CLARIN in Latvia: From the Preparatory Phase to the Construction Phase and Operation

Inguna Skadiņa, Ilze Auziņa, Normunds Grūzītis and Artūrs Znotiņš

Institute of Mathematics and Computer Science, University of Latvia, Raiņa bulv. 29, Riga, Latvia
{inguna.skadina, ilze.auzina, normunds.gruzitis, arturs.znotins}@lumii.lv

Abstract. Qualitative and reliable language resources and natural language processing tools are key elements for research in digital humanities (DH). Several research infrastructures, e.g., CLARIN, DARIAH, provide access to the digital research objects around Europe and beyond. Although these are pan-European research infrastructures, availability of content and the readiness of the particular node varies from country to country. This paper aims to present the current status of the CLARIN research infrastructure in Latvia – key language resources and tools identified, readiness of the technical infrastructure, first steps to collaboration with DH researchers and initiatives on user involvement and education. Being an active participant of the CLARIN initiative during its preparation phase, Latvia joined CLARIN ERIC only four years after its establishment. This four-year gap puts Latvia's CLARIN node in a construction phase, while in many countries CLARIN is already operational. Although CLARIN Latvia is in a construction phase, researchers of Latvia already now can benefit from the language resources and tools from different members of CLARIN ERIC through single sign-on.

Keywords: Latvian Language, CLARIN, Research Infrastructure, Language Resources and Tools, Toolchain for Language Processing.

1 Introduction

The Languages and their diversity have always been among priorities of the European Union multilingualism policy. The ways how to support natural languages, especially less resourced languages, in a digital age have been a hot topic for more than 10 years. In 2012 the META-NET Whitepapers [1] have identified the risk of possible digital extinction for 21 European languages, if these languages will be not sufficiently supported in digital means. Recently published European Parliament resolution on language equality in the digital age calls the Commission "to make as a priority of language technology those Member States which are small in size and have their own language" [2].

There are several long-term initiatives running, that aim not only to support natural languages in AI-centered technologies, but also provide means for digital humanities researchers to research and preserve human languages. CLARIN (Common Language Resources and Technology Infrastructure) is a research infrastructure that was initiated from the vision that all digital language resources and tools from all over Europe and beyond will be accessible through a single sign-on online environment to support researchers in the humanities and social sciences [3].

In 2012 CLARIN has been established as European Research Infrastructure Consortium (ERIC) to create and maintain an infrastructure to support the sharing, use and sustainability of language data and tools for research in the humanities and social sciences [4]. Although Latvia was an active member of CLARIN during the preparatory phase [5], it became a full member of CLARIN ERIC only in June 2016.

The coordinating center of CLARIN Latvia is the Artificial Intelligence Laboratory (AiLab) of the Institute of Mathematics and Computer Science, University of Latvia. The laboratory is conducting research on natural language processing and provides access to different language resources for almost 30 years.

In many countries the CLARIN infrastructure today is fully operational. However, CLARIN Latvia node is in a construction phase - the technical infrastructure needs to be set up and populated with content (language resources and tools). However, already now researchers of Latvia can benefit from CLARIN tools and resources created and maintained by other CLARIN nodes through the single sign-on supported by Latvian academic identity federation LAIFE.

This short paper aims to present an overview of the current status of the CLARIN research infrastructure in Latvia – key language resources and tools identified, development of the technical infrastructure, collaboration with DH researchers in Latvia and initiatives on user involvement and education.

2 National Initiatives

The necessity of language technology support in digital means and importance of language technologies for the long-term survival of the Latvian language has been always recognized at the policy planning documents. Among the four objectives of the State Language Policy Guidelines for 2015–2020, the third objective “Latvian language research and development” envisages support for development of language technologies, databases, corpora and terminological resources. Latvian language and technologies are also included in nine research priorities of Latvian science for 2018–2021.

Research on language resources and technologies has been supported through the State Research Programmes, EU Structural Funds Programmes, grants of the Latvian Science Council, EU Horizon 2020 and CEF Programmes. However, Latvia lacks dedicated language technology program and definite long-term plan for support of language resource infrastructures. As a result, research and development activities in human language technologies and creation of language resources and tools are fragmented and in many cases insufficiently supported.

Where it concerns support for research infrastructures, only limited funding for construction of CLARIN Latvia research infrastructure is being provided since 2018. Interruption in funding for six years and lack of sufficient funding currently are the main obstacles, why Latvia is the only Baltic country, where CLARIN infrastructure is still not in exploitation phase.

3 Latvian Language Resources and Tools

The most important part of the CLARIN research infrastructure are language resources and tools necessary for digital humanities research. During the CLARIN preparation phase 35 language resources and 9 tools have been identified and registered in CLARIN Virtual Language Observatory [5]. In this chapter we present some of the most recent and most important Latvian language resources and tools for humanities researchers, more detailed overview of recent human language technology achievements are presented in papers by Skadiņa et al. [6, 7].

3.1 Corpora and Lexical Resources

Latvian language corpora developed and maintained by the CLARIN Latvia coordinator are listed on *korpuss.lv* website. The list includes both text and speech corpora, general and specialised corpora (e.g., a corpus of historical texts, and a corpus of parliamentary transcripts). The modern Latvian language is represented through the Balanced Corpus of Modern Latvian [8], which has been recently extended to 10 million running words.

The most recent and significant result in the creation of annotated text corpora for Latvian is a syntactically and semantically annotated multi-layered corpus. The corpus contains several annotation layers for the same text units, anchored in widely acknowledged cross-lingual meaning representations (see **Fig. 1**): Universal Dependencies, FrameNet, PropBank, and Abstract Meaning Representation, as well as auxiliary layers of named entity and coreference annotations [9].

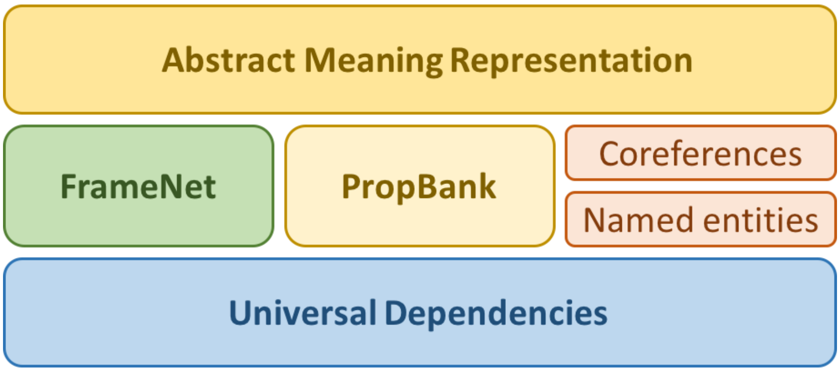


Fig. 1. Annotation layers of NLU corpus [9].

The multi-layered corpus contains about 13 thousand sentences (more than 4,200 paragraphs – text units) annotated at all layers. To ensure that the corpus is balanced not only in terms of text genres and writing styles but also in terms of lexical units, a fundamental design decision is that the text unit is an isolated paragraph. Paragraphs were manually selected from the Balanced Corpus of Modern Latvian: 60% news, 20% fiction, 7% academic texts, 6% legal texts, 5% spoken language, 2% miscellaneous. As for the lexical units, the goal was to cover at least 1,000 most frequently occurring verbs (at least 2,000 lexical units), calculated from the 10-million-word corpus. The primary purpose of the multi-layer corpus is for the development and evaluation of automatic natural language understanding (NLU) means for Latvian, as well as for advanced corpus-based linguistics studies that have not been possible before due to the lack of such resources.

The largest open lexical database for Latvian is *Tezaurs.lv* [10]. It aims to bring together all Latvian words and frequently used multi-word units, allowing for the integration of other language resources and tools. Today it contains more than 315 thousand lexical entries that are compiled from more than 300 sources. *Tezaurs.lv* is popular not only among researchers, translators, developers of language games, and other professionals, but also widely used by the general public: journalists, students and many others. The dictionary is enriched with phonetic, morphological, semantic and other annotations and is enhanced with language processing tools allowing automatic generation of inflectional forms and automatic selection of corpus examples. Additionally, representative FrameNet corpus examples are linked to the entries of frequently used verbs, providing a complementary perspective on the possible sense split [11].

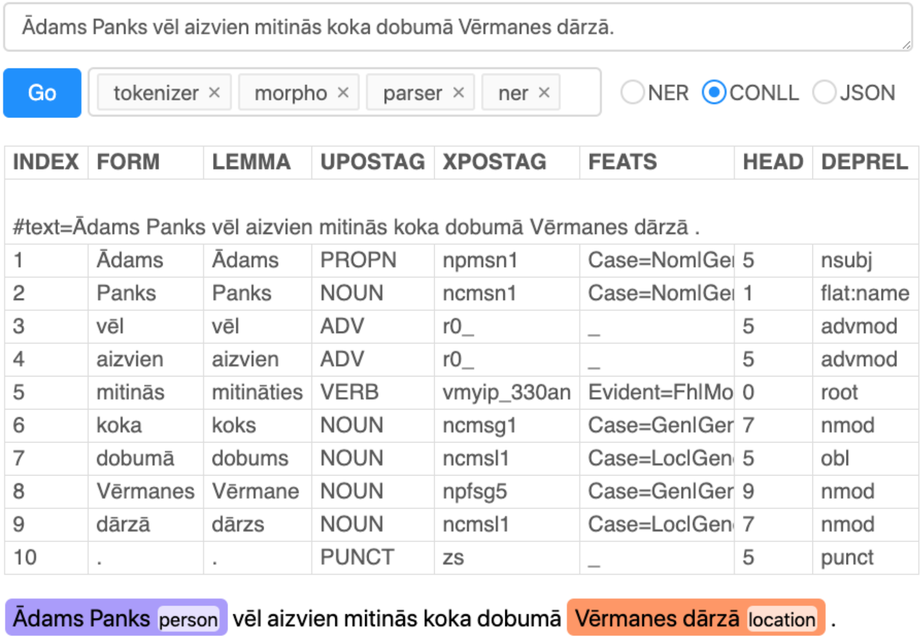
3.2 Toolchain for Latvian Language Processing: NLP-PIPE

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 [12, 13]. It provides the gluing code to combine tools even if they are written in different programming languages and rely on conflicting library versions.

NLP-PIPE was created to make NLP technology more accessible to linguists, and to make new tool creation and integration easier to researchers and software developers. It supports a wide range of annotation services for Latvian, including tokenization, morphological tagging, lemmatisation, 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 (nlp.ailab.lv), 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 Fig. **3**) or exported in several formats.

NLP-PIPE has been used to create a multilayer corpus described in a previous section. The tool also allows post-editing of the annotation results which helps to create reliable datasets.



Ādams Panks vēl aizvien mitinās koka dobumā Vērmanes dārzā.

INDEX	FORM	LEMMA	UPOSTAG	XPOSTAG	FEATS	HEAD	DEPREL
#text=Ādams Panks vēl aizvien mitinās koka dobumā Vērmanes dārzā .							
1	Ādams	Ādams	PROPN	npmsn1	Case=Nom|Ge	5	nsubj
2	Panks	Panks	NOUN	ncmsn1	Case=Nom|Ge	1	flat:name
3	vēl	vēl	ADV	r0_	_	5	advmod
4	aizvien	aizvien	ADV	r0_	_	5	advmod
5	mitinās	mitināties	VERB	vmyip_330an	Evident=Fh|Mo	0	root
6	koka	koks	NOUN	ncmsg1	Case=Gen|Ger	7	nmod
7	dobumā	dobums	NOUN	ncmsl1	Case=Loc|Gen	5	obl
8	Vērmanes	Vērmane	NOUN	npfsg5	Case=Gen|Ger	9	nmod
9	dārzā	dārzs	NOUN	ncmsl1	Case=Loc|Gen	7	nmod
10	.	.	PUNCT	zs	_	5	punct

Ādams Panks person vēl aizvien mitinās koka dobumā Vērmanes dārzā location .

Fig. 2. Fig. **3.** NLP-PIPE applied to the sentence “Adam Punk still lives in a wooden cavity in Vērmane Garden.” The results of the annotation process are displayed in the CONLL-U format with standardised columns. The XPOSTAG column corresponds to the Latvian morphological tag set based on the MULTEXT-East format. For example, the npmsn1 tags for the proper noun Ādams in the first row translates to n – noun, p – proper, m – masculine, s – singular, n – nominative case, 1 – 1st declension. The results of the Named Entity recognition are visualized with highlighted text spans.

4 Technical Infrastructure

The CLARIN research infrastructure is a distributed network of centres. The backbone of CLARIN technical infrastructure is so called CLARIN B-Centres or Service Providing Centres. These centers offer the scientific community access to the language resources and services. According to the Statutes of the CLARIN ERIC (approved by the European Commission on 4 April 2018) each member shall “provide at least one data and service centre”.

While our goal is to become a B-Center, we started with setting up the CLARIN Latvia C-center (C-Centres are the Metadata Providing Centres, their metadata are integrated with CLARIN). Our platform runs under the clarin-dspace repository software developed for the LINDAT/CLARIN repository by the Institute of Formal and Applied Linguistics of the Charles University. Users can sign in via the CLARIN Service Provider Federation, e.g., researchers of Latvia can sign in using their home organization accounts in Latvian academic identity federation LAIFE. This allows Latvian and European researchers benefit from Latvian language resources and tools, including ones described in this paper, already now, i.e., before our repository meets requirements for certification as B-Center.

5 User Involvement

Another important dimension of the CLARIN infrastructure is the knowledge sharing. Knowledge sharing includes activities related to the user involvement and education, such as training (workshops, hands-on sessions), discussions and dissemination activities (conferences, seminars), as well as, everyday support through knowledge center.

5.1 Workshops and Seminars

To involve Digital Humanities and Social Sciences researchers of Latvia and introduce them with CLARIN Latvia, we organize seminars and practical workshops.

The seminar "Tools and Resources for Digital Humanities Research" was organized to showcase the language tools and resources developed at the Artificial Intelligence Laboratory. The 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 were invited to actively participate in the creation of the CLARIN network of expertise in Latvia.

The seminar 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 tools for the analysis and processing of Latvian texts, especially how to use corpus linguistics methods.

The workshop attracted so much interest that not everyone had the chance to participate. 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. 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 two practical seminars were organized to introduce DH researchers to language corpora. In April 2018, a seminar that focused on the Balanced Corpus of Modern Latvian was organized, while in May, 2019 workshop on how to use Latvian treebank in linguistic studies took place. The students of humanities, including future Latvian language teachers, are regularly introduced in workshops with the latest Latvian language corpora and the possibilities for their use.

5.2 CLARIN in Education

In addition to the seminars and workshops, the Latvian language resources and tools are being introduced to humanities students during the master level course on computational linguistics at the University of Latvia. During the course students are also introduced to the research infrastructures, they are asked to explore CLARIN repository and present their findings at a seminar.

Latvian language resources and tools are also explored by doctoral students of the Liepāja University and Ventspils University of Applied Sciences. Materials of these classes have been uptaken by former students of these universities who have became teachers.

Recently the University of Latvia has prepared a professional-oriented higher education study programme "Teacher", where the "Latvian language and literature teacher" sub-programme will have a course "Introduction to Corpus Linguistics".

5.3 Knowledge Centre for Systems and Frameworks for Morphologically Rich Languages

One of the four priorities of the CLARIN strategy for 2018–2020 is a knowledge sharing infrastructure. There are more than 15 CLARIN Knowledge centers established currently to support users and share knowledge: some of them are language or country specific, while others are domain/area specific. Since Latvian is morphologically rich language, we decided to join CLARIN Knowledge Centre for Systems and Frameworks for Morphologically Rich Languages (SAFMORIL). SAFMORIL serves linguists and computational linguists developing and adapting morphologies as well as digital humanities scholars and computer scientists processing language data. Today it brings together researchers and developers from the University of Helsinki, University of Tromsø, Vytautas Magnus University and the Institute of Mathematics and Computer Science (University of Latvia) in the area of computational morphology and its application to language processing. CLARIN Latvia contributes SAFMORIL with its special knowledge about Latvian language tools and corpora.

6 Governance

The Although CLARIN Latvia consortium has not yet been officially established, during the preparatory phase of CLARIN, potential consortium members have been identified. The institutions that expressed interest in the CLARIN research infrastructure include universities and higher education establishments, research institutes, museums and libraries, administration institutions and companies.

The National Advisory board consisting of 17 members from academia, industry and government has also been established during the preparatory phase. Tasks of the Advisory board include monitoring work of CLARIN national contact point, defining priorities and providing recommendations facilitating fulfilment of objectives of CLARIN in Latvia. The National Advisory board will be one of the instruments to collect demands coming from digital humanities and social science researchers. and select more important for implementation.

7 Conclusion and Next Steps

In this short paper we presented current status of CLARIN Latvia research infrastructure for humanities and social sciences. We presented some important Latvian language resources and tools, described the current state of the technical infrastructure, user involvement and knowledge sharing activities and stakeholders interested in CLARIN infrastructure.

At the time when this paper is being submitted, the final steps (e.g., stabilization and localization) are being taken to set a C-centre repository in Latvia (available at repository.clarin.lv). Our next step is to populate it with content and raise awareness between digital humanities researchers. We plan to organize a conference in March, 2020 to introduce DH and social sciences (SS) researchers with CLARIN Latvia. This event, as well as forthcoming CLARIN workshop will be used to collect needs of DH and SS researchers. In addition, to increase understanding of NLP among DH researchers, we will continue workshop series in which we introduce users to the particular tool/language resource and its typical usage scenario.

Where it concerns technical infrastructure, the next step will be to fulfill requirement for certification as a B-Center.

Acknowledgements

Activities of CLARIN Latvia are supported by the European Structural Funds project No. 1.1.1.5/18/I/016. We would like to thank the IT department of the University of Latvia, especially Arnis Voitkāns, Kārlis Kalvišķis and Egija Rapa, for their support to create a DSpace repository for the CLARIN Latvia infrastructure.

References

1. Rehm, G. and Uszkoreit, H. eds. META-NET White Paper Series, Springer (2012).
2. European Parliament resolution of 11 September 2018 on language equality in the digital age, last accessed 2019/10/14.
3. Hinrichs, E. and Krauwer, S.: The CLARIN Research Infrastructure: Resources and Tools for E-Humanities Scholars. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), pp. 1525–31 (2014).
4. De Jong, F., Maegaard, B., de Smedt, K., Fišer, D. and van Uytvanck, D.: CLARIN: Towards FAIR and Responsible Data Science Using Language Resources. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), pp. 3259–3264 (2018).
5. Skadiņa, I.: CLARIN in Latvia: current situation and future perspectives. In: Nordic Perspectives on the CLARIN Infrastructure of Common Language Resources (2009).
6. Skadiņa, I.: Some Highlights of Human Language Technology in Baltic Countries. In: Databases and Information Systems X, pp. 18–30, IOS Press (2019).
7. Skadiņa, I., Auzina, I., Deksne, D., Skadins, R., Vasiljevs, A., Gailuna, M., Portnaja, I.: Filling the gaps in Latvian BLARK: Case of the Latvian IT Competence Centre. In: Human Language Technologies – The Baltic Perspective, pp. 3–11, IOS Press (2016).

8. Levane-Petrova, K.: Līdzsvarotais mūsdienu latviešu valodas tekstu korpuss, tā nozīme gramatikas pētījumos (The Balanced Corpus of Modern Latvian and its role in grammar studies). In: Language: Meaning and Form 10, pp. 131–146, (2019).
9. Gruzitis, N., Pretkalnina, L., Saulite, B., Rituma, L., Nespore-Berzkalne, G., Znotins, A., Paikens, P.: 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), pp. 4506–4513, (2018a).
10. Spektors, A., Auzina, I., Dargis, R., Gruzitis, N., Paikens, P., Pretkalnina, L., Rituma, L. and Saulite, B.: Tezaurs.lv: the largest open lexical database for Latvian. In: Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC), pp. 2568–2571 (2016).
11. Paikens, P., Gruzitis, N., Rituma, L., Nespore, G., Lipskis, V., Pretkalnina, L., Spektors, A.: Enriching an Explanatory Dictionary with FrameNet and PropBank Corpus Examples. In: Proceedings of the 6th Biennial Conference on Electronic Lexicography (eLex), pp. 922–933, (2019).
12. Gruzitis N. and Znotins A.: Multilayer Corpus and Toolchain for Full-Stack NLU in Latvian. In: Proceedings of the CLARIN Annual Conference (2018b).
13. Znotins A. and Cirule E.: NLP-PIPE: Latvian NLP Tool Pipeline. In: Proceeding of Human Language Technologies – The Baltic Perspective, IOS Press (2018).