CLARIN

Common Language Resources and Technology Infrastructure

Tour de CLARIN
SLOVENIA

Edited by **Darja Fišer** and **Jakob Lenardič**

SLOVENIA

Foreword

Tour de CLARIN highlights prominent user involvement activities of CLARIN National Consortia and Knowledge Centres with the aim to increase their visibility, reveal the richness of the CLARIN landscape, and display the full range of activities throughout the CLARIN network that can inform and inspire other consortia and knowledge centres as well as show what CLARIN has to offer to researchers, teachers, students, professionals and the general public interested in using and processing language data in various forms.

The brochure presents Slovenia and is organized in five sections:

- Section One presents the members of the consortium and their work
- Section Two demonstrates an outstanding tool
- Section Three highlights a prominent resource
- Section Four reports a successful event for researchers and students
- Section Five includes an interview with a renowned researcher from the digital humanities or social sciences who has successfully used the consortium's infrastructure in their research

SLOVENIA

Introduction

Written by **Tomaž Erjavec, Jakob Lenardič** and **Nikola Ljubešić**

CLARIN.SI joined CLARIN ERIC in 2015.[1] The seat of CLARIN.SI is the Jožef Stefan Institute, the main Slovenian research organization for applied research in natural sciences and technology, where three units are involved in overseeing its operation: the Department of Knowledge Technologies, the Laboratory for Artificial Intelligence, and the Centre for Networking Infrastructure. CLARIN.SI is organized as a consortium, bringing together partners from all the main organizations that produce or use language resources in Slovenia, in particular the four Slovenian universities (University of Ljubljana, University of Maribor, University of Nova Gorica and University of Primorska), three research institutes (Scientific and Research Centre of the Slovenian Academy of Sciences and Arts, Jožef Stefan Institute, Institute of Contemporary History), three societies (Slovenian Language Technologies Society, Trojina Institute for Applied Slovene Studies, Domestic Research Society), and two HLT companies (Alpineon and Amebis). The national coordinator of CLARIN.SI is Tomaž Erjavec.

CLARIN.SI has very good relations with similar research infrastructures in Slovenia, in particular DARIAH-SI, and ADP, the Slovenian CESSDA node, which are realized through joint work on specific projects, such as the recent ParlaFormat workshop, and partnership in the recently started RDA-Slovenia project.

[1] http://www.clarin.si/

CLARIN.SI is a B-certified centre which offers a LINDAT/D-Space repository that currently contains around 110 language resources for Slovenian as well as for other languages, especially Croatian and Serbian. The repository offers a wide range of large corpora for linguistic research on Slovenian, as well as parallel and manually annotated corpora and lexica for training language tools. Most of the corpora in the repository can also be accessed via two concordancers, KonText and noSketch Engine, both of which are integrated with the repository and serve as versatile online environments for searching and efficiently analysing large and richly annotated corpora. In addition to resources, the centre offers tools for text processing as well, either as open source on GitHub, or as on-line services, such as ReLDIanno, an online tool and web service for, currently, annotating texts in Slovenian, Croatian, and Serbian.

The consortium regularly supports data curation projects, mostly in terms of annotation campaigns, or to prepare existing digital data for inclusion into the repository. A good example of in-kind support is the Kontext.io semantic lexicon of Slovene, Croatian and Serbian, where CLARIN.SI prepared the union of its public corpora, in order to train the word embeddings that are the basis of the lexicon, as well as providing examples of use on the portal. In 2018, support for ad-hoc projects was supplemented by a project call to the consortium partners, through which seven projects were selected for financing. All the projects produced openly available resources or tools for Slovenian, and the call has been repeated in 2019, with six projects accepted for funding.

A major priority of the Slovenian consortium is its outreach activities, many of which have an international scope. The Slovenian Language Technologies Society has been organizing biennial conferences on Language Technologies with online reviewed proceedings for over 20 years. In 2016 the scope of the conference was extended to Digital Humanities, and CLARIN.SI became one of the organizers and supporters of the conference. The 11th edition of the Language Technologies & Digital Humanities conference, which took place in September 2018 in Ljubljana, heard presentations of 47 papers (21 papers in Slovene and 26 in English), including two talks by invited lecturers. The Society also organizes, and has done for almost 15 years, regular JOTA lectures on language technologies; since 2017 CLARIN.SI has supported the recording of these lectures, which are available, together with video-synchronized slides on the VideoLectures portal. CLARIN.SI supports other events that take place in Slovenia and are related to the mission of CLARIN, e.g. in 2018 CLARIN.SI this was the XVIII EURALEX International Congress, and in 2019 the 22nd International Conference “Text, Speech and Dialogue”.

Recently, CLARIN.SI has established the Knowledge Centre for South Slavic languages (CLASSLA). CLASSLA offers expertise on language resources and technologies with its basic activities being (1) giving researchers, students, citizen scientists and other interested parties information on the available resources and technologies via its documentation, (2) supporting them in producing, modifying or publishing resources and technologies via its help desk and (3) organizing training activities. The K-Centre also offers a FAQ for Slovene, Croatian and Serbian and documentation on how to use ReLDIanno CLARIN.SI web services.

The vibrant CLARIN.SI community gathered at the 11th Slovenian Language Technologies and Digital Humanities Conference in 2018

Tool | **CSMTiser**

Written by **Nikola Ljubešić** and **Tomaž Erjavec**

A well-known problem with using text annotation tools that have been trained on datasets of standard language for texts written in non-standard language, such as dialects, historical varieties, or user-generated content, is that the results are drastically decreased. A common approach to overcome this problem is to first normalize (i.e., modernize or standardize) the non-standard text and only then proceed with further processing. As an additional benefit, normalization of non-standard texts also simplifies searching in such text collections.

CSMTiser, available on the CLARIN.SI GitHub site,[2] is a supervised machine learning tool that performs word normalization by using Character-level Statistical Machine Translation. The tool is a wrapper around the well-known Moses SMT package, which enables non-computer-scientists to train and run a text normalizer by editing the configuration file, running a script for training the normalizer, and then another one for applying it.

The tool has been very efficient in modernizing historical Slovene (Scherrer and Erjavec, 2016) and Slovene user-generated content (Ljubešić et al. 2016). It has also been successfully applied to normalize Swiss dialects to a common denominator (Scherrer and Ljubešić, 2016) and to modernize historical Dutch for the purposes of further processing (Tjong Kim Sang, 2017) within a shared task in which the CSMTiser ranked first among eight teams, many of which applied neural approaches. The success of the CSMTiser shows the strength of a simple, yet powerful approach to text normalization. Even today, after significant improvements in the area due to deep learning, the neural approaches outperform the CSMTiser by only 1 to 2 accuracy points, which is low given a large increase in the complexity of processing (Lusetti et al. 2018, Ruzsics and Samardžić 2019).

18B	Al	ta	nar bòl	vashna	reſsniza	je	moja	lubęsen	prut	Nęshki.	
(1790)	*ali*	*ta*	*najbolj*	*važna*	*resnica*	*je*	*moja*	*ljubezen*	*proti*	*nežki*	
19A	poboshnim	ſerzam	in	veſtjo	pridnoſt	in	ljubesin	k	ſvojimu	ſtanu	sdrushi
(1843)	*pobožnim*	*srcem*	*in*	*vestjo*	*pridnost*	*in*	*ljubezen*	*k*	*svojemu*	*stanu*	*združi*
19B	Otroška	ljubezen	naj	zmír	te	navdaja	Za	starše,	za	brate,	Bogá	in	cesarja
(1872)	*otroška*	*ljubezen*	*naj*	*zmeraj*	*te*	*navdaja*	*za*	*starše,*	*za*	*brate,*	*boga*	*in*	*cesarja*

Figure 1: *Slovene text from three different periods. The column in bold shows the slice the text belongs to and, in brackets, its year of publication. Each example gives the original text in the first line and the modernized word tokens in the second line, to illustrate the kind of phenomena that must be handled in the modernization of words.*

[2] https://github.com/clarinsi/csmtiser

The importance of text normalization can clearly be seen through the improvements in downstream text processing on the basic task of part-of-speech tagging: while 18th century Slovene processed without normalization gives a PoS tagging accuracy of 58%, 93% is achieved on the normalized text. Less drastically but still very noticeably, PoS tagging user-generated content without prior normalization achieves an accuracy of 83%, while normalizing the text prior to tagging produces an accuracy of 89% (Zupan et al. 2019).

We expect that new tools and approaches will emerge that will outperform the CSMTiser both in terms of higher accuracy and lower complexity, which is why CLARIN.SI focuses on providing publicly available training datasets. For text normalization, the repository offers datasets for learning normalization of Slovene, Croatian and Serbian user-generated content, as well as datasets for normalizing historical Slovene in two distinct historical periods.

References:

Scherrer, Y. and Erjavec, T. 2016. Modernising historical Slovene words. *Natural Language Engineering* 22 (6): 881–905.

Ljubešić, N., Zupan, K., Fišer, D., and Erjavec, T. 2016. Normalizing Slovene data: historical texts vs. user-generated content. In *Proceedings of the 13th Conference on Natural Language Processing* (KONVENS), 146–155, September 19–21, 2016, Bochum, Germany.

Scherrer, Y. and Ljubešić, N. 2016. Automatic normalisation of the Swiss German ArchiMob corpus using character-level machine translation. In *Proceedings of the 13th Conference on Natural Language Processing* (KONVENS), 248–255, September 19–21, 2016, Bochum, Germany.

Kim Sang, T. et al. 2017. The CLIN27 shared task: Translating historical text to contemporary language for improving automatic linguistic annotation. *Computational Linguistics in the Netherlands Journal 7*: 53–64.

Lusetti, M., Ruzsics, T., Göhring, A., Samardžić, T., and Stark, E. 2018. Encoder-Decoder Methods for Text Normalization. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects* (VarDial 2018), 18–28.

Ruzsics, T., and Samardžić, T. 2019. Multilevel Text Normalization with Sequence-to-Sequence Networks and Multisource Learning.

Zupan, K., Ljubešić, N., and Erjavec, T. 2019. How to Tag Non-standard Language: Nominalization vs. Domain Adaptation. *Natural Language Engineering* 25 (5): 651–674.

Resource | **Emoji Sentiment Ranking 1.0**

Written by **Petra Kralj Novak, Jasmina Smailović, Borut Sluban** and **Igor Mozetič**

Emoji are Unicode graphic symbols, used as a shorthand to express concepts and ideas, and can play an important role in social media text analytics. In 2015, Petra Kralj Novak, Jasmina Smailović, Borut Sluban and Igor Mozetič from the Jožef Stefan Institute in Ljubljana, Slovenia released the first emoji sentiment lexicon, called Emoji Sentiment Ranking 1.0, and published it as a resource in the public language resource repository CLARIN.SI.[3] With 78,500 downloads to date, the lexicon is the most downloaded resource in the CLARIN.SI repository.

The sentiment of the emoji was computed from the sentiment of the tweets in which they occur based on the labelling of sentiment polarity (negative, neutral, or positive) of about 1.6 million tweets in 13 European languages by 83 human annotators. About 4% of the annotated tweets contained emoji. The sentiment score of each emoji was computed based on its estimated probability of appearing in a tweet with each sentiment.

The process and analysis of the Emoji Sentiment Ranking is described in detail by Kralj Novak et al. (2015). The authors draw a sentiment map of the 751 emoji (see Figure 2), formalize sentiment and present a novel intuitive visualization of sentiment distribution in the form of a sentiment bar (Figure 3). Furthermore, they compare the sentiment of tweets with and without emoji, and find that tweets with emoji tend to be more positive. They also found differences between the more and less frequent emoji: the more frequently used emoji tend to be more positive. Another interesting aspect is the position of emoji in tweets: the more sentimental charge an emoji has, the more likely it is to appear at the end of tweets (see Figure 4). An exception is the soccer ball emoji, which is commonly used to replace a word but has a very positive sentiment associated with it.

[3] http://hdl.handle.net/11356/1048

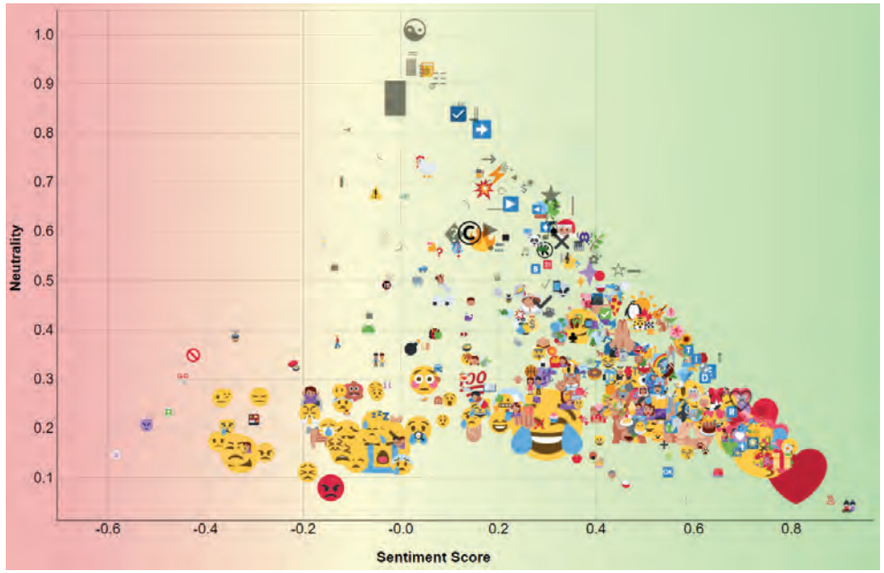


Figure 2: *Sentiment map of the 751 most frequently used emoji. The position of the emoji denotes its sentiment score and neutrality, while its size is proportional to the frequency of its usage. An interactive version is available here: http://kt.ijs.si/data/Emoji_sentiment_ranking/emojimap.html*

Image [twemoji]	Unicode codepoint	Occurrences [6...max]	Position [0...1]	Neg [0...1]	Neut [0...1]	Pos [0...1]	Sentiment score [-1...+1]	Sentiment bar (c.i. 95%)	Unicode name
😂	0x1f602	14622	0.805	0.247	0.285	0.468	0.221		FACE WITH TEARS OF JOY
❤	0x2764	8050	0.747	0.044	0.166	0.790	0.746		HEAVY BLACK HEART
♥	0x2665	7144	0.754	0.035	0.272	0.693	0.657		BLACK HEART SUIT
😍	0x1f60d	6359	0.765	0.052	0.219	0.729	0.678		SMILING FACE WITH HEART-SHAPED EYES
😭	0x1f62d	5526	0.803	0.436	0.220	0.343	-0.093		LOUDLY CRYING FACE
😘	0x1f618	3848	0.854	0.053	0.193	0.754	0.701		FACE THROWING A KISS
😊	0x1f60a	3186	0.813	0.060	0.237	0.704	0.644		SMILING FACE WITH SMILING EYES
👌	0x1f44c	2925	0.805	0.094	0.249	0.657	0.563		OK HAND SIGN
💕	0x1f495	2400	0.766	0.042	0.285	0.674	0.632		TWO HEARTS
👏	0x1f44f	2336	0.787	0.104	0.271	0.624	0.520		CLAPPING HANDS SIGN
😁	0x1f601	2189	0.796	0.127	0.296	0.577	0.449		GRINNING FACE WITH SMILING EYES
☺	0x263a	2062	0.799	0.062	0.218	0.720	0.657		WHITE SMILING FACE

Figure 3: *The sentiment distribution of each emoji is visualized in form of a sentiment bar. http://kt.ijs.si/data/Emoji_sentiment_ranking/index.html*

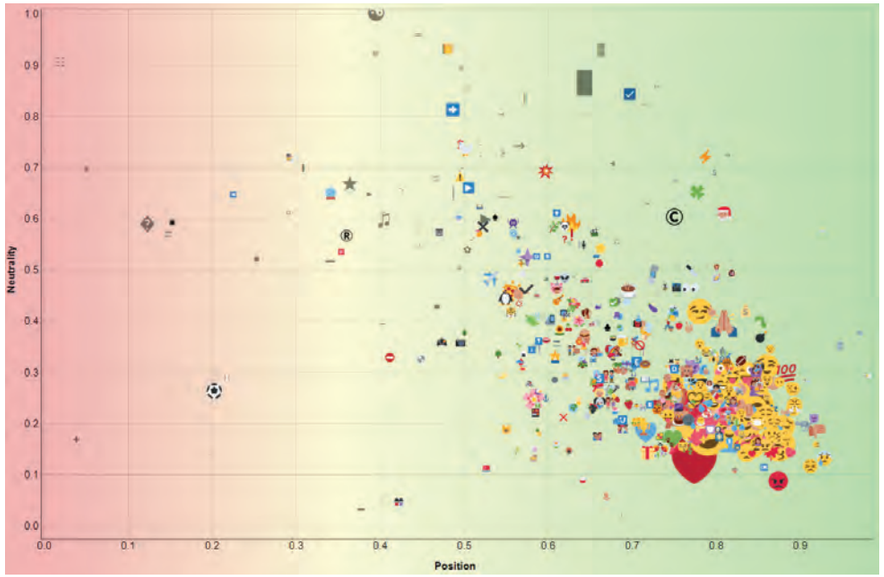


Figure 4: *Emoji position in tweets. The horizontal axis represents the length of a tweet. The vertical axis represents the neutrality of the emoji: top for very neutral and bottom for very emotional, either positive or negative. Emoji that act as word replacements, thus positioned in the middle of the tweets, tend to have a neutral sentiment. The emoji that act as sentiment conveyers are more likely positioned at the end of tweets.*

As a further analysis, the authors investigated whether the Emoji Sentiment Ranking can be considered as a universal language-independent resource, at least for European languages. They made independent rankings of emoji sentiment for each of the 13 languages and showed that there is no evidence of significant differences between emoji sentiment between the languages.

The information about the sentiment of emoji can be used in the automated sentiment classification of informal texts. A basic distinction between positive and negative emoji can be used to automatically label positive and negative samples of texts. These samples can then be used to train and test sentiment-classification models using machine learning techniques. Such emoji-labelled sets can be used to automatically train sentiment classifiers. Emoji can also be exploited to extend the more common features used in text mining, such as sentiment-carrying words.

Reference:

Kralj Novak, P., Smailović, J., Sluban, B., and Mozetič I. 2015. Sentiment of Emoji. PLoS ONE 10 (12): e0144296. doi:10.1371/journal.pone.0144296.

Event | **JANES Express**

Written by **Darja Fišer** and **Tomaž Erjavec**

The past decade has witnessed rapid growth of user-generated content, such as blogs, forums and social media. This type of content offers an important source of information to diverse fields, such as social sciences, economics and computer science, both for research and business. But when dealing with user-generated content it is necessary to come to grips with the language of computer-mediated communication which is, due to its social and technical characteristics, often very different from the standard language, characterized by colloquialisms and borrowings, dialect-specific phonetic orthography and syntax, specific abbreviations, fast uptake of new vocabulary, and so on.

This was achieved in the scope of the Slovene basic research project JANES, which compiled a large and representative corpus that covered a large portion of publicly available user-generated text in Slovene, in particular tweets, blogs, forums posts, news comments and Wikipedia talk pages (Fišer, Ljubešić and Erjavec 2018). The corpus is linguistically annotated with standardized spelling, lemma, part-of- speech, and names and is freely available via the two CLARIN.SI concordancers to make it useful for theoretical and applied linguistic research. The project further produced a series of manually annotated datasets, which were used to develop methods for automatic processing of non-standard Slovene texts. Finally, the project developed a dictionary of non-standard Slovene, available through a web portal. The dictionary should be useful for teachers, students, linguists, lexicographers and the general public. All the developed resources have been made openly available for download under the Creative Commons license through the CLARIN.SI repository for research and development in computational linguistics and other automatic data processing fields.

Apart from hosting the developed resources and tools, CLARIN.SI has also contributed to several user involvement events which presented the results of the project to different user groups: two summer camps on Slovene Netspeak for high school students from 20 high schools all over Slovenia, one summer school on Internet linguistics for university students of Slovene linguistics from Slovenia and abroad and four workshops on the resources, tools and methods for analysing non-standard language for researchers and university lecturers.

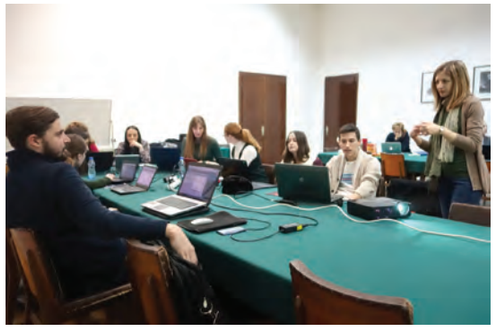
Particularly notable was the JANES Express seminar series for fellow researchers in corpus and computational linguistics which have been organized in Ljubljana (Slovenia), Zagreb (Croatia) and Belgrade (Serbia).[4] It was organized in collaboration with the Regional Linguistic Data Initiative. The seminar series presented the guidelines for manual annotation of training corpora of non-standard language varieties and the annotation platform WebAnno. As a result, three comparable gold standard corpora for tagging, lemmatization and normalization of non-standard Slovene, Croatian and Serbian have been developed and are available from the CLARIN.SI repository. Apart from testing and adapting the methodology originally developed for Slovene to two additional closely related languages, the JANES Express seminar series is also a best-practice example of knowledge transfer in the region.

More information about all these events as well as teaching materials are available on the project website.[5]

Summer camp on Slovene Netspeak

Summer school on Internet Linguistics

JANES Express seminar in Zagreb



JANES Express seminar in Belgrade

[4] http://nl.ijs.si/janes/dogodki/janes-ekspres/

[5] http://nl.ijs.si/janes/dogodki/

Interview | **Kaja Dobrovoljc**

Kaja Dobrovoljc is a Slovenian corpus linguist who works at the Centre for Language Resources and Technologies at the University of Ljubljana and regularly collaborates with CLARIN.SI and uses its infrastructure.

Could you please introduce yourself – your research background, your national and international research networks and current projects?

<

I am a linguist with an undergraduate degree in translation studies and a doctoral degree in Slovene linguistics, awarded in 2018. As a researcher at the Centre for Language Resources and Technologies of University of Ljubljana, my main research interests lie in the design, annotation and evaluation of machine-readable language resources, and their use in descriptive language research. I am currently also involved in two nationally-funded projects aimed at setting up the methodological foundations for a corpus-based grammar of Slovene (in collaboration with the Jožef Stefan Institute) and an interactive online portal for Slovene language learning (in collaboration with the University of Maribor).

>

How did you get involved with CLARIN.SI? How has CLARIN.SI supported your research? How have the results of your collaboration contributed to your research community?

<

Most of the projects I have collaborated on so far have been dedicated to publishing their results under open licenses to be freely available to anyone interested in their use and further modification. The establishment of the CLARIN.SI consortium in 2013 and the creation of the CLARIN.SI repository that followed soon after was therefore a very welcome addition to the Slovenian language infrastructure in general. On the one hand, it has enabled me and my colleagues to publish and disseminate fundamental language resources, such as the Sloleks morphological lexicon, the ssj500k training corpus or the Thesaurus of Modern Slovene in a stable online repository with long-term technical support and assistance. On the other hand, I have also benefited from the ease of access to resources developed by others, such as the GOS corpus of spoken Slovene and the JANES corpus of computer-mediated Slovene, the key language resources in my PhD research on the usage of speech-specific discourse markers in online communication.

In addition to the repository, CLARIN.SI also provides several online services, such as the noSketchEngine web concordancer and the WebAnno annotation tool. I find these particularly useful in my everyday linguistic research, and was therefore happy to join CLARIN.SI's initiative to organize hands-on training sessions for other researchers within the community as well. As the secretary of the Slovenian Language Technologies Society, I am also very grateful and proud of CLARIN.SI's continuing support of JOTA, a monthly series of talks held by Slovenian and foreign researchers on topics related to languages technologies, which are also accessible online.

>

Despite being an early-career researcher, you're one of the most prolific contributors of resources to the CLARIN.SI repository. Among others, you've created several sets of n-grams from various Slovene corpora. Could you discuss the importance of these resources for your own research as well as for the research community?

<

Although the lists of frequently recurring sequences of words in a language (also known as word n-grams) have traditionally been associated with the domain of natural language processing, where they are used in language modelling and other computational tasks, these sequences are gaining increasing importance in linguistics as well. In addition to the most commonly studied groups of expressions, such as idioms and collocations, the lists of n-grams with outstanding

frequency of usage (also known as formulaic sequences or lexical bundles) reveal an abundance of other multi-word expressions that are not necessarily fixed and idiomatic in the traditional phraseological sense, such as the expressions te dni 'these days', *v sodelovanju z* 'in collaboration with', *po drugi strain pa* 'but on the other hand' in written Slovenian, or *ali pa nekaj takega* 'or something like that', *gremo naprej* 'let's move on', *veš kaj* 'you know what' in spoken Slovenian. Phrases like these often seem uninteresting and self-evident to native speakers of a language, but they have been shown to have a special cognitive status in our brain nevertheless, and are also one of the key indicators of native-like fluency in language learners.

In my PhD work, I was mostly interested in formulaic sequences that contribute to discourse organization in spoken Slovenian. However, I applied the same extraction tool to several other reference corpora, such as written, computer-mediated and historical Slovene, producing the lists of most frequently recurring words, lemmas, PoS tags and other feature combinations with two kinds of frequency counts. These open the way to numerous interesting explorations of the nature and use of formulaic expressions in the future in various linguistic disciplines, from language teaching and lexicography to psycholinguistics and diachronic language studies.

>

You've also been part of the team that created the manually annotated ssj500k corpus.[6] Could you describe your role in its creation and annotation? Why is this corpus important for Slovenian linguistics?

<

In a way, this corpus has been pivotal to my career as a researcher, as I first came into contact with language resources and technologies as a student annotator, checking for tokenization, lemmatization and tagging mistakes performed by the automatic morphosyntactic tagger. In subsequent projects, I continued working on this dataset by manual annotation of surface syntax with the JOS dependency labels and their subsequent conversion to the complementary Universal Dependencies scheme. In addition to these layers of linguistic annotation, ssj500k has also been annotated for named entities, semantic role labels and multi-word expressions. With more than 500,000 tokens or 27,000 sentences in total, ssj500k is the largest and most extensively manually annotated corpus of Slovenian, and thus an invaluable resource for the development of fundamental language technologies, such as tokenizers, lemmatizers, taggers and parsers, which build their knowledge of the Slovenian language by observing its behaviour in such datasets. At the same time, this resource has had an important impact on Slovene linguistics as well, since many of the traditional linguistic categorizations of language phenomena in Slovenian had to be re-evaluated and improved in the annotation process, not only to meet the specific needs of machine-based applications, but also to enable systematic application to large amounts of authentic, real-world language data.

>

Together with Joakim Nivre you have worked on annotating the Treebank of Spoken Slovenian[7] following the Universal Dependencies framework. What are the benefits of the Universal Dependencies framework and why is it important for Slovene to be part of the initiative? What are the challenges of creating a treebank of spoken language data? Why is it important for Slovene linguists and the society at large to have access to a treebank of the spoken language?

<

Universal Dependencies is an international initiative aimed at a cross-lingually consistent annotation scheme for morphological and syntactic annotation, which has already been applied to over 100 treebanks in more than 80 languages, including the written and spoken treebanks of Slovenian.

Harmonizing the annotation of linguistic phenomena that are similar across languages has many important advantages for language technologies, since it enables the development of multilingual tools, such as taggers and parsers, and promotes consistent cross-lingual language technology research and evaluation in general. Many of these benefits are already visible, as several state-of-the-art tools have emerged based on this dataset and are directly applicable to all participating languages. This is especially important for small language communities that cannot necessarily afford the continuous development of high-performing language technology tools, in particular the era of fast-paced computational progress. At the same time, the large number of treebanks annotated in a unified way offers exciting opportunities for contrastive linguistic research, such as quantitative investigations into typological differences and similarities between different languages or language groups.

[6] http://hdl.handle.net/11356/1210

[7] https://universaldependencies.org/treebanks/sl_sst/index.html

This comparative aspect was also the motivation behind the construction of the spoken Slovenian UD treebank, which, in contrast to its automatically converted written counterpart, has been manually annotated from scratch, using the CLARIN.SI WebAnno installation. In the process, many speech-specific phenomena had to be addressed, such as repairs, restarts, hesitations and other types of disfluencies. Interestingly, a comparison of the annotated written and spoken treebanks of Slovenian revealed that it is not just these obvious structural particularities that distinguish speech from writing, but that the two modes also differ in terms of sentence- and phrase-structure in general. For example, spoken data consists of shorter and more elliptic sentences, fewer and simpler nominal phrases, and more relations marking interaction, deixis and modality. Just like the written ssj500k treebank, the Spoken Slovenian Treebank thus represents an important language resource for future explorations in spoken language research and spoken language technologies alike, especially given the fact that it is the spoken language that is the primary and prevalent form of human communication.

>

How can research infrastructures such as CLARIN best serve early-stage researchers and how can they best contribute to the research infrastructure?

<

Undoubtedly, research infrastructures such as CLARIN represent an invaluable source of easily accessible resources, services and support for early-stage researchers, who are usually restricted to very limited funding and need help navigating the complex landscape of digital language resources. This is certainly the case with CLARIN.SI, where Tomaž Erjavec and his team provide continuing support with language data management, such as help with annotation tools, format conversions and validations, untrivial tasks for researchers in the Humanities and Social Sciences with little computational background. At the same time, online repositories, such as the one maintained by CLARIN.SI, offer early-stage researchers a unique chance to publish and disseminate our own research results in a stable online environment, which not only contributes to increased visibility, but also creates opportunities for future collaborations.

>

Fig. 1. Slovene text from three different periods. The column in bold shows the slice the text belongs to and, in brackets, its year of publication. Each example gives the original text in the first line and the modernised word tokens in the second line, to illustrate the kind of phenomena that must be handled in the modernisation of words.

COLOPHON

This brochure is part of the ‘Tour de CLARIN’ volume II
(publication number: CLARIN-CE-2019-1537, November 2019).

Coordinated by
Darja Fišer and **Jakob Lenardič**

Edited by
Darja Fišer and **Jakob Lenardič**

Proofread by
Paul Steed

Designed by
Tanja Radež

Online version
www.clarin.eu/Tour-de-CLARIN/Publication

Publication number
CLARIN-CE-2019-1537
November 2019

ISBN
9789082990911

Contact
CLARIN ERIC
c/o Utrecht University
Drift 10, 3512 BS Utrecht
The Netherlands

www.clarin.eu

CLARIN 2019