that domain or language.

The technologies based on machine learning do require manually annotated datasets, but not much more than that. This opens up the space for training

26 Coordination is a crucial ingredient for top-down infrastructure building as well. However, in top-down environments coordination tends to be present from the beginning and tends to be a key ingredient behind the very existence of a top-down infrastructure.

developed technologies for multiple languages if comparable training data are available.

Technology is becoming ever more available, and our advice is that most energy, especially for bottom-up infrastructure building, should be invested in the production of manually annotated datasets. Once high-quality manually annotated data are available, it is quite easy to train different tools on that data. On top of that, machine-learning-based technology is nowadays developing at an unprecedented pace and the best bet for making an infrastructure future-proof is to invest in high-quality manually-annotated data. While we are still improving, and heavily using the CLASSLA pipeline, it is obvious that BERT-like pre-trained language models will be production-ready in the very near future. The code base for this new paradigm will be developed by large infrastructures and companies, and smart small infrastructures, especially the bottom-up ones, will be waiting in the wings with high-quality training data, ready for when the technology ripens and is easily offered to the users of the infrastructure.

6.3 Funding bottom-up infrastructure building

On the question of funding and running infrastructures, the situation is rather simple for the infrastructures being built top-down – these mostly receive significant national funding and have the necessary organizational capacity. For those infrastructures that have to be built bottom-up, we suggest the following.

International funding is much preferred, as national funding can be very hard to obtain, which is likely one of the reasons for the specific infrastructure not being built top-down in the first place. The Croatian and Serbian bottom-up infrastructure efforts were mostly supported by international funding.

Collaboration with other infrastructures-to-be that are in a similar bottom-up situation is highly advisable on the financial level as well. Any funding is much more likely to be obtained with joint forces. The good example are Croatian and Serbian joint efforts in obtaining funding.

There is no such thing as bad or too little funding. Work on the Croatian and Serbian user-generated content infrastructure was started on a project that only received a few thousand euros in funding.

It is worth coordinating efforts with top-down infrastructures as well, as this type of coordination effort might bring you by far the most return. Do not feel like you are exploiting someone: the other side will benefit from the collaboration as well, just as the Slovenian infrastructure has benefited from working on Croatian and Serbian.

Most activities need to be performed in an iterative manner. This is often the case even for top-down approaches, and when the funding is lacking, and conditions are far from optimal, the probability of obtaining a major single-run result is rather low. The Croatian standard-language training dataset came to its current size through three expansions and many more quality improvement iterations, another one being performed as we write.

Performing linguistic research alongside infrastructure building activities will inform these activities enormously. Research infrastructure building is full of pitfalls and identifying them early on is crucial. In our case, the user-generated-content infrastructure building profited enormously from the early observation that most data in this type of content is actually standard. This seemingly simple observation significantly changed the direction of the infrastructure development for all three languages.

7 Conclusion

This chapter has described the rather different roads to language technology development taken by three South Slavic languages. While the development in Slovenia has been predominantly top-down, relying on strategic documents and targeted funding, the developments in Croatia and Serbia have been rather bottom-up and most results have been achieved via smaller projects not formally embedded in any wider-scope strategy of infrastructure building. We have also shown the benefits of two types of collaboration between infrastructures(-to-be). The first type is between two bottom-up initiatives, for Croatian and Serbian, that was mostly driven by international funding, breaking the vicious circle related to the lack of national strategy, political will and funding for infrastructure building. The second type of collaboration, between top-down and bottom-up infrastructures, was illustrated on the collaboration between JANES and ReLDI projects. These two types of collaboration, together with CLARIN.SI as an overarching institutional framework, resulted in a crucial aggregation of resources and competences, which can now be streamed towards efficient future joint developments. The direction for scaling up these future developments is set by the recent establishment of the CLASSLA CLARIN knowledge centre.

We hope that this contribution will motivate further research in infrastructure development methodology in general, and especially on the coordination of infrastructure developments for related languages. We hope even more that it will enhance the practice of coordinating infrastructure developments, especially in the case of communities and languages that lack the socio-economic

support necessary for the development of a top-down language technology infrastructure. The types of coordination that we have described in this chapter are, in our opinion, the best chance communities and languages have to kick-start an infrastructure development and ensure the functioning of a language in the digital age.

Acknowledgment: The results presented in this work have been funded by the Swiss National Science Foundation grant IZ74Z0_160501 (ReLDI), the European Union Seventh Framework Programme FP7/2007–2013 under grant agreement PIAP-GA-2012-324414 (Abu-MaTran), the Slovenian Research Agency within the national basic research project "Resources, Tools and Methods for the Research of Nonstandard Internet Slovene" (J6-6842, 2014–2017), and the Slovenian research infrastructure CLARIN.SI.

The authors also acknowledge the financial support from the Slovenian Research Agency (research core funding No. P6-0411 Language resources and technologies for Slovene language), from the European Union's Rights, Equality and Citizenship Programme (2014–2020) project IMSyPP (Innovative Monitoring Systems and Prevention Policies of Online Hate Speech, grant no. 875263), the Slovenian Research Agency and the Flemish Research Foundation bilateral research project LiLaH (The linguistic landscape of hate speech on social media, grant no. ARRS-N6-0099 and FWO-G070619N), and the Slovenian Research Agency and the Serbian Ministry of Education bilateral project "The Construction of Corpora and Lexica of Non-standard Serbian and Slovenian" (BI-RS/14-15-068).

Bibliography

Aepli, Noëmi, Ruprecht von Waldenfels & Tanja Samardžić. 2014. Part-of-speech tag disambiguation by cross-linguistic majority vote. *Proceedings of the first workshop on applying NLP tools to similar languages, varieties and dialects*, 76–84. Dublin, Ireland: Association for Computational Linguistics and Dublin City University.

Agić, Željko & Nikola Ljubešić. 2015. Universal Dependencies for Croatian (that work for Serbian, too). *The 5th workshop on Balto-Slavic natural language processing*, 1–8. Hissar, Bulgaria: INCOMA Ltd. Shoumen, BULGARIA.

Agić, Željko, Nikola Ljubešić & Danijela Merkler. 2013. Lemmatization and morphosyntactic tagging of Croatian and Serbian. *Proceedings of the 4th biennial international workshop on Balto-Slavic natural language processing*, 48–57. Sofia, Bulgaria: Association for Computational Linguistics.

Baevski, Alexei, Henry Zhou, Abdelrahman Mohamed & Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*.

Batanović, Vuk, Nikola Ljubešić & Tanja Samardžić. 2018. SETimes.SR – a reference training corpus of Serbian. *Proceedings of the Conference on Language Technologies & Digital Humanities 2018 (JT-DH 2018)*, 11–17. Ljubljana: Ljubljana University Press.

Božović, Petar, Tomaž Erjavec, Jörg Tiedemann, Nikola Ljubešić, Vojko Gorjanc et al. 2018. Opus-MontenegrinSubs 1.0: First electronic corpus of the Montenegrin language. *Proceedings of the Conference on Language Technologies & Digital Humanities 2018* (JT-DH 2018), 24–28. Ljubljana: Ljubljana University Press.

Chakravarthi, Bharathi Raja, Mihaela Gaman, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Ruba Priyadharshini, Christoph Purschke et al. 2021. Findings of the VarDial Evaluation Campaign 2021. *Proceedings of the 8th VarDial Workshop on NLP for Similar Languages, Varieties and Dialects*, 1–11. Kiyv, Ukraine: Association for Computational Linguistics.

Erjavec, Tomaž. 2012. MULTEXT-East: morphosyntactic resources for Central and Eastern European languages. *Language Resources and Evaluation* 46 (1): 131–142.

Erjavec, Tomaž, Darja Fišer, Simon Krek & Nina Ledinek. 2010. The JOS Linguistically Tagged Corpus of Slovene. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA).

Erjavec, Tomaž, Darja Fišer, Jaka Čibej & Špela Arhar Holdt. 2016. CMC training corpus Janes-Norm 1.2. Slovenian language resource repository CLARIN.SI.

Erjavec, Tomaž, Darja Fišer, Jaka Čibej, Špela Arhar Holdt, Nikola Ljubešić, Katja Zupan & Kaja Dobrovoljc. 2019. CMC training corpus Janes-Tag 2.1. Slovenian language resource repository CLARIN.SI.

Erjavec, Tomaž, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Vladislava Grigorova, Michał Rudolf, Andrej Pančur, Matyáš Kopp, Starkaður Barkarson, Steinhór Steingrímsson, Henk van der Pol, Griet Depoorter, Jesse de Does, Bart Jongejan, Dorte Haltrup Hansen, Costanza Navarretta, María Calzada Pérez, Luciana D. de Macedo, Ruben van Heusden & M. 2021. Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 2.0. Slovenian language resource repository CLARIN.SI.

Fišer, Darja, Tomaž Erjavec, Nikola Ljubešić & Maja Miličević. 2015. Comparing the nonstandard language of Slovene, Croatian and Serbian tweets. *Simpozij Obdobja* 34: 225–231.

Fišer, Darja, Nikola Ljubešić & Tomaž Erjavec. 2020. The Janes project: language resources and tools for Slovene user generated content. *Language resources and evaluation* 54 (1): 223–246.

Gimpel, Kevin, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan & Noah A. Smith. 2011. Part-of-speech tagging for Twitter: Annotation, features, and experiments. *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, 42–47. Portland, Oregon, USA: Association for Computational Linguistics.

Hammarström, Harald. 2009. Unsupervised Learning of Morphology and the Languages of the World. PhD dissertation, University of Gothenburg.

Hennelly, Martin, Langa Khumalo, Juan Steyn & Menno van Zaanen. 2022. Training of digital language resources skills in South Africa. In Darja Fišer & Andreas Witt (eds.), *CLARIN. The infrastructure for language resources.* Berlin: De Gruyter.

Holdt, Špela Arhar, Iztok Kosem & Nataša Logar Berginc. 2012. Izdelava korpusa Gigafida in njegovega spletnega vmesnika. *Proceedings of Eighth Language Technologies Conference IS-LTC*, Volume 12.

Ivić, Pavle. 1985. *Dijalektologija srpskohrvatskog jezika. Uvod i štokavsko narečje*. 2nd edition. Novi Sad: Matica srpska.

Kapović, Mate. 2017. The position of kajkavian in the South Slavic dialect continuum in light of old accentual isoglosses. *Zeitschrift für Slawistik* 62 (4): 606–620.

Krek, Simon, Kaja Dobrovoljc, Tomaž Erjavec, Sara Može, Nina Ledinek, Nanika Holz, Katja Zupan, Polona Gantar, Taja Kuzman, Jaka Čibej, Špela Arhar Holdt, Teja Kavčič, Iza Škrjanec, Dafne Marko, Lucija Jezeršek & Anja Zajc. 2019. Training corpus ssj500k 2.2. Slovenian language resource repository CLARIN.SI.

Lindén, Krister, Tommi Jauhiainen, Mietta Lennes, Mikko Kurimo, Aleksi Rossi, Tommi Kurki & Olli Pitkänen. 2022. Donate Speech: Collecting and sharing a large-scale speech database for Social Sciences, Humanities and Artificial Intelligence research and innovation. In Darja Fišer & Andreas Witt (eds.), *CLARIN. The infrastructure for language resources*. Berlin: De Gruyter.

Ljubešić, Nikola. 2019a. Inflectional lexicon hrLex 1.3. Slovenian language resource repository CLARIN.SI.

Ljubešić, Nikola. 2019b. Inflectional lexicon srLex 1.3. Slovenian language resource repository CLARIN.SI.

Ljubešić, Nikola. 2020. Word embeddings CLARIN.SI-embed.mk 0.1. Slovenian language resource repository CLARIN.SI.

Ljubešić, Nikola & Kaja Dobrovoljc. 2019. What does neural bring? Analysing improvements in morphosyntactic annotation and lemmatisation of Slovenian, Croatian and Serbian. *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, 29–34. Florence, Italy: Association for Computational Linguistics.

Ljubešić, Nikola & Tomaž Erjavec. 2016. Corpus vs. lexicon supervision in morphosyntactic tagging: the case of Slovene. *Proceedings of the tenth international conference on language resources and evaluation (LREC'16)*, 1527–1531. Portorož, Slovenia: European Language Resources Association (ELRA).

Ljubešić, Nikola & Tomaž Erjavec. 2021. Montenegrin web corpus meWaC 1.0. Slovenian language resource repository CLARIN.SI.

Ljubešić, Nikola, Tomaž Erjavec, Vuk Batanović, Maja Miličević & Tanja Samardžić. 2019a. Croatian Twitter training corpus ReLDI-NormTagNER-hr 2.1. Slovenian language resource repository CLARIN.SI.

Ljubešić, Nikola, Tomaž Erjavec, Vuk Batanović, Maja Miličević & Tanja Samardžić. 2019b. Serbian Twitter training corpus ReLDI-NormTagNER-sr 2.1. Slovenian language resource repository CLARIN.SI.

Ljubešić, Nikola, Tomaž Erjavec & Darja Fišer. 2017. 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*, 60–68. Valencia, Spain: Association for Computational Linguistics.

Ljubešić, Nikola, Miquel Espla-Gomis, Filip Klubička & Nives Mikelić Preradović. 2015. Predicting inflectional paradigms and lemmata of unknown words for semi-automatic expansion of morphological lexicons. *Proceedings of the International Conference Recent Advances in Natural Language Processing*, 379–387. Hissar, Bulgaria: INCOMA Ltd. Shoumen, BULGARIA.

Ljubešić, Nikola, Darja Fišer & Tomaž Erjavec. 2014. TweetCaT: a tool for building Twitter corpora of smaller languages. *Proceedings of LREC*, 2279–2283. Reykjavik, Iceland: European Language Resources Association (ELRA).

Ljubešić, Nikola & Filip Klubička. 2014. {bs, hr, sr} WaC – Web corpora of Bosnian, Croatian and Serbian. *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, 29–35. Gothenburg, Sweden: Association for Computational Linguistics.

Ljubešić, Nikola, Filip Klubička, Željko Agić & Ivo-Pavao Jazbec. 2016. New inflectional lexicons and training corpora for improved morphosyntactic annotation of Croatian and Serbian. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 4264–4270. Portorož, Slovenia: European Language Resources Association (ELRA).

Ljubešić, Nikola, Filip Markoski, Elena Markoska & Tomaž Erjavec. 2021. Comparable corpora of South-Slavic Wikipedias CLASSLA-Wikipedia 1.0. Slovenian language resource repository CLARIN.SI.

Ljubešić, Nikola, Katja Zupan, Darja Fišer & Tomaž Erjavec. 2016. Normalising Slovene data: historical texts vs. user-generated content. *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, Volume 16, 146–155.

Ljubešić, Nikola, Maja Miličević Petrović & Tanja Samardžić. 2018. Borders and boundaries in Bosnian, Croatian, Montenegrin and Serbian: Twitter data to the rescue. *Journal of Linguistic Geography* 6 (2): 100–124.

Markoski, Filip, Elena Markoska, Nikola Ljubešić, Eftim Zdravevski & Ljupco Kocarev. 2021. Cultural topic modelling over novel Wikipedia corpora for South-Slavic languages. *Proceedings of the international conference on recent advances in natural language processing (ranlp 2021)*, 910–917. Held Online: INCOMA Ltd.

Miličević, Maja & Nikola Ljubešić. 2016. 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): 156–188.

Miličević, Maja, Nikola Ljubešić & Darja Fišer. 2017. Birds of a feather don't quite tweet together: An analysis of spelling variation in Slovene, Croatian and Serbian Twitterese. In Darja Fišer & Michael Beißwenger (eds.), *Investigating Computer-Mediated Communication: Corpus-based Approaches to Language in the Digital World*, 14–43. Ljubljana: Ljubljana University Press.

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton & Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Samardžić, Tanja, Mirjana Starović, Željko Agić & Nikola Ljubešić. 2017. Universal Dependencies for Serbian in comparison with Croatian and other Slavic languages. *Proceedings of the 6th workshop on Balto-Slavic natural language processing*, 39–44. Valencia, Spain: Association for Computational Linguistics.

Stanojčić, Živojin & Ljubomir Popović. 2008. *Gramatika srpskog jezika za gimnazije i srednje škole*. 11th. Belgrade: Zavod za udžbenike i nastavna sredstva.

Stiperski, Zoran & Jelena Lončar. 2008. Changes in levels of economic development among the states formed in the area of former Yugoslavia. *Hrvatski geografski glasnik* 70 (2): 5–32.

Straka, Milan & Jana Straková. 2017. Tokenizing, PoS tagging, lemmatizing and parsing UD 2.0 with UDPipe. *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 88–99. Vancouver, Canada: Association for Computational Linguistics.

Tadić, M. 2003. *Jezične tehnologije i hrvatski jezik*. Zagreb: Ex libris.

Tiedemann, Jörg & Nikola Ljubešić. 2012. Efficient discrimination between closely related languages. *Proceedings of COLING 2012*, 2619–2634. Mumbai, India: The COLING 2012 Organizing Committee.

Zampieri, Marcos, Liling Tan, Nikola Ljubešić & Jörg Tiedemann. 2014. A report on the DSL shared task 2014. *Proceedings of the first workshop on applying NLP tools to similar languages, varieties and dialects*, 58–67. Dublin, Ireland: Association for Computational Linguistics and Dublin City University.