CLARIN

Common Language Resources and Technology Infrastructure

Tour de CLARIN LATVIA

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

LATVIA

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 Latvia 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

LATVIA

XPOSTAG
FEATS
ajā mācību gadā Aizk
nāzijas 8. klasē mācījās Marisa Butnere no Am
Šajā šis pd3msln Skaitlis=Vienska 3 det
mācību mācība ncfpg4 Skaitlis=Daudzsl 3
gadā gads ncmsl1 Skaitlis=Vienska 9 obl
Aizkraukles Aizkraukle PROPN npfsg5 Skaitlis=Vienska 5
5 novada novads NOUN ncmsg1 Skaitlis=Vienska 6 nmod
6 ģimnāzijas ģimnāzija NOUN ncfsg4 Skaitlis=Vienska 8 nmod
8. 8. ADJ xo Reziduāļa_tips=I 8 amod
8 Skaitlis=Vienska 9 obl
9 Laiks=PagātneI 0 root
Skaitlis=Vienska 9 nsubj
10

Introduction

Written by **Inguna Skadiņa**

Latvia joined CLARIN ERIC in June 2016. The national coordinator of CLARIN Latvia is Inguna Skadiņa, the user involvement activities are led by Ilze Auziņa, while Roberts Darģis is involved in the Centre Committee.[1] The coordinating centre of CLARIN Latvia is the Artificial Intelligence Laboratory of the Institute of Mathematics and Computer Science, University of Latvia. The laboratory has been conducting research on natural language processing and has provided access to different language resources, including corpora and lexicons (e.g. tezaurs.lv), for almost 30 years. Prominent corpora offered by CLARIN Latvia, most of which are available through online concordancers like noSketchEngine, include:

- the LKV2018, a morphologically annotated 10-million-word corpus of modern Latvian;
- Senie, a 900-word-corpus of Latvian texts from the 16th to the 18th centuries; and
- Saeima, a corpus of parliamentary data.

CLARIN Latvia has participated in long-term national and international cooperation with different research organisations on language resource creation and maintenance – for instance, experts from the Lithuanian consortium and CLARIN Latvia have together developed LiLa, a parallel corpus of Latvian and Lithuanian. The centre also cooperates with companies in different projects on Latvian language processing tasks. To involve Digital Humanities and Social Sciences researchers, CLARIN Latvia organises practical workshops aimed at introducing its language corpora. In April 2018, a seminar was hosted that focused on LKV2018, the balanced corpus of Modern Latvian Texts. The participants of the workshop were linguists who were introduced different usage scenarios of corpus in language studies.

The Latvian CLARIN consortium has not yet been officially established.
However, during the preparatory phase of CLARIN (FP7 project), potential partners have been identified. These include providers of language resources and tools, researchers and students from the humanities and social sciences, public and government organisations and companies. The institutions that expressed interest in the CLARIN research infrastructure include universities and higher education establishments (University of Latvia (UL), Riga Stradiņš University, Liepaja University, Daugavpils University, Ventspils University College and Rēzekne Academy of Technologies), research institutes (Latvian Language institute (UL), Institute of Literature, Folklore and Art (UL) and Institute of Mathematics and Computer Science (UL)), National Library of Latvia, State Language Commission, Latvian Language agency, State Language Centre and companies - Tilde and LETA.

The activities of CLARIN Latvia are supported through the European Structural Funds project "University of Latvia and its institutes in European research space – excellence, activity, mobility and capacity" (No. 1.1.1.5/18/I/016).

Members of the Artificial Intelligence Laboratory at a brainstorming session

[1] http://clarin.lv/lv/

Tool | **NLP-PIPE**

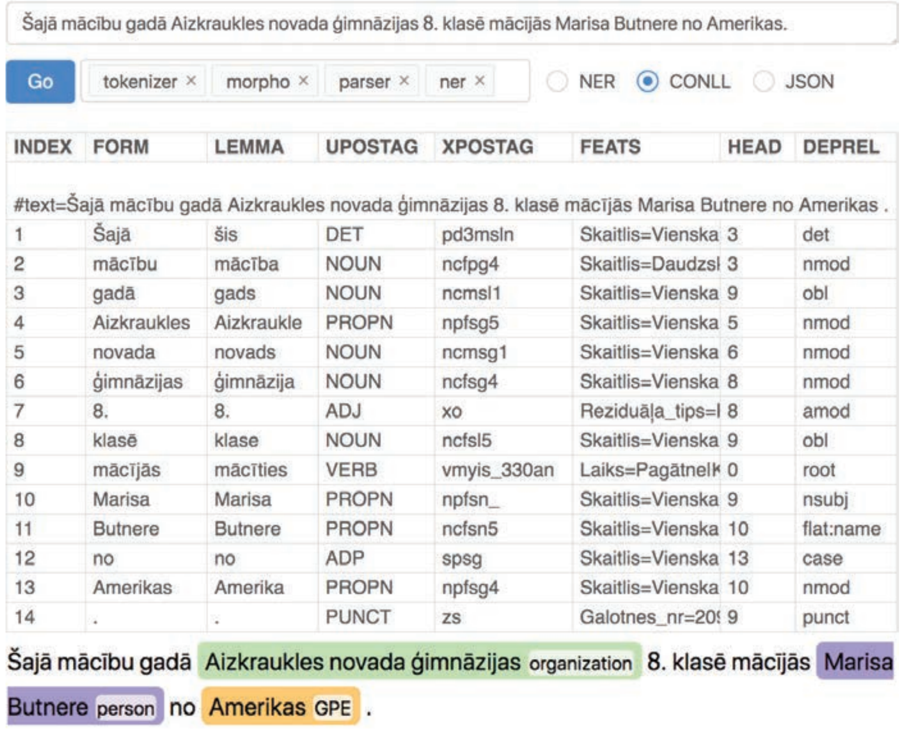
Written by **Artūrs Znotiņš**

Working with large volumes of texts usually requires multiple linguistic annotation steps which are increasingly difficult to integrate if they are based on different technologies. NLP-PIPE is a modular toolchain that allows researchers to combine multiple natural language processing tools in a unified framework. It provides the gluing code that is used to combine tools even if they are written in different programming languages and rely on conflicting library versions. It was created to make NLP technology more accessible to linguists, and to make new tool creation and integration easier for researchers and software developers.

NLP-PIPE supports a wide range of annotation services for Latvian, including tokenization, morphological tagging, lemmatization, universal dependency parsing, and named entity recognition. The easiest way to start using the toolchain is via the on-line demo version. In the web based interface, a user simply selects the required processing tools and inputs the text they want to annotate. The results can then be viewed either directly on the website (see Figure 1) or exported in several formats.

The NLP-PIPE web interface has been successfully used to perform named entity recognition on autobiographical texts, as well as to extract person mentions from an archive of photo descriptions. NLP-PIPE has also been used by CLARIN Latvia to create a multilayer corpus for Full-Stack natural language understanding (NLU), which is of crucial importance for advancing machine reading comprehension. The tool also allows post-editing of the annotation results, which helps to create reliable training datasets.

NLP-PIPE is developed at the Institute of Mathematics and Computer Science at University of Latvia and can be freely used for non-commercial purposes from GitHub. For more details on the NLP-PIPE, see Znotins and Cirule (2018) and Gruzitis and Znotins (2018).



Šajā mācību gadā Aizkraukles novada ģimnāzijas 8. klasē mācījās Marisa Butnere no Amerikas.

Go | tokenizer × | morpho × | parser × | ner × | ○ NER | ● CONLL | ○ JSON

INDEX	FORM	LEMMA	UPOSTAG	XPOSTAG	FEATS	HEAD	DEPREL
#text=Šajā mācību gadā Aizkraukles novada ģimnāzijas 8. klasē mācījās Marisa Butnere no Amerikas .							
1	Šajā	šis	DET	pd3msln	Skaitlis=Vienska	3	det
2	mācību	mācība	NOUN	ncfpg4	Skaitlis=Daudzsl	3	nmod
3	gadā	gads	NOUN	ncmsl1	Skaitlis=Vienska	9	obl
4	Aizkraukles	Aizkraukle	PROPN	npfsg5	Skaitlis=Vienska	5	nmod
5	novada	novads	NOUN	ncmsg1	Skaitlis=Vienska	6	nmod
6	ģimnāzijas	ģimnāzija	NOUN	ncfsg4	Skaitlis=Vienska	8	nmod
7	8.	8.	ADJ	xo	Reziduāļa_tips=I	8	amod
8	klasē	klase	NOUN	ncfsl5	Skaitlis=Vienska	9	obl
9	mācījās	mācīties	VERB	vmyis_330an	Laiks=PagātnelK	0	root
10	Marisa	Marisa	PROPN	npfsn_	Skaitlis=Vienska	9	nsubj
11	Butnere	Butnere	PROPN	ncfsn5	Skaitlis=Vienska	10	flat:name
12	no	no	ADP	spsg	Skaitlis=Vienska	13	case
13	Amerikas	Amerika	PROPN	npfsg4	Skaitlis=Vienska	10	nmod
14	.	.	PUNCT	zs	Galotnes_nr=20!	9	punct

Šajā mācību gadā Aizkraukles novada ģimnāzijas organization 8. klasē mācījās Marisa Butnere person no Amerikas GPE .

Figure 1: *NLP-PIPE applied to the sentence* "In this school year Marisa Butnere from America was studying in the 8th grade of Aizkraukle County gymnasium." *The results of the annotation process are displayed in the CONLL-U format with standardized columns. The XPOSTAG column corresponds to the Latvian morphological tagset based on the MULTEXT-East format. For example, the npfsg5 tags for the proper noun Aizkraukles in the fourth row translates to n – noun, p – proper, f – feminine, s – singular, g – genitive case, 5 – 5th declension. The results of the Named Entity recognition are visualised with highlighted text spans.*

References:

Znotins, A. and Cirule, E. 2018. NLP-PIPE: Latvian NLP Tool Pipeline. In *Proceedings of the CLARIN Annual Conference 2018* – The Baltic Perspective, IOS Press, 2018. doi: 10.3233/978-1-61499-912-6-183.

Gruzitis, N. and Znotins, A. 2018. Multilayer Corpus and Toolchain for Full-Stack NLU in Latvian. In *Proceedings of the CLARIN Annual Conference 2018*, 61–65. https://office.clarin.eu/v/CE-2018-1292-CLARIN2018_ConferenceProceedings.pdf. CLARIN2018_ConferenceProceedings.pdf.

Resource | **Latvian FrameNet**

Written by **Normunds Grūzītis**

The Latvian FrameNet-annotated text corpus is a balanced, multi-layered corpus that shows how words are used and what they mean. In natural language processing, it is used in applications such as information extraction, machine translation, event recognition, and sentiment analysis, while in linguistics it can be used as a valence dictionary that shows the combinatorial properties of vocabulary. Latvian FrameNet is being created within a larger industry-driven R&D project by the Institute of Mathematics and Computer Science at University of Latvia (IMCS UL) and the national news agency LETA (Grūzītis et al. 2018), which relies on natural language understanding and information extraction technologies for efficient and innovative media monitoring and content production. It is the corpus with the most annotation layers in the repository CLARIN Latvia. It is well suited for this as it is anchored in several cross-lingual syntactic and semantic representations:

- Universal Dependencies (Nivre et al. 2016), which provide the framework for the syntactic parsing of the corpus;
- FrameNet (Fillmore et al. 2003), a human- and machine-readable lexical inventory based on frame semantics for semantic role labelling;
- PropBank (Palmer et al. 2005), which provides basic predicate-argument relations such as thematic roles (e.g., agent, patient, recipient, theme, etc.);
- Abstract Meaning Representation (Banarescu et al. 2013), which are graph representations of “who is doing what to whom” in a sentence;
- auxiliary layers for named entity and coreference annotation.

Latvian FrameNet is annotated according to the latest frame inventory of Berkeley FrameNet on top of the underlying UD layer, using the CLARIN-D annotation tool WebAnno (Eckart de Castilho et al. 2016). Thus, the annotation of frames and frame elements is guided by the dependency structure of a sentence. Currently, Latvian FrameNet consists of 7,581 annotation sets (frame instances) which cover 454 different semantic frames and 834 different target verbs (lexemes), making 1,580 lexical units (LU).

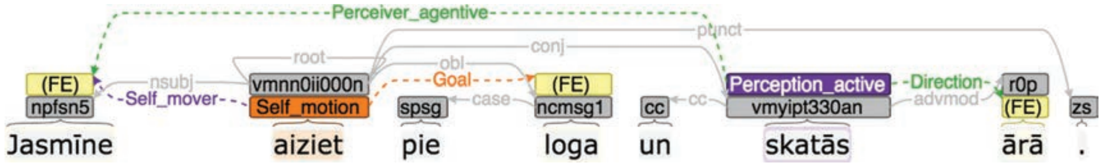


Figure 2: *Latvian FrameNet annotation*

The figure above shows how the Latvian variant of the sentence *Jasmine goes to the window and looks outside* is annotated with frame semantic labels and relations. This sentence consists of two coordinated clauses that share the same grammatical subject. The verb *aiziet* (‘go’) in the first clause is labelled with the semantic frame *self_motion* (triggered by this particular context), while the verb *skatīties* (‘look’) in the second evokes the frame *Perception_active*. Since the frame semantics are built on top of the underlying syntactic dependencies, the noun *Jasmine* gets specified with the relations *Self-mover* and *Perceiver_agent*, which are connected to the two verbs.

The dataset is available on GitHub. By the end of the project, CLARIN Latvia expects to double the size of the Latvian FrameNet corpus. The overall aim is to acquire a balanced and representative medium-sized multilayer corpus: around 10,000 sentences annotated at all the above-mentioned layers, including FrameNet. To ensure that the corpus is balanced not only in terms of text genres and writing styles but also in terms of LUs, a fundamental design decision is that the text unit is an isolated paragraph. Paragraphs were manually selected from a balanced 10-million-word text corpus: 60% news, 20% fiction, 7% academic texts, 6% legal texts, 5% spoken language, and 2% miscellaneous. As for the LUs, the goal is to cover at least 1,000 most frequently occurring verbs, calculated from the 10-million-word corpus.

Acknowledgements

This work has received financial support from the European Regional Development Fund under the grant agreement No. 1.1.1.1/16/A/219 (Full Stack of Language Resources for Natural Language Understanding and Generation in Latvian).

References:

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., and Schneider, N. 2013. Abstract Meaning Representation for Sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, Sofia, Bulgaria, 178–186.

Eckart de Castilho, R., Mujdricza-Maydt, E., Yimam, S.M., Hartmann, S., Gurevych, I., Frank, A., and Biemann, C. 2016. A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities,* Osaka, Japan, 76–84.

Fillmore, C.J., Johnson, C.R., and Petruck, M.R.L. 2003. Background to FrameNet. *International Journal of Lexicography* 16 (3): 235–250.

Grūzītis, N., Nespore-Berzkalne, G., and Saulite, B. 2018. Creation of Latvian FrameNet based on Universal Dependencies. In Proceedings of the International FrameNet Workshop 2018: *Multilingual FrameNets and Constructicons* (IFNW), Miyazaki, Japan, 23–27.

Grūzītis, N., Pretkalnina, L., Saulite, B., Rituma, L., Nespore-Berzkalne, G., Znotins, A., and Paikens, P. 2018. Creation of a Balanced State-of-the-Art Multilayer Corpus for NLU. In *Proceedings of the 11th International Conference on Language Resources and Evaluation* (LREC), Miyazaki, Japan, 4506–4513.

Nespore-Berzkalne, G., Saulite, B., and Gruzitis, N. 2018. Latvian FrameNet: Cross-Lingual Issues. In *Human Language Technologies – The Baltic perspective: Proceedings of the Eighth International Conference Baltic HLT 2018*, Tartu, Estonia, 96–103.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C.D., McDonald, D., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation* (LREC), 1659–1666.

Palmer, M., Gildea, D., Kingsbury, P. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics* 31 (1): 71–106.

Event | **Tools and Resources for Digital Humanities Research Seminar**

Written by **Ilze Auziņa**

The seminar *Tools and Resources for Digital Humanities Research* was organised by the staff of the Artificial Intelligence Laboratory (AILab) of the Institute of Mathematics and Computer Sciences, University of Latvia to showcase the language tools and resources developed at AILab. The seminar took place on February 1, 2018 and brought together a wide range of humanities researchers, including philologists, journalists, political scientists, translators, librarians, historians and other representatives of the Humanities and Social Sciences. Among the audience were both students and experienced researchers who wanted to find out what tools were available for the analysis and processing of Latvian texts and how to use corpus linguistics methods, for example, in Literary Studies.

Normunds Grūzītis opening the seminar

During the workshop, CLARIN national coordinator Inguna Skadiņa introduced the attendees to CLARIN and outlined the plans to establish the CLARIN infrastructure in Latvia. Although CLARIN was already actively promoted during the preparatory phase, which ended in 2012, this seminar was the first event in which CLARIN Latvia was presented to a wider audience after Latvia joined CLARIN ERIC. The participants were introduced to the national and international aims of CLARIN, and invited to actively participate in the creation of the CLARIN network of expertise in Latvia.

CLARIN: no priekšdarbiem pie darbiem
Inguna Skadiņa
CLARIN Latvija nacionālā koordinatore
Common Language Resources and Technology Infrastructure
Vienota valodas resursu un tehnoloģiju infrastruktūra
CLARIN

Inguna Skadiņa introducing CLARIN

Other speakers presented the tools, resources and research projects which are to serve as the backbone of CLARIN Latvia. Artūrs Znotiņš and Pēteris Paikens presented different types of text analysis, such as lexical, semantic and sentiment analysis, and the tools available for such analyses. Baiba Saulīte and Ilze Auziņa introduced the on-going project *Full Stack of Language Resources for Natural Language Understanding and Generation in Latvian.* The project aims to create multi-layered semantically annotated language resources for Latvian, anchored in the widely acknowledged multilingual representations of lexico-grammatical relations, such as PropBank, FrameNet and Universal Dependencies, and showcase their use by developing an advanced Latvian abstractive text summarizer to be evaluated on the media monitoring use case. Roberts Darģis introduced the corpora developed at AILab and demonstrated their use for digital humanities research. Finally, Ilmārs Poikāns turned to methods and tools for digitizing language and history materials.

The audience at the Lavtian Tools and Resources for DH research seminar

The workshop attracted so much interest that not everyone had the chance to participate: the maximum number of participants was 50, but nearly 100 people signed up for the seminar. The great number of participants from diverse research backgrounds showed that there is much interest in the use of language tools and resources among Latvian researchers. What is more, after the seminar several participants registered for the master's course Introduction to Computational Linguistics, taught at the Faculty of Humanities, University of Latvia. In addition, experts from CLARIN Latvia discussed possible opportunities for collaboration with political scientists from Riga Stradiņš University and translational scientists from Ventspils University of Applied Sciences.

LATVIA

Interview | **Sanita Reinsone**

Sanita Reinsone is a leading researcher at the Institute of Literature, Folklore and Art at the University of Latvia.

Please describe your research background. What sparked your interest in Digital Humanities?

<

I work at the Institute of Literature, Folklore and Art at the University of Latvia, where I research life writing, oral history and digital participatory practices. I hold a PhD in philology, which I obtained in 2012 from the University of Latvia. I am leading several Digital Humanities and Cultural Heritage initiatives at the Institute. This gives me the opportunity to collaborate with passionate and talented researchers from diverse fields, such as folkloristics, literary studies, music and theatre research, as well as history and linguistics. Concretely, my work mainly involves the development and curation of different crowdsourcing initiatives within Digital Humanities and Cultural Heritage.

I became interested in digital approaches to the humanities when I was a first-year philology student and started working at the Artificial Intelligence Laboratory of the Institute of Mathematics and Computer Sciences at the University of Latvia, the same team which is now heading CLARIN Latvia, where the creation of some of the first Latvian corpora was already underway at the time. I helped with digitizing Latvian literary classics and folklore publications, from which I learned how digital methods can be used in the study of cultural heritage. This experience served as a foundational background for my future research at the Archives of Latvian Folklore (part of the Institute of Literature, Folklore and Art), since digital methodologies were not taught at the University at that time.

>

You are a leading researcher at the Institute of Literature, Folklore and Art at the University of Latvia. What does it mean to apply a Digital Humanities approach to folklore? Could you give a concrete example of how folklore studies can be complemented by such an approach?

<

A digital approach to folklore collections essentially means that we are able to work with tools that can automatically analyse unstructured collections and provide new ways of visualizing, indexing and classifying the texts and other types of folklore material. Implementing such an approach has greatly sped up the initial process of collecting and sorting the data, which in our field are very diverse in terms of the type of material. It is now easier than ever to answer complex research questions, as computation tools allow us to examine textual and stylistic variation observed in different periods and dialects or the geographical distribution of vernacular expressions in a precise and very time efficient manner. This has only become possible now that our digitized folklore texts are enriched with metadata such as geographical information and interlinked with other materials in the Archives, such as photographs and sound recordings.

For instance, Sandis Laime, who is a post-doctoral researcher at the Archives, has used geospatial analysis tools to examine the geographical distribution of legends related to witches and witchcraft in Latvia. His research turned out to be of crucial importance for the better understanding of the historical aspects of the tradition as well. Before he started working on this topic, an opinion existed that witchcraft beliefs were more or less uniform in the whole country and did not differ much from the European tradition. However, Sandis was able to prove, by using digital methods, that the Latvian witchcraft belief system is not at all as homogeneous as previously believed. Along with the character of the diabolized witch, which was present in most parts of Latvian at the turn of the 20th century, he was able to determine several conservative areas in the peripheral regions of the country which were not influenced by Christian demonology. Finally, the geographical aspect of the research turned out to be historically significant.

>

How does the collection, processing, analysis and archiving of research material in folklore differ from other DH disciplines? What are the main obstacles with respect to the technologies applied to your material? How could CLARIN be of help in this respect?

<

Since 2000, my research has mostly focused on oral history. Related to this, together with my colleagues I've recently launched an initiative at the Archives to create the Autobiography Collection, which consists of written diaries, memories and life stories. Such materials are relatively unstudied phenomena, at least in comparison to oral life stories. They provide a very personal insight into the lives of ordinary people and a direct perception of a historical event, especially since these texts are not written and edited by professional biographers. There are often spelling mistakes, for example, but this just means that we're dealing with pure, unaltered text that provides a unique and often colourful perspective. The textual materials of the Autobiography Collection are very diverse, possibly more so than in other disciplines. Life writing texts such as diaries are often accompanied by additional contextual materials such as photographs and audio interviews with the authors or contributors.

In general, the Archives of Latvian Folklore hold a lot of the dialectal speech, and the quality of older sound recordings often isn't the best, which is then a problem for speech-to-text transcription software. The archiving of the material is also very challenging because of its diversity, so we are planning on future collaboration with the Latvian CLARIN consortium to streamline the collection and digitization process. Additionally, vernacular expressions are often used on social media these days – in a way, such informal language is a type of modern folklore – and I believe CLARIN could provide us with help to mine such data.

>

Does your Institute collaborate with the Latvian CLARIN consortium in the digitization of folklore and the curation of digital folklore archives?

<

We collaborate with CLARIN Latvia at two levels. The first, of course, is the personal level, which means that we often consult with their experts on how to use a specific language tool or resource. The second, which I think is crucial, is the institutional level. This involves communication on how to improve our Archives' infrastructure and align it with CLARIN standards.

At the Archives, we are currently creating a corpus of life writing. First, however, we have to reach out to the general public and get in touch with museums and other archives in order to get the materials. In relation to such outreach, we already have close cooperation with the CLARIN Latvian team, as we have successfully organized several awareness raising and knowledge sharing events for researchers and students of the Humanities and Social Sciences. I see that such educational initiatives are appreciated and very much needed, since they provide direct showcases on how language tools and resources can be applied within qualitative research and bridge the gap between computational experts and Humanities and Social Sciences researchers. For instance, one such successful event was a Digital Humanities workshop which members of our Institute and the National Library of Latvia organized together with CLARIN Latvia. The interest was unexpectedly high, and we couldn't provide enough seats for everyone who wanted to attend.

For the future, we very much look forward to incorporating some of CLARIN Latvia's automated services for language processing at the Archives. We especially want to implement their tools for speech-to-text transcription and the automatic annotation of spoken data, since conducting interviews with informants can be a very laborious process if you have to do the transcriptions by hand. I would also appreciate an automatic image annotator, given the very large number of photographs in the Archives.

>

Your Institute has also been successfully involved in crowdsourcing. Could you please describe this? Why is crowdsourcing important for Digital Humanities?

<

The crowdsourcing initiative began five years ago, when we set up our Archives' online repository. We were faced with a very large number of handwritten manuscripts that were not yet converted to a computer-readable format. Since we wanted the Archives to be not only openly accessible, but also involved with the general public, we decided to reach out and find volunteers who would be willing to transcribe the manuscripts, which were made available on the platform.

In the first year, the volunteers managed to transcribe around 1,000 handwritten pages. This wasn't a very large number, but at that point we had not yet managed to fully promote the initiative, since we were mostly focusing on the further development and maintenance of the repository. Soon after, we started collaborating with the Latvian branch of UNESCO, and together we launched a special outreach campaign with which we invited schoolchildren to participate in transcribing the handwritten texts. It was a wonderful experience that lasted for a little more than two months. During this relatively short period, schoolchildren managed to transcribe around

15,000 pages which is a lot of text, especially in comparison to the first round. This inspired us to continue with the initiative, which gradually built an active community of transcribers who are passionate about our materials. They regularly communicate with us and send helpful suggestions for potential future implementations to the Archives. A concrete result of our collaboration with the transcribers is that we managed to establish a new and improved online platform for transcription which is very user friendly and minimises the need for technical knowledge – the volunteers only need to log in, select one of the 10 languages that the manuscripts are in, and then immediately begin transcribing one of the manuscript pages. There is also an option to add comments to the text, which further solidifies our collaboration.

I think the reason as to why this crowdsourcing initiative has been a success is the fact that many people take pride in their local lore. Perhaps what's important here is that folklore does not only encompass such genres as folk tales, legends and folk songs; it also includes a lot of regional knowledge and memories of the old ways of life and traditions in rural areas that are disappearing from the modern world. Hence the reason why many people are so willing to engage with our materials.

In addition, we have recently started several other crowdsourcing initiatives. For instance, a children's poetry reading campaign, in which we invited the public to add to the database of Latvian literature by reading poems out loud, recording their voices for the enjoyment of future generations and for research. The poetry chosen for this project was written at least one hundred years ago by well-known poets, loved by many generations, and also lesser-known poets worthy of attention. This initiative, which was also supported by the National Library of Latvia, was very successful in that it basically led to the creation of a speech corpus of poetry, which we now use to study the different ways in which poetry is read; that is, the different manners, and thus whether it is recited or sung, and so forth. Another initiative, which will be launched on 15 February, is called Sing with the Archives, with which we aim to popularize the musical recordings of the Archives and to collect modern musical versions that will be performed by the participants. Additionally, a campaign called Contemporary Calendar invites the public to record their special calendar events and thus help researchers to study the contemporary ritual year.

>

How can research infrastructures such as CLARIN benefit from crowdsourcing?

<

I believe that CLARIN-related research could also be complemented by crowdsourcing, especially if it involves, for instance, building a spoken language corpus. In order to ensure that such a corpus is representative of the spoken language, it should also contain dialect samples. I think it wouldn't be too difficult to motivate people to provide their own recordings, given that a person's dialect is part of his or her personal identity, much in the same way as history and folklore are. What's crucial, though, is that CLARIN should focus on making their tools, platforms and interfaces as user-friendly as possible, which means that CLARIN experts should actively engage with the external community, be they established researchers or passionate amateurs, and try to meet their needs and expectations. As the success of our own crowdsourcing initiative shows, communication from both sides goes a long way to establishing fruitful cooperation.

>

How are the Digital Humanities represented in Latvian research institutions and universities?

<

Digital Humanities in Latvia are fairly new. Although computational linguistics has quite a long tradition in Latvia, other disciplines have only recently started to adopt digital methods. I think that crucial to its promotion in our country is the digitalhumanities.lv initiative, which involves voluntary collaboration among research institutions like our Archives and CLARIN Latvia. The initiative is currently organizing the 2019 Baltic Summer School of Digital Humanities. In 2018, Riga Technical University launched the first master's programme in Digital Humanities, which has turned out to be quite popular among students. In addition, the Faculty of Humanities at the University of Latvia has started to offer foundational courses in Digital Humanities, which often get filled to full capacity. Generally, I think the younger generation is keen on learning how to apply digital methodologies in their work or use them in their studies, even if they come from traditionally non-digital fields like history or philology.

For the future, we plan on further collaborating with other Latvian research institutes like Riga Technical University and the National Library of Latvia to promote Digital Humanities, computational linguistics and computational folkloristics in Latvian Universities, and plan on including additional subjects in school curricula.

>

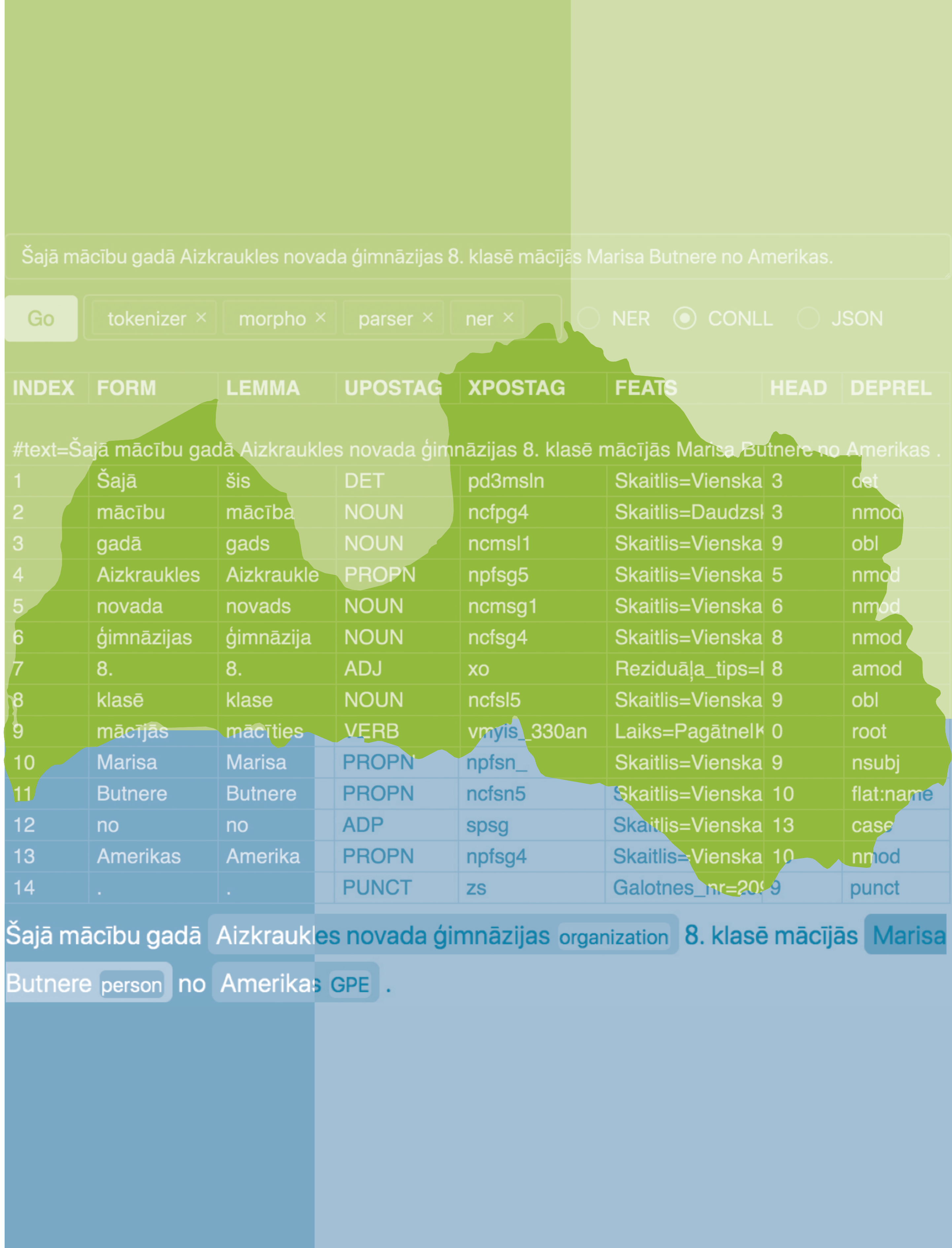
How in your opinion could CLARIN Latvia help promote computational methods and the use of research infrastructures in traditional fields such as your own (i.e., folklore studies)?

<

I think educational activities should be a major priority for CLARIN Latvia at this stage, as this is the most efficient way for experienced and novice researchers to learn how to integrate the CLARIN infrastructure in their own work. What is more, such activities can also spark new collaboration opportunities among researchers from different disciplines.

Another important topic on CLARIN's agenda, in my opinion, should be copyright issues. For example, many of the materials in our Archives are challenging from the perspective of copyright, since collecting life writing such as diaries and memoirs means that we store a lot of personal and sensitive data. Although we try to be very rigorous in securing copyrights and discussing this with our informants, it would be very helpful if there were more joint discussions about the legal implications related to the creation and maintenance of such collections, as there are many other institutes who are dealing with materials that fall into a kind of legal grey area. This is why I think CLARIN could be very helpful in this respect by providing researchers with some helpful and easily reusable scenarios and guidelines.

>



Šajā mācību gadā Aizkraukles novada ģimnāzijas 8. klasē mācījās Marisa Butnere no Amerikas.

#text=Šajā mācību gadā Aizkraukles novada ģimnāzijas 8. klasē mācījās Marisa Butnere no Amerikas .

INDEX	FORM	LEMMA	UPOSTAG	XPOSTAG	FEATS	HEAD	DEPREL
1	Šajā	šis	DET	pd3msln	Skaitlis=Vienska	3	det
2	mācību	mācība	NOUN	ncfpg4	Skaitlis=Daudzsl	3	nmod
3	gadā	gads	NOUN	ncmsl1	Skaitlis=Vienska	9	obl
4	Aizkraukles	Aizkraukle	PROPN	npfsg5	Skaitlis=Vienska	5	nmod
5	novada	novads	NOUN	ncmsg1	Skaitlis=Vienska	6	nmod
6	ģimnāzijas	ģimnāzija	NOUN	ncfsg4	Skaitlis=Vienska	8	nmod
7	8.	8.	ADJ	xo	Reziduāļa_tips=I	8	amod
8	klasē	klase	NOUN	ncfsl5	Skaitlis=Vienska	9	obl
9	mācījās	mācīties	VERB	vmyis_330an	Laiks=PagātnelK	0	root
10	Marisa	Marisa	PROPN	npfsn_	Skaitlis=Vienska	9	nsubj
11	Butnere	Butnere	PROPN	ncfsn5	Skaitlis=Vienska	10	flat:name
12	no	no	ADP	spsg	Skaitlis=Vienska	13	case
13	Amerikas	Amerika	PROPN	npfsg4	Skaitlis=Vienska	10	nmod
14	.	.	PUNCT	zs	Galotnes_nr=20	9	punct

Šajā mācību gadā Aizkraukles novada ģimnāzijas organization 8. klasē mācījās Marisa Butnere person no Amerikas GPE .

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
Common Language Resources and
Technology Infrastructure

CLARIN 2019