CLARIN

Common Language Resources and Technology Infrastructure

Tour de CLARIN
ESTONIA

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

ESTONIA

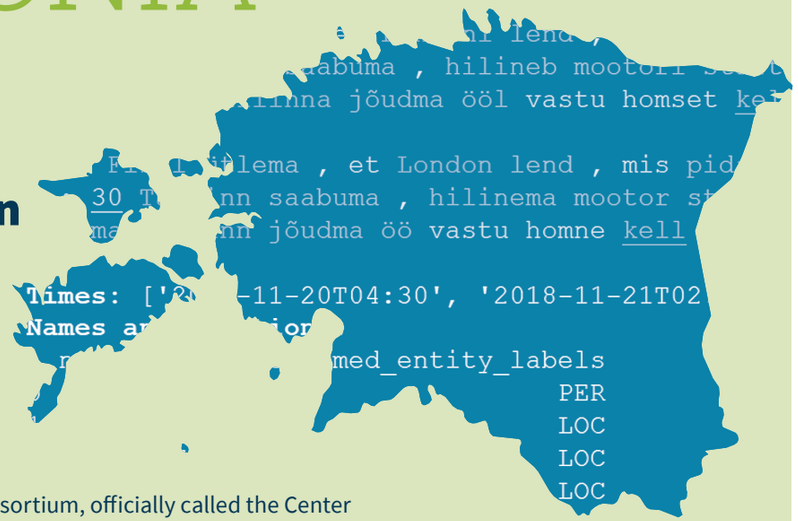
Foreword

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 Estonia 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

ESTONIA



Introduction

Written by **Kadri Vider**

The Estonian CLARIN consortium, officially called the Center of Estonian Language Resources (CELR), is a founding member of CLARIN ERIC.[1] It is a B-certified centre that involves four Estonian research institutions – the University of Tartu, Tallinn University of Technology, Institute of the Estonian Language and Estonian Literary Museum. The National Coordinator of CLARIN ERIC in Estonia is Kadri Vider. Aleksei Kelli, an Estonian legal expert, is the chair of the CLARIN Legal and Ethical Issues Committee (CLIC).

CELR provides access to Estonian language resources and language technology software (dictionaries, text and speech corpora, language databases, language software) for everyone working with digital language materials. The consortium also coordinates and organizes the registration and archiving of the resources as well as draws up necessary legal contracts and licences for different types of users.

The CELR LR META-SHARE registry currently contains 152 registered and published records in Estonian as well as 24 other languages with VLO-harvestable metadata, each of them having DataCite DOI. These comprise 59 lexical-conceptual resources, 66 corpora, 25 tools and services and two language descriptions. Among the many resources provided by CELR are several monolingual as well as multilingual dictionaries, such as the Dictionary of Standard Estonian and the dynamically updated English-Estonian Machine Translation Dictionary, all of which can be queried online.

[1] https://keeleressursid.ee/en/

The language tools include text and speech processing services, such as the Android Newsreader, which reads aloud the news articles in Estonian, and a comprehensive rule-based morphology toolkit which consists of modules for syllabification, paradigm recognition, morphological analysis and synthesis.

In addition to collecting, registering and archiving language resources, CELR also introduces the resources to potential users. The most successful outreach events in recent years were the workshops and tutorials about Estonian text corpora in KORP and lexical resources from the Institute of Estonian Language. Through the promotion of KORP usage we have reached out to the broader community of DH researchers in Estonia. Literary scholars have become interested in data analysis methods in literary studies, which has resulted in a collaborative project whose aim is to compile a corpus for literary studies. The collaboration is significant for the current stage of Estonian DH data digitisation, which needs to become more machine-analysable so that close-reading of digitised texts and a more sophisticated searching for tendencies in the bigger data collections become possible for the DH scholars.

The centre is also involved in the National Programme for Estonian Language Technology, whose aim to support the development of new language technologies for Estonian and associated initiatives. CELR is responsible for archiving the outcomes of the projects and introducing the resulting developments in language technology to the widest possible audience.

The Estonian CLARIN team

Tool | **EstNLTK**

Written by **Krista Liin**

When working with texts it is often difficult to extract the necessary information, especially if the texts are in a morphologically complex language such as Estonian. To find out which locations or individuals are mentioned in the text, you'd need to perform a full language processing workflow, from tokenization and finding base word forms up to detecting named entities. The next challenge is getting all those steps to work together.

EstNLTK, the Estonian Natural Language Toolkit, brings together previously developed Estonian NLP tools and resources in a common environment, making them easily accessible.[2] The toolkit is a set of Python libraries that has been created following the example of NLTK. It provides the following NLP components for the processing and analysis of the Estonian language: tokenization, spelling correction, pronunciation clues for stress and palatalization, the detection of paragraph, sentence and clause boundaries, verb chain tagging (such as *'oli läinud tooma'* - *'had gone to bring'*), morphological synthesis, named entity recognition, and a WordNet module.

EstNLTK is open source and is available for Linux, MacOS and Windows. Anaconda packages are available for researchers who want to use the toolchain as part of the Anaconda data science distribution. EstNLTK can also be used as a docker image, which allows researchers to skip the installation process and access the toolchain directly from a web browser in the Jupyter notebook (and copy any tutorials to work with). In addition to Python libraries, parts of EstNLTK can also be accessed as a webservice, or a WebLicht service.

The documentation that accompanies EstNLTK includes tutorials that cover several NLP tasks, from basic text operations such as finding base word forms (which is not very easy for a morphologically rich language such as Estonian) to more interesting tasks such as mapping the time expressions, recognising named entities or querying the Estonian WordNet and tagging words in text with their meanings and related synsets. Throughout the years people have created several morphological and syntactic analysers for Estonian, and EstNLTK has made an attempt to incorporate them all. To make it easier to work with large text corpora, EstNLTK has a database module that integrates with Elastic so you can use elasticsearch. There are also tutorials available on how to use the Estonian Reference corpus or Wikipedia data in EstNLTK.

[2] https://estnltk.github.io/estnltk/1.4.1/

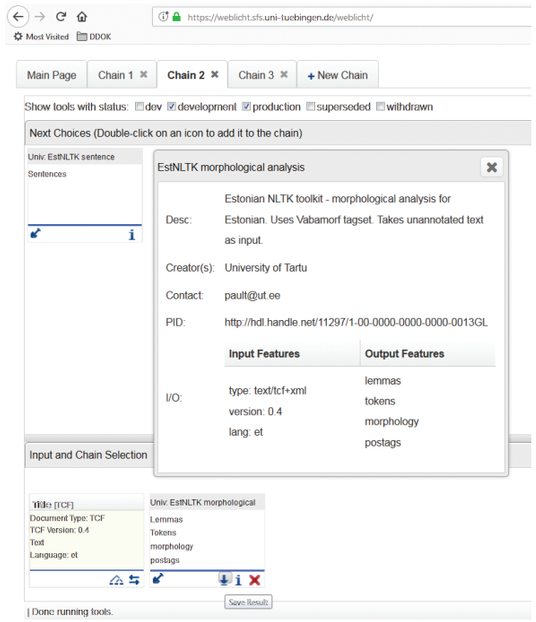


Figure 1: *Using EstNLTK in a WebLicht workflow for morphological analysis*

```
# default: use MaltParser model
text = Text('Maril oli väike tall')
text.tag_syntax()
# use EstCG instead
parser = VISLCG3Parser( vislcg_cmd='C:\\cg3\\bin\\vislcg3.exe' )
text2 = Text( 'Maril oli väike tall', syntactic_parser=parser )
text2.tag_syntax()
```

	English	word	POS	form	MaltP	head	EstCG	head
0	Mary	Maril	S	sg ad	@SUBJ	1	@ADVL	2
1	had/was	oli	V	s	ROOT	-1	@FMV	0
2	little	väike	A	sg n	@AN>	3	@AN>	4
3	lamb	tall	S	sg n	@PRD	1	@SUBJ	2

Figure 2: *Using the different integrated dependency syntax parsers*

Although the newer versions of EstNLTK allow researchers to choose among different tools in the language processing workflow, it is also possible to simply use the default options and get the end results. As can be seen in Figure 2, the default option for dependency syntax is the statistical MaltParser model. However, it is also possible to work with the rule-based Constraint Grammar parser EstCG instead.

```
Mark Fišel ütles , et Londoni lend , mis pidi täna hommikul kell
4 : 30 Tallinna saabuma , hilineb mootori starteri rikke tõttu
ning peaks Tallinna jõudma ööl vastu homset kell 02 : 20.

Mark Fišel ütlema , et London lend , mis pidama täna hommik kell
4 : 30 Tallinn saabuma , hilinema mootor starter rike tõttu ning
pidama Tallinn jõudma öö vastu homne kell 02 : 20.

Times: ['2018-11-20T04:30', '2018-11-21T02:20']
Names and locations:
  named_entities named_entity_labels
0     Mark Fišel                PER
1         London                LOC
2        Tallinn                LOC
3        Tallinn                LOC
```

Figure 3: *The NLP tasks performed by the EstNLTK toolkit - Part-of-Speech tagging and Named Entity recognition*

Figure 3 shows an example of the standard NLP tasks performed by EstNLTK. In this example, the toolkit is applied to the sentence *"Mark Fišel ütles, et Londoni lend, mis pidi täna hommikul kell 4:30 Tallinna saabuma, hilineb mootori starteri rikke tõttu ning peaks Tallinna jõudma ööl vastu homset kell 02:20"* ("Mark Fišel said that the flight from London, which was scheduled to land to Tallinn today morning at 4:30, is late due to an engine starter malfunction and is about to arrive to Tallinn tomorrow night at 02:00"). The text formatting chosen here shows lemmas with annotation for persons (red), locations (green), verbs (magenta), nouns (blue) and time expressions (underlined). The example was run on 2018-11-20, so the time values were calculated with respect to that date.

EstNLTK is highly interoperable and is used in several widely used applications, such as Feelingstream, which uses it in the processing of opinion mining, and the TEXTA toolkit, which takes advantage of the morphological analysis and NER for text mining. The toolkit is fairly robust, and it has also been used to work with non-contemporary texts, such as communal court minute books from the late 19th century, which did not follow modern spelling and were often written in local dialects. Kersti Lust from the National Archives of Estonia, Kadri Muischnek from the chair of language technology in University of Tartu and several of their colleagues worked together to make the collection of almost 3,000 texts from 22 different parishes browsable by annotating it with (standardised) lemmas and named entities, which makes it easier to study the interactions between different people mentioned in the minutes. Although manual correction is still needed, the automatic annotation worked very well, except for the Southern Estonian dialects, which differ a lot from contemporary Estonian, even syntactically.

EstNLTK has been developed under the NPELT programme by Sven Laur and colleagues.

Resource | **The Place Names Database (KNAB)**

Written by **Peeter Päll** and **Kairi Tamuri**

The Place Names Database of the Institute of the Estonian Language (KNAB) is a multilingual and multiscriptual systematic database of geographical names covering Estonia and other countries.[3] Its purpose is to facilitate the study and standardisation of geographical names by providing information on their history and modern use. It has been planned as a linguistically oriented database.

KNAB currently contains approximately 46,000 entries related to Estonia and 108,000 entries related to other countries. Estonian geographical names include the following:

- street names;
- names of populated places;
- names of former manor houses;
- farm names (partially);
- names of administrative units (both modern and historic);
- names of natural features (rivers, lakes, islands, bogs, capes etc.).

The geographical names of other countries cover at least 1st-level administrative divisions of each country, some autonomous administrative units of Russia (notably North Caucasus) and some minority names from other parts of the world (e.g.. Basque, Tibetan, Welsh). KNAB also collects exonyms or conventional foreign names from many languages of the world, which are also published separately.

Please note that the database is not an authoritative source of official names in Estonia. While some feature types (e.g. street names of Tallinn, names of populated places in Estonia) are fully covered, others might not be. The official register of Estonian place names is maintained by the Land Board of Estonia.

The database is continuously updated. By giving access to both modern and historic records, the database provides researchers with the possibility to identify name forms across different languages and study their diachronic development. Uniquely, the database also provides geographical names in different scripts; besides Latin, there are names in Burmese, Chinese, Cyrillic, Devanagari, Greek, Japanese, Mongolian, Tibetan and many other scripts, strictly encoded according to Unicode. In the case of foreign names, it should be borne in mind that the

[3] http://portaal.eki.ee/knab

data often reflect the de facto situation in a given country, so the names do not always correspond to the de iure status of certain regions. By contrast, country names follow the international naming conventions.

The users of the database include editors, translators, researchers, geographers and other specialists. The Estonian edition of the database (where the Estonian variants of the place names are listed as keywords) is used for example by the Estonian Wikipedia and the media when there is a need for more comprehensive listings than those given by dictionaries. In the English edition of the database, preference is given to local official names. The English data have been used in international research projects, which required multilingual name variants. For instance, in the Named Entity Recognition and Classification project of the Joint Research Centre of the European Commission, Pouliquen et al. (2006) used KNAB to develop a tool that recognizes geographical information in texts, which can be then visualized by tools such as Google Earth.

Figure 4: *Visualizing geographical information provided by KNAB in Google Earth with a tool developed by Pouliquen et al. (2006)*

Reference:

Pouliquen, B., Kimler, M., Steinberger, R., Ignat, C., Oellinger, T., Blackler, K., Fluart, F., Zaghouani, W., Widiger, A., Forslund., A-C., and Best, C. 2006. Geocoding Multilingual Texts: Recognition, Disambiguation and Visualisation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation* (LREC'06), 53–58.

Event | **Workshops at the Estonian Digital Humanities Conference 2017**

Written by **Kadri Vider** and **Olga Gerassimenko**

The Centre of Estonian Language Resources (CELR) actively reaches out to local researchers in order to address the recent challenges in digital humanities, map out possible solutions and offer personalized help. CELR specialists regularly offer workshops at a variety of events and conferences that bring together digital humanists and computational experts. The annual Estonian Digital Humanities conference, which is organized by multiple CELR partner organizations (including the Estonian Society for Digital Humanities, Estonian Literary Museum, Centre of Excellence in Estonian Studies and Wikimedia Estonia), is a key event attended by both Estonian and European scholars in DH and a perfect place to address the challenging issues related to the interaction of scholars within different fields.

At the 2017 conference "Open licences, open content, open data: tools for developing digital humanities", which took place between November 1 and 3 in 2017 at the Estonian National Museum, Aleksei Kelli held the workshop "Copyright and cultural heritage".[4] The workshop focused on the interaction of intellectual property (IP) and cultural heritage, with special attention given to copyright and related rights. Professor Kelli presented issues related to the free use of heritage works by public archives, museums or libraries, quotation rights and the right to use copyrighted materials for educational and research purposes.

Workshop participants were invited in advance to send a description of a (potentially) problematic case in their research related to copyright. For instance, the attending etymologists and folklorists raised the following issue related to the ambiguous legal nature of folklore. Although folklore is not inherently copyrighted, the recordings of folk songs or the retellings of stories do get protected under copyright law, in that the contributors who share folk stories or sing old folk songs retain the rights to their performance. However, in many cases, such performances are very old and often the folklorist who recorded them did not explicitly ask the contributors for their consent, as there was no such legal requirement in the past. Consequently, it is often unclear whether digitized collections of folklore can be made publicly available, although research exceptions in copyright regulations make them available for academics.

[4] https://dh.org.ee/events/dhe2017/programme/

Aleksei Kelli leading the workshop “Copyright and Cultural Heritage”

Another workshop CELR presented at this conference was the hands-on demonstration “Language annotation workflows in your browser” by Krista Liin. Estonian and foreign participants could try annotating their texts in the workflow managers Keeleliin (for Estonian) and Weblicht (built by CLARIN-D and available for several European languages) and to learn the possibilities of automatic annotation accessible through a web-browser and how to use the annotated texts in their research. Such browser-driven annotation was welcomed by the participants, all of whom worked with morphologically rich free word order languages such as Estonian. After lemmatizing different Estonian texts (i.e., standard language data and spoken language/data) with Keeleliin, participants created simple workflows and experimented with Weblicht’s easy-mode chains for tokenising and parsing texts in English or German. Some of the participants also familiarised themselves with the open framework for interoperable NLP web services Galaxy, the multilingual text similarity analysis system WebSty, some NLP and visualization tools in Textimager, and UDpipe, a parsing pipeline for CONLL-U texts.

Interview | **Marin Laak**

Marin Laak is a senior researcher and principal investigator of the Estonian Literary Museum. She was one of the developers of the Estonian cultural history web portal “Kreutzwald’s Century” that gives access to a vast amount of the language’s digitised literary legacy.

Could you briefly tell us about your academic background?

<

I studied at the University of Tartu and got my bachelor’s degree in Estonian language and literature and then obtained my M.A. and Ph.D. degrees in literary studies. I have always been interested in collaboration with linguists, since I believe that literary studies are not possible without paying close attention to the language, which is clearly both the material and the base for the creation of literature. Throughout my career as a literary scholar I have been interested in large content-based models and literary environments, which I explored in my doctoral dissertation called “Non-linear Models of Literary History: The Problems of Text and Context in the Digital Environment”. I worked on the first Estonian project that developed a hyper-text environment which linked various types of texts together. While the links were created manually at first, in the next project we developed software to generate them automatically, which required close work with the textual resources. The issue of accessibility and usability in a wide range of scientific and educational purposes has always been one of my priorities. In this sense, the collaboration with the Estonian CLARIN consortium has been a dream project for me.

>

How did you get involved with the Estonian CLARIN consortium?

<

The Estonian Literary Museum has been a member of the Estonian CLARIN consortium since 2016 and they have always supported my research. Together we have created a small and efficient working group that is developing the first tagged literary corpus for Estonian. I am very grateful to the team and the synergy that we have established working together. I have been involved in the Digital Humanities since 1997 when my old friend Neeme Kahusk (now a member of the CLARIN Estonia staff) advised me to participate in the call for the Tiigrihüpe (Est. Tiger's Leap) project proposals. This was an initiative of the Estonian government that started in 1997 and heavily invested into the development and expansion of computer skills and network infrastructures in Estonia, with a particular emphasis on education. I was then the only non-linguist in the team, and in a couple of years we put together an extensive corpus of literary criticism texts, which was linked with textual interconnections to a larger hypertext network.

Having observed the work of linguists since the 1990s, I have witnessed a huge qualitative leap in their research. The potential of textual resources that are tagged morphologically and syntactically has grown significantly, and has led to countless new possibilities for contextual research. For this reason, I believe that computational linguists should strive to make their tools more helpful and user-friendly for literary scholars. To make this possible, we first need to overcome the challenges set by the diachronic changes of language.

>

You are one of the authors of the Estonian cultural history web portal Kreutzwald's Century. Why is this portal important for Digital Humanities in Estonia?

<

Kreutzwald's Century is a unique project that is named after a literary exhibition dedicated to the cultural legacy of the Estonian writer and publicist Friedrich Reinhold Kreutzwald.[5] The portal was created as a non-linear environment model for new literary history studies and is actually the starting point of the digitisation of all the books ever published in Estonian. It is an immense leap forward in the context of the massive digitisation of cultural legacy that is taking place nowadays. Currently, the portal gives access to 268 author biographies, more than 10,000 photos and more than 2,000 event descriptions based on newspaper material. More than 300 older fictional works in Estonian are accessible in the e-pub format, and the publicly available text corpora contain 13,808 pages or 24,859,487 characters. The portal is widely used in education: in 2018, we registered around a million clicks monthly (which is almost comparable to the Estonian population, which is slightly over a million people) and around 2,000 unique visitors. We have manually controlled and corrected the optical character recognition (in spite of the large amount of work this entailed). As a result, the portal is the biggest and most accurate literary textual resource portal in Estonia.

>

How can corpus linguistics be applied to the research of cultural and literary history? Why is textual annotation relevant for literary studies? How does CLARIN Estonia help researchers in non-technical fields like literary theory to apply computational methodologies?

<

With the support of the Estonian government, all types of cultural legacy (printed books, archival documents, etc.) are being massively digitised and made accessible as open data. Consequently, the quantity of texts is becoming exponentially larger and larger. However, the methods that literary scholars use are still the same as those from decades ago – they are mostly based on close reading, which is a time-consuming method with a narrow focus and a lot of limitations for large-scope research. It is not a local problem, as I see the same tendencies at the international level. Literary scholars worldwide already have access to large amounts of data and create new resources themselves, but our vision for textual resources and the possibilities of their usage has not yet reached the level of computer linguists. Linguists have worked with morphologically, syntactically and even semantically tagged resources for decades, and have developed new annotation layers and new research methods to meet new opportunities. That is exactly the challenge literary scholars are facing now. We need to work out proper annotation layers and tagsets to address the content-driven research questions that are in our focus. We need to address the challenges of having simultaneous access to large collections of data where we can, by relying on linguistic information, trace the connections between texts and authors, the developments of literary means, changes in poetics, and so forth. We need the expertise of linguists to develop the theory and practice of annotation. At the same time, we need to learn how to pose new research questions and solve research problems in literary studies and humanities in the digital framework. We already strive to make the materials we work with broadly accessible, and our next step is to enhance their quality for scientific usage.

>

[5] http://www.folklore.ee/dh/en/dhe_2013/mikkel_laak/

Could you describe how the Estonian Literary Museum collaborates with CLARIN Estonia on the digitization of textual cultural heritage and its transformation into machine-readable research data?

<

As a pilot project, we have put together a morphologically tagged corpus out of approximately thousand pages of handwritten letters by two Estonian writers, Johannes Semper and Johannes Barbarus, from 1910 to 1940. The corpus is publicly available via the Estonian interface of the corpus query system KORP. Our work is described in our DHN2019 paper, titled "Literary Studies Meet Corpus Linguistics: Estonian Pilot Project of Private Letters in KORP" (authors Marin Laak and Kaarel Veskis from Estonian Literary Museum; Olga Gerassimenko, Neeme Kahusk and Kadri Vider from the University of Tartu). We are going to use this corpus to test the possibilities that linguistic annotation opens for the studies of literary content and literary history. Together with our international colleagues, we will discuss how research questions in literary studies relate to KORP collections and the possible adaptations of KORP functionalities for literary scholars at DHN2019, as well as at the Research Data and Humanities conference in 2019.

Estonia is expecting an explosive growth of digital heritage and textual resources. Preparations for massive digitisation of cultural heritage started in 2018 as part of the national programme, and the creation of different digital resources is the current priority of Estonian memory institutions.

Additionally, our institution already has a lot of digitised contemporary data for life-writing studies. The crucial question for us is how to bridge the gap between the research possibilities offered by contemporary language technologies on the one hand and the ever-increasing volumes of texts and other digital data produced by memory institutions on the other. We therefore need to rethink the approach to defining the empirical object in literary studies in general and proposing new research questions. The ability to compare text strategies, rhetorical and stylistic patterns in literary, religious and political text corpora should give us new insights into the way ideology, rhetoric and identity presentations interact.

To do this, we have to learn to search for not only linguistic patterns but for the cultural threads in literary texts. Such threads show how ideas and thoughts travel from one text to another and from one period to the next. We need to unite the expertise of literary scholars, linguists and computational experts to make this possible, and we need to organize our textual resources wisely according to their genre, creation period and other metadata.

Thankfully, the Estonian CLARIN centre offers the needed expertise for transforming our data into valuable and reliable text resources, which was already achieved in the case of the Kreutzwald's Century materials and is currently taking place with the Corpus of Estonian Literary Criticism.

My collaboration with CLARIN Estonia is a continuation of my work in the European Union East project CULTOS: Cultural Units of Learning Tools and Services. I lead the project "Formal and informal networks of literature based on sources of cultural history" and I believe that the new technical opportunities offered by the consortium are helping us advance our research. Our interdisciplinary practical work, which has involved the preparation of a literary corpus for KORP, has been a synergetic team effort, and I have the best hopes for our future work together.

>

Are there any tools and resources provided by the Estonian consortium that you use in your work and you would like to single out as inspiring for other Digital Humanities researchers?

<

The tool I am currently fascinated by is the corpus query system KORP. We learned a lot about the KORP functionalities, such as flexible search options and statistics. We would love to promote the research possibilities with KORP among our colleagues and adapt KORP functionalities for literary studies. I would love to work on the further development of KORP together with the international community.

>

In your opinion, how can research infrastructures like CLARIN help museums (staff and visitors alike)?

<

The Estonian Literary Museum is not really a visitor-type museum; rather, it functions as a leading memory institution and research centre. Along with the Centre of Excellence in Estonian Studies, we will benefit from our partnership with CLARIN by being able to rely on CLARIN's ability to create, maintain and enhance the usability of data collections.

>

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

Mark Fišel ütles , et Londoni lend , mis pidi täna hommikul kell 4 : 30 Tallinna saabuma , hilineb mootori starteri rikke tõttu ning peaks Tallinna jõudma ööl vastu homset kell 02 : 20.

Mark Fišel ütlema , et London lend , mis pidama täna hommik kell 4 : 30 Tallinn saabuma , hilinema mootor starter rike tõttu ning pidama Tallinn jõudma öö vastu homne kell 02 : 20.

Times: ['2018-11-20T04:30', '2018-11-21T02:20']
Names and locations:

	named_entities	named_entity_labels
0	Mark Fisel	PER
1	London	LOC
2	Tallinn	LOC
3	Tallinn	LOC

CLARIN 2019