CLARIN

Common Language Resources and Technology Infrastructure

Tour de CLARIN

VOLUME ONE

Edited by Darja Fišer and Jakob Lenardič

Table of contents

Foreword

Tour de CLARIN is an initiative started by CLARIN ERIC in 2016 that has been periodically highlighting prominent user involvement activities of CLARIN national consortia in the form of blog posts published on the CLARIN webpage, disseminated through the CLARIN news flash and on social media. By focusing a different national consortium every two months and showcasing their outstanding language resources, text processing tools, user involvement events and researchers, we have been aiming to increase the visibility of the various consortia, reveal the richness of the CLARIN landscape, and display the full range of activities throughout the network that can not only inform and inspire other consortia, but also 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.

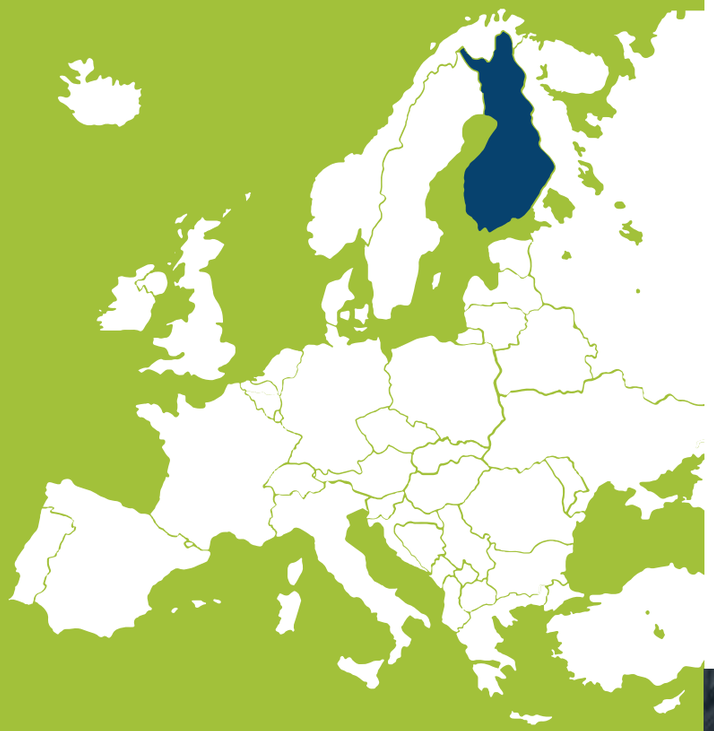
In the two years we have been running the initiative, and having visited nearly half of all the CLARIN member countries, we can say that Tour de CLARIN has proved to be one of the flagship user involvement initiatives by CLARIN ERIC; highly valuable for our network and incredibly popular with our readers. This is why have decided to collect the blog posts in a printed volume. The first volume presents all the nine countries which we have visited so far: Finland, Sweden, Austria, the Netherlands, Poland, Belgium, the Czech Republic, Greece and Lithuania. While we did start the initiative with Finland which had the largest number of user involvement events in 2016, the order of the rest of the visits took into account a combination of factors, such as good geographic distribution of the highlighted countries, a balanced focus on older as well as more recent CLARIN members, and the availability of the consortium members and researchers to provide input for the posts. The volume is organized in the same way as our on-line initiative: each chapter focuses on a CLARIN member country and in five sections presents the members of the consortium and their work, one of their key resources, an outstanding tool, a successful event for the researchers and students in their network, and an interview with a renowned researcher from the digital humanities or social sciences who has successfully used the consortium's infrastructure in their research. Where relevant, the texts have been updated with the most recent information on the research projects, events, publications and versions of the highlighted tools and resources.

Tour de CLARIN would not have been possible without the contributions and dedication of the national user involvement representatives and national coordinators: Mietta Lennes and Krister Lindén from Finland, Eira Brandby and Lars Borin from Sweden, Tanja Wissik and Karlheinz Mörth from Austria, Arjan Van Hessen and Jan Odijk from the Netherlands, Jan Wieczorek and Maciej Piasecki from Poland, Vincent Vandgehinste and Griet Depoorter from DLU/Flanders, Barbora Vidová Hladká and Eva Hajičová from the Czech Republic, Maria Gavriilidou and Stelios Piperidis from Greece, and Jolanta Kovalevskaite and Jurgita Vaičenonienė from Lithuania. We would also like to thank all the researchers who have kindly agreed to be interviewed for their time and invaluable insights: Tommi Jantunen, Maria Ågren, Stephan Procházka, Melvin Wevers, Maciej Maryl, Cora Pots, Radim Hladík, Vassiliki Georgiadou and Erika Rimkutė.

Tour de CLARIN will continue to visit CLARIN member countries and present their success stories online as well as in future printed volumes. In addition to presenting the work of national consortia, we also plan to expand the initiative to CLARIN K-centres in order to highlight their work and make them more visible to a wider audience.

Darja Fišer and Jakob Lenardič
Ljubljana, Slovenia
November 2018

Finland



Written by Darja Fišer and Jakob Lenardič

The Finnish national consortium FIN-CLARIN[1] has been a CLARIN member since 2015. The members of the consortium are the University of Helsinki, the University of Eastern Finland, the University of Jyväskylä, the University of Oulu, the University of Tampere, the University of Turku, the University of Vaasa, the Institute for the Languages of Finland, the Helsinki Institute of Technology and the IT Centre for Science (CSC). The national coordinator for FIN-CLARIN is Research Director Krister Lindén.

FIN-CLARIN has been actively engaged in developing tools and resources that have become a staple of Finnish researchers working with language data. Through the Language Bank of Finland, which is a certified CLARIN B-Centre, researchers can access dozens of Finnish corpora, which are in most cases available through online interfaces such as Korp.

A flagship resource provided by the Finnish consortium is the Suomi 24 Sentences Corpus, a corpus that compiles texts from discussion forums of the popular Suomi24 online networking website. The data from the corpus is currently being analysed in the framework of the Citizen Mindscapes project, which seeks to uncover "trends and shifts in attitudes in connection to societal phenomena" in Finland, thus making the corpus an extremely important resource that highlights how corpus-based linguistics can lead to a greater understanding of society at large. Read more about this corpus on page 7.

The Finnish consortium is actively engaged with ground-breaking researchers working in Digital Humanities and Social Sciences who make use of the consortium's resources and tools. The Language Bank hosts a "Researcher of the Month" archive, intended to highlight both the work of the prominent researchers and the tools and resources of potential use to researchers.

In 2016, Finland organised 22 user involvement events. A very successful event was the Roadshow organised to celebrate 20th anniversary of the Language Bank of Finland. It consisted of a series of seminars at all the member organisations of the FIN-CLARIN consortium.

[1] https://www.kielipankki.fi/organization/

The Language Bank of Finland hosts a variety of language tools, for instance the following publicly available ones:

- Finnish Parse, which is a powerful dependency parser that is capable of tokenisation, sentence splitting, morpho-syntactic tagging and parsing, and can be applied to plain Finnish text with extremely high accuracy
- Aalto-ASR, a continuous speech recogniser that can handle a large amount of Finnish vocabulary
- the Helsinki Finite-State Transducer Technology that provides software for morphological analyses of various European languages, and
- the Proto-Indo-European Lexicon, which acts as a generative etymological dictionary, providing data on word origins and historical changes for the hundred most ancient Indo-European languages.

FIN-CLARIN Team | Back row: Atro Voutilainen, Senka Drobac, Pekka Kauppinen. Middle row: Maria Palolahti, Erik Axelson, Jyrki Niemi; front row: Krister Lindén (Research Director), Mietta Lennes, Jussi Piitulainen. Not in the photo: Tero Aalto, Imre Bartis, Ute Dieckmann, Sam Hardwick, Martin Matthiesen.

Helsinki, Finland | photo by Alexandr Bormotin | Unsplash

AaltoASR

Written by Darja Fišer and Jakob Lenardič

The AaltoASR project,[2] which is led by Professor Mikko Kurimo at Aalto University, focuses on the development of an Automatic Speech Recognition system that is able to transcribe spoken Finnish language with a very high accuracy rate. The system, which started as a relatively simple spoken-language recogniser in the 1980s that was at first capable of handling around 1,000 Finnish words, is today a complex piece of software that can recognise and transcribe not only isolated words but also spontaneous speech. The AaltoASR system comprises of complex procedures that accurately transform audio signals into linguistically-modelled speech units on the basis of a complex network of probabilistic distributions, thus making the system easily adaptable to various domains and styles.

By focusing on complex agglutinative languages such as Finnish and Estonian and under-resourced languages such as the Sami ones, the AaltoASR team continues to make groundbreaking progress in the development of a successful large vocabulary speech recogniser that is able to tackle complex inflectional and compounding systems, which otherwise make it difficult to perform rule-based morphology analysis and, by extension, speech recognition. The AaltoASR tool is open source, and the developer version can be found on the tool's GitHub page. AaltoASR is available for research use via the Language Bank. ASR systems built on top of AaltoASR tools are also used by companies for subtitling TV broadcasts in Finland and Sweden.

The most recent papers by Prof. Kurimo and his colleagues include:

Mansikkaniemi, A., Smit, P., and Kurimo, M. (2017). Automatic Construction of the Finnish Parliament Speech Corpus. In Interspeech 2017. http://dx.doi.org/10.21437/Interspeech.2017-1115.

Smit, P., Virpioja, S., and Kurimo, M. (2017). Improved subword modeling for WFST-based speech recognition. In Interspeech 2017. http://dx.doi.org/10.21437/Interspeech.2017-103.

[2] https://github.com/aalto-speech

The Suomi24 Corpus

Written by Darja Fišer and Jakob Lenardič

The Suomi24 corpus[3] is a comprehensive collection of texts from discussion forums of Suomi24, which is Finland's largest and most popular social media website and is used by 86% of Finns every month. The corpus contains more than 2.6 million tokens of texts from 2001 to 2016 and is tokenised and morpho-syntactically tagged with the Turku Dependency Parser. A version of the corpus where the sentences are scrambled is publicly available through the web interface Korp under the CLARIN ACA - NC licence, while researchers who have a username and a password can download the entire corpus in the VRT format.

The corpus is used by researchers working in the Citizen Mindscapes project (2016–2019), funded by the Academy of Finland. The aim of the project is a far-reaching socio-political and linguistic analysis of the everyday discourse that is part and parcel of the Finnish society. By applying a wide range of quantitative and qualitative methods such as statistical data analysis and thematic interviews and by making use of advanced language tools to process the data within the corpus, Citizen Mindscapes researchers, who are led by Professors Jussi Pakkasvirta and Krista Lagus, seek not only to uncover the current societal and political trends in Finland, but also pinpoint those features of the online discourse that may very well hint at the prospective evolution of the Finnish society as a whole.

To make the presentation of the complex data within the Suomi24 corpus as optimal as possible for socio-political analysis, the Citizen Mindscapes team is developing the Social Thermometer. This novel visualisation method helps researchers detect deep-rooted views related to complex issues such as nationalism, which often begin in and are shaped by discussions on the Internet.

The Citizen Mindscapes project is thus a pivotal multidisciplinary endeavour that has brought together and established long-term collaborations between researchers from diverse fields such as natural language processing, sociolinguistics, sociology and political studies. Wishing to promote open science and open data and thereby support the development of novel approaches in social sciences and NLP, the researchers in Citizen Mindscapes plan to make their data sets and tools available within the Language Bank of Finland in collaboration with FIN-CLARIN and the Centre for Science.

Follow Citizen Mindscapes on Twitter: @mindscapes24

Researcher and Helsinki Challenge semi-finalist Krista Lagus with the Citizen Mindscapes research project team (photo by Linda Tammisto).

[3] http://urn.fi/urn:nbn:fi:lb-2017021506

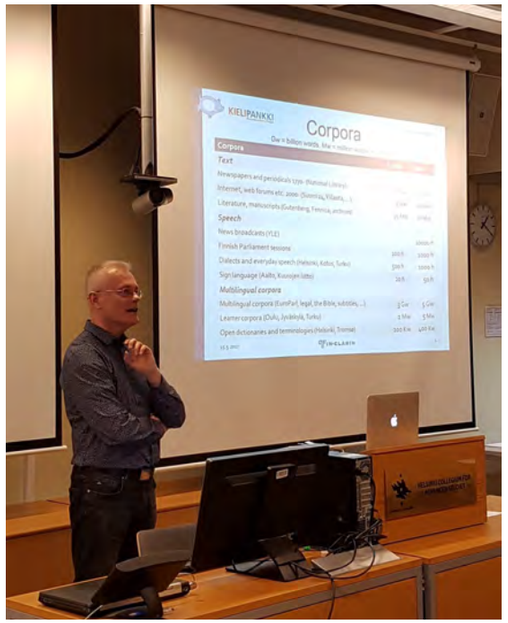
The Language Bank of Finland's Roadshows

Written by Darja Fišer and Jakob Lenardič

The Finnish consortium FIN-CLARIN is one of the most active ones in the CLARIN network in the field of user involvement. In 2016, they organised the greatest number of user involvement events. Among these, their roadshow events, which were organised to celebrate the 20th anniversary of the founding of the Language Bank of Finland, stand out in particular. Through the course of the roadshows, FIN-CLARIN presented the tools and resources of the Language Bank at all the Finnish universities that are members of the Finnish consortium, beginning with the University of Turku on 28 April 2016 and ending with the University of Jyväskylä on 15 November 2016. Each presentation began with a demonstration of the Korp interface and of various ways in which researchers can use it to search through the text and speech corpora that are available in the Language Bank. The presentation materials are available both in English and Finnish.

To conclude each event, presentations were given by researchers who work at the hosting universities and who have used the tools and resources of the Language Bank. For instance, when the roadshow stopped at Aalto University on 27 September, André Mansikkamäki presented speech recognition tools developed within the AaltoASR project, which we highlighted on page 6. When the roadshow concluded for 2016 at the University of Jyväskylä, Tommi Jantunen, whose interview can be read on page 9, gave a presentation on the Finnish Sign Language and Jarmo Jantunen on the usage of the Suomi24 corpus, which we present on page 7.

Because of the success of the 2016 events, FIN-CLARIN continued the roadshows in 2017. After a brief pause in 2018 when FIN-CLARIN was busy carrying out other UI initiatives, the roadshows are expected to continue in 2019 with a focus on introducing newly developed tools and corpora.



Speaker: Krister Lindén, National Coordinator of FIN-CLARIN at the Helsinki Collegium of Advanced Studies, 15 May 2017 (photo by Mietta Lennes).

Tommi Jantunen

Tommi Jantunen is a linguist specialising in the Finnish Sign Language working at the University of Jyväskylä. The following interview took place via Skype on Tuesday, 23 May 2017 and was conducted and transcribed by Jakob Lenardič and edited by Darja Fišer.

1. Could you tell us a little bit about yourself, your background and your current work?

My name is Tommi Jantunen and I am an academic research fellow at the Department of Language and Communication Studies at the University of Jyväskylä. I have an MA in General Linguistics from the University of Helsinki and a PhD in Finnish Sign Language from the University of Jyväskylä. I am currently running the last year of my five-year-long project ProGram, which is funded by the Academy of Finland and in which my colleagues and I are investigating the grammar and prosody of the Finnish Sign Language. At the University of Jyväskylä, the Finnish Sign Language is one of the main fields that people can get a degree in, and I was one of the first people to obtain a PhD degree in Finnish Sign Language at our university almost ten years ago in 2008. I don't think that there are many universities in Europe and even in the world in general where one can fully focus on studying sign languages as a major subject, on par with more traditional fields like Finnish or English studies, so it's really excellent that our university offers this opportunity.

2. How did you hear about CLARIN and how did you get involved?

I think I knew about CLARIN ERIC even before I became directly involved with our national consortium. This is probably so because all Finnish universities are part of the FIN-CLARIN network so if you are a researcher working with languages it's almost impossible not to hear about CLARIN. I think I became involved a few years ago at the start of our current project ProGram. We needed to publish the data that we were starting to work on, so I contacted our FIN-CLARIN UI representative Mietta Lennes from the University of Helsinki and she was happy to get our project involved with the consortium.

3. How has CLARIN influenced your way of working and how was this received in your research community?

I think the easiest way to approach this question is from the following point of view. If you're a researcher and you are in the process of applying for research funding, one of the prerequisites for obtaining funding is that you draft a plan of how the data that you are working on can be related to and included within research infrastructures such as FIN-CLARIN. In other words, it is the funding agencies that require researchers to connect their data with a ready-made infrastructure so that the data become openly available. Related to this, FIN-CLARIN has proven itself to be an exceptionally good collaborator – whenever we write new applications for research funding, it is extremely easy for us to connect our work with the FIN-CLARIN and CLARIN ERIC infrastructures.

When it comes to sign language, we unfortunately can't use much of the existing services within the infrastructures because of the specificity of our field, which requires specialised tools and resources. However, we are building our own tools and services in collaboration with FIN-CLARIN, which we of course plan to make available within the FIN-CLARIN repository – that is, the Language Bank of Finland. For instance, my colleagues are currently compiling a very exhaustive Finnish Sign Language corpus, which is planned to be made available through FIN-CLARIN and which will mark the end of the project.

[4] http://users.jyu.fi/~tojantun/ProGram/ProGram.html

4. Could you describe the status of the Finnish Sign Language?

I think its status is relatively good, especially in comparison with some other countries. The Finnish Sign Language is recognised in the Finnish Constitution as a national minority language, and there's a sign language law that clarifies certain linguistic etc. rights for sign language users. The Finnish Sign Language came into the spotlight when a deaf rap artists called Signmark[5] won second place in Finland's national qualifications for Eurovision in 2009. Apart from the Finnish Sign Language, there's also the Finland-Swedish Sign Language, which is a kind of minority-within-a-minority language spoken by relatively few people in the Swedish speaking areas of Finland. Linguistically, these two sign languages are very close to each other. Ultimately, it is a political decision that governs the question what is a specific language and what isn't.

5. To what extent in your opinion is the sign language research community benefiting from digital tools and resources, such as the ones provided by CLARIN infrastructure?

I think one of the most important, and also the most invisible, aspects of what CLARIN ERIC and the national consortia have been doing is the work related to standardisation and the resolution of metadata issues. I don't know if we would even be able to do our corpus work if it weren't for the standards created in FIN-CLARIN. The LAT platform is also extremely important, as it allows us to publish our visual data. For instance, on the LAT platform we have already published a richly-annotated video corpus of the Finnish Sign Language. This corpus comprises a rather small sample that is much more deeply annotated than our final exhaustive corpus is going to be, and I also use this corpus to showcase Finnish Sign Language data to my students and colleagues.

6. What kind of specific methodological and technical challenges does a researcher working on sign language face with respect to the available infrastructure?

The biggest challenge is related to the videos, which serve as basically the only way you can record and compile sign language data. On the one hand, working with videos is challenging from a technical perspective; it is difficult to search through visual data and they take up a lot of storage space. On the other hand, videos raise many privacy and legal issues, and we have to be extremely careful when it comes to getting informed consent. One of the problems we had in connection with this is that there was very little information about the necessary steps we needed to take in order to fulfil all the legal requirements. Additionally, it has been my experience that annotating the sign language data is extremely time consuming, more so than typical speech data.

7. How would you recommend your colleagues to get involved with CLARIN and start using the available infrastructure?

My advice to all researchers who are interested in working with languages is that they contact their respective UI representatives. In Finland, Mietta Lennes has been doing an excellent job in promoting FIN-CLARIN and CLARIN-ERIC. She has taken part of practically every linguistic conference in Finland where she has kept FIN-CLARIN very visible. In terms of visibility, I think we introduce CLARIN even to our BA students, so a researcher in Finland can't really avoid knowing something about CLARIN. I think a more important question here is when is it that a researcher needs to know something about CLARIN. This is at the point when you start doing your own research and when you apply for funding – as I've said, the funding agencies require that the data obtained in your research get connected with the existing infrastructures, and FIN-CLARIN, as well as CLARIN ERIC in general, provides an excellent infrastructure for this.

[5] https://en.wikipedia.org/wiki/Signmark

8. What resources, tools and services from CLARIN would you recommend to your colleagues? What would you recommend CLARIN to do in order to attract more researchers from your community?

I'm mostly familiar with the previously mentioned LAT platform and the ELAN tools, which we have been using to annotate our visual data. It is also great that such tools are available as web services, which makes them very easy to use. As to the second question, I think that FIN-CLARIN has been doing an exemplary job already, so I think it's impossible to recommend anything in terms of improvement. Last year, the roadshow event organised by FIN-CLARIN, which took place in autumn, also reached our university and I gave a presentation on Finnish Sign Language then.

9. What's your vision for CLARIN 10 years from now?

I would like to see more and more video materials and tools related to sign language processing in the repository.

10. Describe CLARIN in three words.

Internationality, openness and user-friendliness!

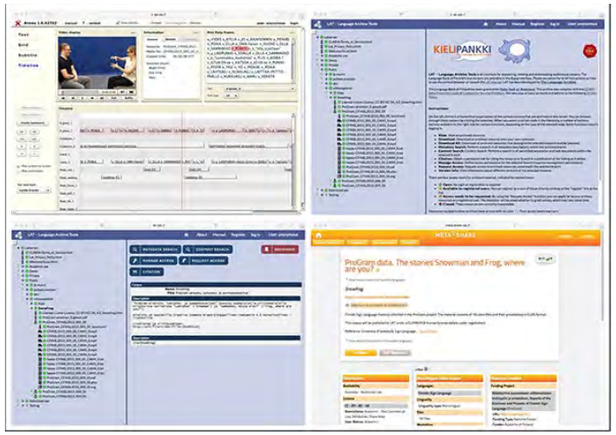
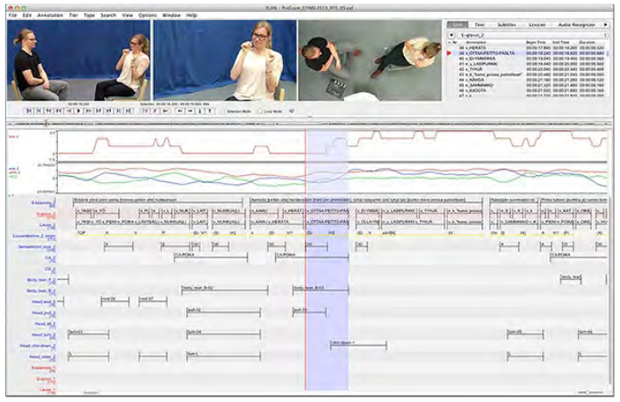
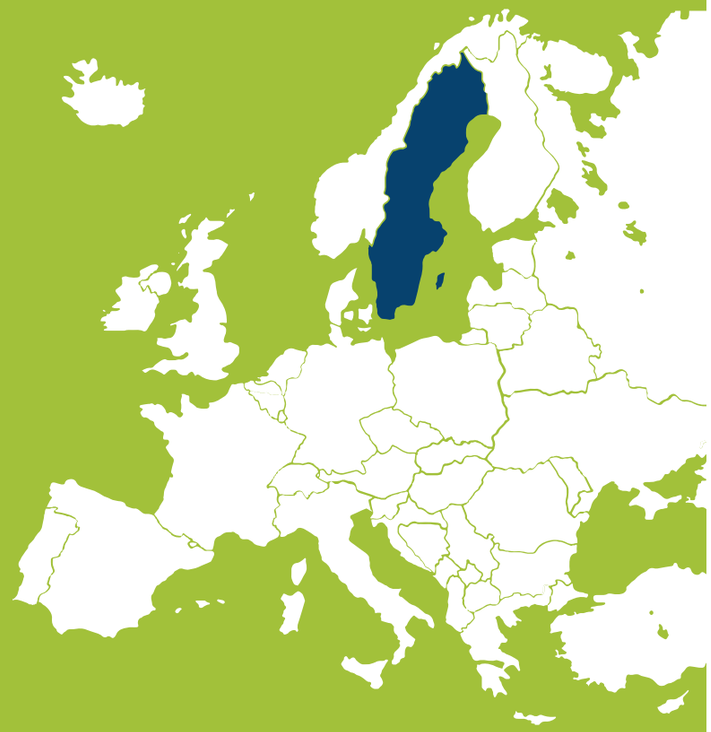


Figure 1: The ProGram corpus accessed via the FIN-CLARIN infrastructure.

Figure 2: An annotated video corpus of the Finnish Sign Language.



Sweden



Written by Darja Fišer and Jakob Lenardič

The Swedish consortium Swe-Clarin,[6] which has been a member of CLARIN ERIC since 2014, is a collaboration between the national archive Riksarkivet, the Swedish National Data Service, the Swedish language council Språkrådet, the KTH School of Computer Science and Communication and a variety of language technology research units at five universities – the Department of Computer and Information Science at Linköping University, the Humanities Lab at Lund University, Språkbanken (the Swedish Language Bank) at the University of Gothenburg, the Department of Linguistics at Stockholm University, and the Computational Linguistics Group at Uppsala University. The national coordinator of Swe-Clarinis Lars Borin, professor of natural language processing at the University of Gothenburg and co-director of Språkbanken.

The coordinating centre of Swe-Clarin is Språkbanken (the Swedish Language Bank), a major national and international research centre on computational approaches to language in Sweden, established already in the 1970s, which provides researchers with access to language resources, including an extremely wide range of Swedish texts, as well as state-of-the-art computational tools for the processing, compilation and linguistic analysis of corpora. The rapidly-increasing number of corpora, which are in the majority of cases available for download in standard formats, not only comprise a comprehensive collection of contemporary Swedish texts representing a wide variety of formal and informal discourse produced both in Sweden and Finland, where Swedish is an official language, but also include historical texts from most periods of written Swedish.

[6] https://spraakbanken.gu.se/eng

Related to the corpora are the tools of Språkbanken; for instance, the corpus infrastructure Korp, used for accessing both the above mentioned corpora, the Finnish corpora made available by FIN-CLARIN in the Language Bank of Finland, the Saami corpora provided by the Norwegian CLARIN Giellatekno node in Tromsø, and Estonian corpora available through the Estonian CLARIN ERIC centre, or the annotation tool Sparv, presenting a web-based interface to the Korp annotation toolchain, offering part-of-speech tagging, compound analysis, word sense disambiguation, named entity recognition and dependency parsing of Swedish text. Additionally, Språkbanken researchers are working in a great number of research projects – one such endeavour is the research program "Towards a knowledge-based culturomics", among whose goals is "to advance the state of the art in language technology resources and methods for semantic processing of Swedish text, in order to provide researchers and others with more sophisticated tools for working with the information contained in large volumes of digitised text, by, for instance, being able to correlate and compare the content of texts and text passages on a large scale".

Swe-Clarin has also organised successful user involvement events. One such event was the Second National Swe-Clarin workshop held in connection with the Swedish Language Technology Conference in November 2016, where two invited presentations were given, one on the text mining project BiographyNet and the other on the impact language technology has on scholarship in the humanities, followed by a poster session featuring Swedish CLARIN-supported research.

Lars Borin, National Coordinator of Swe-Clarin.

Korp

Written by Darja Fišer and Jakob Lenardič

A concordancer is one of the key tools of a language resource provider, as it serves as the main entry point to language in context. One of the best known and widely used concordancers is that provided by Swe-Clarin's Korp.[7] A versatile and user-friendly tool, it is the main corpus infrastructure of Språkbanken and is used extensively by the Swedish and Finnish consortia, as well as in an Estonian and a Norwegian CLARIN centre. Through Korp, researchers can access some of the consortia's most important language resources, such as Swe-Clarin's Riksdagsen öppna data corpus (see page 16) and FIN-CLARIN's Suomi24 corpus (see page 7).

Korp has been developed by a team of about eight people at Språkbanken at the University of Gothenburg and consists of three components:

- the Korp corpus pipeline, which is used for the import, annotation and export of corpora;
- the Korp backend, which consists of a series of web services used for searching and retrieving both the corpora and their associated annotations and metadata; and
- the Korp frontend, which is the graphical user interface communicating with the backend.

The exhaustive corpus collection of Språkbanken, which is accessed through Korp, consists of over 400 corpora with more than 13 billion tokens and almost one billion sentences representing mainly modern written Swedish, but also the older language, going back all the way to the Old Swedish of the Middle Ages.

Through the Korp corpus pipeline, researchers can import and annotate their own data. A pivotal characteristic of the pipeline is its dynamic nature, which allows researchers to integrate their existing annotations into the Korp infrastructure and use it as the basis for other types of annotation. The pipeline also provides researchers with a series of automatic annotation options – tokenisation, sentence splitting, links to the lexical persistent identifiers, lemmatisation, compound analysis, PoS/MSD tagging, and syntactic dependency parsing.

The Korp frontend is a graphical search interface, and thus the aspect of the corpus that researchers usually first come in contact with. The Korp frontend gives users the flexibility to search through the corpora by giving them the option to use simple queries or the CQP-corpus query language. After performing a search, users can then find the concordances under the KWIC tab (Figure 1), which also brings up a sidebar on the right-hand side that shows how the relevant token is annotated. Other functions of Korp include the ordbild (the word picture) tab (Figure 2), which shows the most relevant syntactic collocates of a lemma or text word; the related words tab, where a list of semantically-related lemmas is given; and the statistics tab, which provides users with a statistical overview of the token, as a table with a row for every unique hit and a column for every selected corpus, or in the form of a graph showing frequency of one or more linguistic phenomena over time (Figure 3). The Korp backend, which provides access to corpora, their annotations and their metadata, can be downloaded here, while most of the corpora that can be searched through Korp are available for download in Språkbanken.

[7] https://spraakbanken.gu.se/swe/node/1535

Figure 3: Concordances for the lemma "forskning" (“research”).

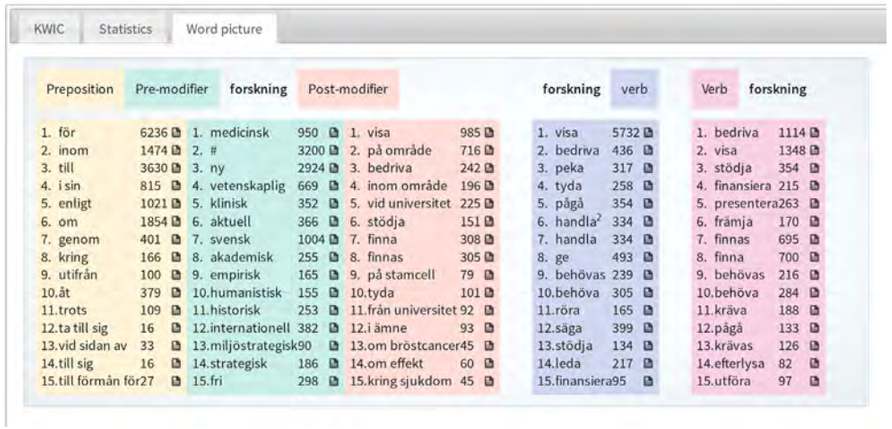


Figure 4: The word image for the lemma "forskning" (“research”).

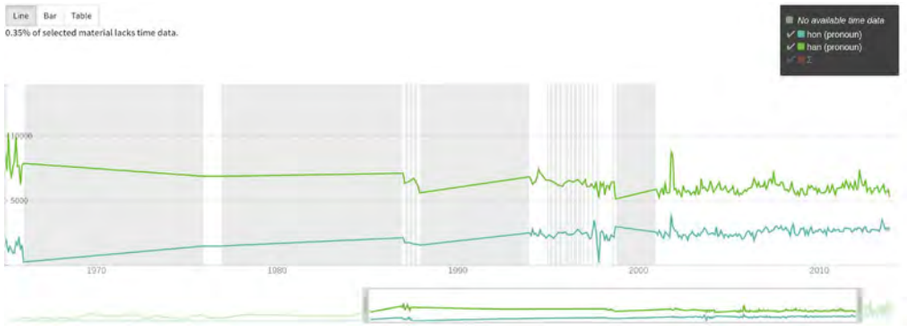


Figure 5: The trend diagram for the personal pronouns "hon" (“she”) and "han" (“he”) in the modern newspaper corpus.

The Riksdag's Open Data Corpus

Written by Darja Fišer and Jakob Lenardič

Since parliamentary speech has a great societal impact on account of its language and content, the creation and availability of large parliamentary multimodal corpora—a topic that was the subject of a CLARIN-PLUS workshop in 2017—play a pivotal role in humanitarian and social research.

The Riksdag's open data[8] is one such corpus. It is the digitised collection of Swedish parliamentary data and consists of roughly 30,000 documents pertaining to Sweden's national political decision processes. It has been made available for download on the website of the Swedish parliament. In addition, the Swedish National Library has digitised and published the public reports of inquiry for the period between 1922 and 1999 under the CC0 licence on the parliamentary website, with newer reports now being digitised from the very outset.

This parliamentary corpus is available in Korp and consists of 1.25 billion tokens. It can also be downloaded in the XML format from the resource page of Språkbanken. The annotation was performed with the Swe-Clarin's tool Sparv and consisted of tokenization, lemmatization, as well as lemgram (inflectional paradigm) and word sense identification, and compound splitting.

The resource has been successfully used by scholars working in the Social Sciences and Digital Humanities. Fredrik Norén from the Department of Culture and Media Studies at Umeå University has researched how social information in Sweden was structured in the period between 1965 and 1975, with a focus on uncovering how the government informed its citizens and communicated with them during this period. He has used Korp to search through SOU, a subset of the parliamentary corpus that contains the official reports of the government.

Norén has also collaborated with Roger Mähler from the Centre of Digital Humanities at Umeå University to analyse the changes in governmental discourse on the basis of the distribution of nouns. Using topic modelling, they were able to identify how information discourse arose in the 1960s and infiltrated governmental policies. Norén and Pelle Snickars from Umeå University have also used similar methods to analyse policies related to Swedish film in the 20th century on the basis of 4,500 reports in the SOU corpus. All in all, digitised language data like the Riksdag's open data corpus have made it possible to study the evolution of concepts like information in great detail, and by extension, they unveil historical change in a more precise and nuanced manner than ever before.

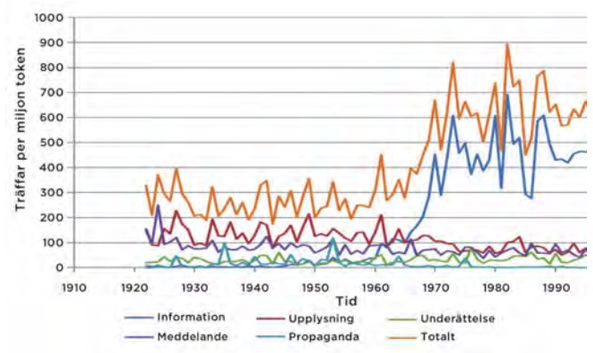


Figure 6: Frequencies of the lemmas "information", "upplysning" ("information"), "underrättelse" ("notification"), "meddelande" ("message") and "propaganda" throughout the 20th century.

[8] https://data.riksdagen.se/in-english/

Tutorial and Workshop on Automatic Sentence Selection from Corpora

Written by Darja Fišer and Jakob Lenardič

One of the most valuable aspects of an international research infrastructure such as CLARIN ERIC is the knowledge sharing that occurs among the national consortia. A successful example of this is the tutorial[9] and workshop[10] on automatic sentence selection for dictionary construction. The event, which was organised by Ildikó Pilán and Elena Volodina from Språkbanken, took place at the University of Gothenburg from 26 May to 1 June 2017 and brought together researchers from the Swedish, Estonian and Slovenian consortia.

The aim of the tutorial was to give an introduction to corpus data processing with Python and machine-learning approaches for lexicography, as well as offer opportunity for practical hands-on sessions with scikit learn and WEKA.

At the workshop, Ildikó Pilán described the HitEx extraction system, which is being developed at Språkbanken and is tailored to the automatic identification of corpus sentences for the exercises aimed at learners of Swedish as a second language. Adapted to various language-proficiency levels on the basis of the CEFR criteria, HitEx is a powerful system that allows for dynamic machine-assisted learning as it provides teaching professionals, lexicographers and students with options to set their own parameters, such as the difficulty level of the words they wish to learn. Iztok Kosem presented how the automatic extraction of corpus data has been successfully implemented in Slovene lexicography. He introduced the Collocations Dictionary of Slovene project, which has just released the first corpus-based dictionary of collocations for Slovene. Kristina Koppel presented the on-going work on compiling the Estonian Collocations Dictionary, which is in development will primarily be aimed at learners of Estonian at the B2-C1 levels.

Figure 7: The HitEx user interface for sentence selection with advanced search options.

Results

Rank	Score	Sentence

No sentences matched your criteria. See below for sentences with violations.

Results with violations

Rank	Score	Sentence
1	-1	När vi äter fet fisk får vi i oss fetter som är bra för kroppen .
2	-1	Han fick för sig att människor med munnen öppen såg ut som fiskar .
3	-1	Det finns risk att fisken försvinner från havet .
4	-1	Marocko tillåter att EU fiskar utanför Västsahara .

Different CEFR level: B2

Contains proper names: Marocko, EU, Västsahara

Typicality: 0

Figure 8: Corpus example sentence selection results for "fisk" ("fish") at B1 (intermediate) level.

[9] https://sweclarin.se/swe/sentence-selection-tutorial

[10] https://sweclarin.se/swe/mini-workshop-collocations-and-sentence-selection

Maria Ågren

Maria Ågren is a professor of history working at Uppsala University, Sweden. She leads the Gender and Work (GaW) project in which she has collaborated with Swe-Clarin researchers to create the GaW database, a collection of annotated historical language data that reveal the ways men and women supported themselves in the early modern history of Sweden. The interview was conducted by e-mail correspondence by Jakob Lenardič and edited by Darja Fišer.

1. Could you please briefly describe your background and tell us what your recent research is about?

I received my first degree in history and Swedish. I also have a diploma for teaching these two subjects in upper-secondary school; however, I never worked as a teacher because I enrolled as a graduate student in history instead. Since 2001, I have been a professor of history and my most recent research has been the Gender and Work (GaW)[11] project that I am leading.

2. How did you get involved with Swe-Clarin and what impact has this collaboration had on your research?

In the Gender and Work project, we are interested in finding snippets of information about people's jobs in historical documents, such as farm accounts, diaries, and court protocols. These snippets usually take the form: mend boat, sell eggs, take care of old people, and so on. At an early stage of the project, we told a linguist about our interest in building a database in which this type of information could be stored. She then exclaimed: "Aha! You are interested in verbs!" This comment had two far-reaching effects for our research project. First, we realised that we should call our method verb-oriented because it is a short and efficient way of explaining our approach that everyone immediately understands. Second, this linguist encouraged us to contact Professor Joakim Nivre from Swe-Clarin, which has led to a fruitful collaboration.

3. Which tools and corpora have you used and how did you integrate them into your existing research?

I did not use any existing corpora. Instead, the project has built its own corpus, the GaW database. Project members have gathered and classified thousands of fragments of information from a variety of handwritten historical sources that describe the ways people sustained and provided for themselves. The first stage of the project, which ran between 2010 and 2014 focused on the historical period from 1550 to 1800. The project now continues (from 2017 to 2021) with a focus on the period between 1720 and 1880. The GaW database is accessible to researchers, students, and the general public.

4. Have corpus data helped you reveal any interesting societal and linguistic trends of the periods you are interested in that would have been more difficult to uncover were it not for corpus-based methodology?

Yes, if one accepts my claim that the GaW database is a form of corpus then its data have been absolutely vital to the project. I would even say that most of our results could not have been achieved without it. Likewise, if one accepts that the verb-oriented method is a corpus-based methodology, then the answer is most definitely "yes". We have made many interesting discoveries about early modern society.

[11] http://gaw.hist.uu.se/what-is-gaw/research+project/

5. Could you describe the project in more detail? How did the language differ from contemporary Swedish? Are there any interesting differences from a socio-historical point of view? In what way have gendered roles/expectations changed from that time until now?

Gender and Work is a combined research and digitisation project at the Department of History at Uppsala University. The aim of the project is to acquire knowledge about the work of both men and women in the past. With the project we have been able to show the importance of the two-supporter model in early modern society; that is, that there was an expectation and practical reality of both men and women contributing to the household's survival. The project has also shown that what people did for a living in the past had more to do with marital status than with gender. The difference between what married and unmarried people did for a living was larger than that between what men and women did for a living.

One could say that early modern gender roles were more similar to the ones we have today – that is, both spouses worked, both spouses were expected to take care of children, even if the mother was thought to have a somewhat larger responsibility in this respect, people worked long days and could have to travel far to earn a living – than the ones that developed within the nineteenth-century bourgeoisie.

The Swedish language at the time was of course quite different from modern Swedish. There were no spelling rules, for instance, which makes for varied and, one might say, unorthodox spelling practices. It happened that German words were used in Swedish sentences. For researchers who use the corpus, the language itself is not the largest problem, since all scholars involved in the project are historians specialising in the early modern period and they are therefore all accustomed to reading early modern Swedish. The handwriting, on the other hand, is more of a problem; sometimes, the handwriting is so bad that you simply cannot make out what the text is about.

6. Has your field in general embraced the available digital text collections and language technologies? Do Swedish historians make use of language technology or collaborate with research infrastructures such as Swe-Clarin?

I think the answer to this question must be "no". In my opinion the Gender and Work project has been a pioneer within the historical disciplines in Sweden, especially because the collaboration with researchers working with language technology has allowed us to overcome a variety of technical difficulties we faced when dealing with the historical documents.

7. Could you elaborate on these methodological and technical challenges that a researcher working with historical text collections faces with respect to the available infrastructure?

There are two large problems: (1) the inconsistent spelling and (2) the fact that a majority of documents are only available in the original, handwritten form. The former problem is less daunting. In fact, there has been a highly successful collaboration with Professor Joakim Nivre and his now former PhD student Eva Pettersson from Swe-Clarin, which has led to substantial progress in overcoming the inconsistencies in spelling. For more in-depth information regarding this, see Eva Pettersson's doctoral dissertation "Spelling Normalisation and Linguistic Analysis of Historical Text for Information Extraction".

The latter problem is much more difficult to overcome. In the early modern period, state bureaucracies swelled and this led to a big increase in the production of documents, all of which are valuable to historians. But most of these sources are only available in handwritten form; rarely have they been printed (in which case they can be OCR-read) or digitised directly. If they are not available in digitised form, they cannot be processed automatically. If there were an easy way of transforming these handwritten documents to digital texts, then the corpus of early modern text material would grow enormously.

Since this is not the case, collaborative interdisciplinary projects like the one between Nivre and Pettersson on the one hand, and GaW on the other hand, are very rare. In our case, the historians read and annotated the texts manually, but at the same time also digitised them. This provided Pettersson with the language material on which she could train her normalisation tool. This tool identifies verbs, and particularly verbs describing work activity. The tool is not yet developed to perfection, but hopefully it will one day be possible to run it on digitised texts from the early modern period, and in this way speed up the processing of historical texts.

If you are interested in our approach to the extraction of information from historical texts, I suggest that you check the paper "HistSearch – Implementation and evaluation of a web-based tool for automatic information extraction from historical text" by Eva Pettersson, Jonas Lindström, Benny Jacobsson and Rosemarie Fiebranz.

8. What's your vision for CLARIN 10 years from now? What in your opinion should CLARIN focus on providing?

That it will be a permanent collection of resources and will contain more text corpora from the period between the Middle Ages and ca. 1800. This is the period during which many more documents were produced than in the Middle Ages, and most of them were not printed. After around 1800, handwriting became more similar to that we see today, and more documents were written on typewriters, making them easier to process automatically. The period from 1500 to 1800, on the other hand, is the period that is still largely unsupported in terms of corpora and text processing tools.

Stockholm, Sweden | photo by Davids Kokainis | Unsplash

Austria

Written by Darja Fišer and Jakob Lenardič

The Austrian CLARIN group is part of CLARIAH-AT,[12] a network of Austrian institutions participating in the two European research infrastructure consortia CLARIN and DARIAH. The network is comprised of eleven research departments at leading Austrian universities and heritage institutions, was a founding member of CLARIN ERIC, and is coordinated by Karlheinz Mörth, director of the Austrian Centre for Digital Humanities at the Austrian Academy of Sciences.

The main Austrian infrastructure for language resources is the Resource Centre for Humanities Related Research in Austria (ARCHE) CLARIN Centre Vienna, which is hosted by the Austrian Centre for Digital Humanities of the Austrian Academy of Sciences. Operating as a B-certified repository that received the Data Seal of Approval in 2014, ARCHE is a language-resource portal that offers researchers the opportunity to deposit and host language resources, language data and tools.

The focus of many activities of the group has been non-standard and historical language varieties. Among the corpora that can be accessed through the centre are the historical Mecmua corpus, which consists of language data from the early modern Ottoman period, the VICAV corpus, which pools language data for the research of Modern Spoken Arabic, and the Austrian Baroque Corpus (described on page 25), which is a specialised historical corpus of German texts from the memento mori genre written in the Baroque period.

[12] http://digital-humanities.at/en/dha/

Part of CLARIAH-AT | Back row: Gerlinde Schneider, Walter Scholger, Claudia Resch, Johannes Spitzbart, Friedrich Neubarth, Martin Hagmüller, Christiane Fritze, Gernot Kubin. Front row: Vesna Lušicky, Tanja Wissik, Helmut Kowar, Ursula Brustmann, Karlheinz Mörth. (Photo by Mehmet Emir, CC-BY 4.0).

Additionally, ARCHE also hosts DictGate, a platform for exchanging lexicographic tools and freely accessible lexical data such as small dictionaries of vernacular Arabic and a bilingual Persian-English dictionary that is still in development.

State-of-the-art tools that have been developed at the Austrian consortium can be freely obtained from the Centre as well. One such tool is the Viennese Lexicographic Editor (described on page 24), an XML-based dictionary writing system which serves as a user-friendly and dynamic tool for compiling and editing digital dictionaries. Another one is the SMC browser, which is a web application that offers users the ability to efficiently explore the Component Metadata Infrastructure by visualising its data.

As a very successful user involvement activity, the Austrian consortium hosts the ACDH Tool Galleries (described on page 27) three times a year. The Tool Galleries are workshops in which established scholars give lectures on the usage of digital tools relevant in digitally grounded humanities research. The consortium thereby promotes knowledge-sharing among its participants and propagates a multifaceted approach to research that is crucial when working in the digital humanities.

Viennese Lexicographic Editor

Written by Darja Fišer and Jakob Lenardič

The Viennese Lexicographic Editor[13] has been developed by the Austrian Centre for Digital Humanities and is a standalone XML editing system that is designed for collaborative work on lexicographic data. The tool can be freely downloaded and updated versions are uploaded regularly.

A powerful and adaptable tool, the Viennese Lexicographic Editor provides a flexible environment for navigating through and working with complexly-annotated dictionary entries. Researchers can use the tool either to directly access the data in XML (Figure 1) or to edit them by means of an easy-to-use graphical database interface (Figure 2). Furthermore, the tool offers different ways to visualise the data and checks for well-formedness whenever researchers save their entries. Through a special module, the Viennese Lexicographic Editor allows its users to access and integrate external language resources, such as corpora and other dictionaries.

The Viennese Lexicographic Editor has established quite a tradition since it was first used within a glossary-building student project. It has been used as the key piece of software for the compilation of electronic dictionaries, such as the TUNICO Dictionary, within the international Viennese Corpus of Arabic Varieties and the Linguistic dynamics in the Greater Tunis Area projects. In turn, these dictionaries, which are freely available through the DictGate platform, have served as an important resource for comparative dialectological research of language varieties and have been used in language teaching courses, thereby facilitating a cooperative approach to lexicography and lexicology within the digital humanities and social sciences.

```
<entry xmlns="http://www.tei-c.org/ns/1.0" xml:id="amaandla_005">
   <form type="lemma">
      <orth>amandla</orth>
   </form>

   <gramGrp>
      <gram type="pos">pluralNoun</gram>
   </gramGrp>

   <form type="stem">
      <orth>andla</orth>
   </form>

   <sense>
      <cit type="translation" xml:lang="en">
         <quote>strength</quote>
      </cit>
      <cit type="translation" xml:lang="en">
         <quote>power</quote>
      </cit>
   </sense>
```

Figure 9: Editor for TEI dictionaries (screenshot by Karlheinz Mörth).

[13] https://www.oeaw.ac.at/acdh/tools/vle/

The ABaC:us Corpus

Written by Darja Fišer and Jakob Lenardič

The Austrian Baroque Corpus (ABaC:us)[14] is a digital collection of printed texts from the Baroque era, with the bulk of the data from the period between 1650 and 1750. Since 2015, the core corpus has been freely available through the ABaC:us web application, which is provided by the Austrian Centre for Digital Humanities and serves as the first corpus-based application for viewing well-documented language data from the Baroque period. The texts within the collection are predominantly characterised by religious topics, and include morality lectures by Abraham a Sancta Clara, who was one of the most successful preachers in the German-speaking area in the 17th century.

Figure 10: ABaC:us web application (screenshot by Claudia Resch).

The collection consists of 200,000 tokens. Its rather smaller size is due to the fact that computer-generated annotations of Baroque texts, whose stylised orthography and variations in spelling cause a large amount of mismatches by extant taggers, requires copious amounts of additional manual editing. However, the data that are part of the corpus are very richly annotated with mark-up applied to chapters, headings, paragraphs, and named-entities. Apart from PoS-tagging, the corpus is annotated with lemma information, which means that each word form is linked to its base form. Lemma information is a crucial part of the corpus, as it enables researchers to easily identify all occurrences of a word despite the existence of many competing spelling variants and inflected forms.

Because of ABaC:us, scholars are for the first time able to explore the vocabulary as well as linguistic structures of the works attributed to Abraham a Sancta Clara in a corpus-based approach. Moreover, the detailed linguistic annotation allows for unbiased research of Sancta Clara, who is portrayed as a particularly linguistically-talented writer in literary history.

[14] https://www.oeaw.ac.at/acdh/tools/abacus/

The ABaC:us project has broken new ground since it allows scholars to combine methods from historical studies—whether they be literary, theological or purely historical—with a corpus-based approach that enables access to richly-annotated data. Indeed, reactions from literary scholars, (computer) linguists, religious scholars and historians have shown how interdisciplinary the interest in ABaC:us is and how many different fields of research across the digital humanities hope to benefit from the free availability of this enriched resource.

The following is a selection of published papers on the ABaC:us corpus:

Resch, C. (2017). "Etwas für alle" – Ausgewählte Texte von und mit Abraham a Sancta Clara digital. In Zeitschrift für digitale Geisteswissenschaften 2017. http://www.zfdg.de/2016_005.
Resch, C. and Czeitschner, U. (2017). Morphosyntaktische Annotation historischer deutscher Texte: Das Austrian Baroque Corpus. In Digitale Methoden der Korpusforschung in Österreich (= Veröffentlichungen zur Linguistik und Kommunikationsforschung, 30, 39–62.
Resch, C., Czeitschner, U., Wohlfarter, E., and Krautgartner, B. (2016). Introducing the Austrian Baroque Corpus: Annotation and Application of a Thematic Research Collection. In Proceedings of the Third Conference on Digital Humanities in Luxembourg with a Special Focus on Reading Historical Sources in the Digital Age.
Resch, C. and Wolfgang, U. Dressler. (2016). Zur Pragmatik der Diminutive in frühen Erbauungstexten Abraham a Sancta Claras. Eine korpusbasierte Studie. In Linguistische Pragmatik in historischen Bezügen, 235–249.

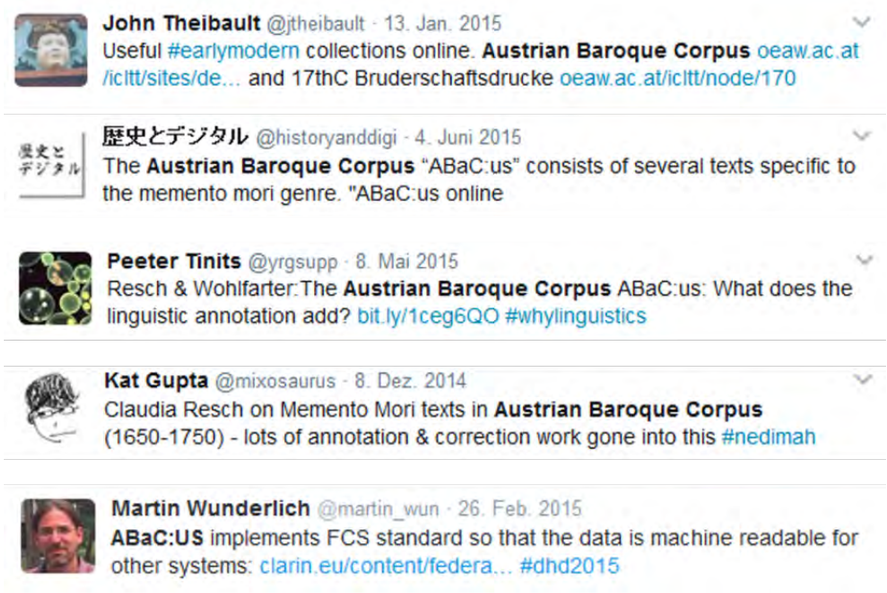


Figure 11: Tweets on ABaC:us by humanities scholars.

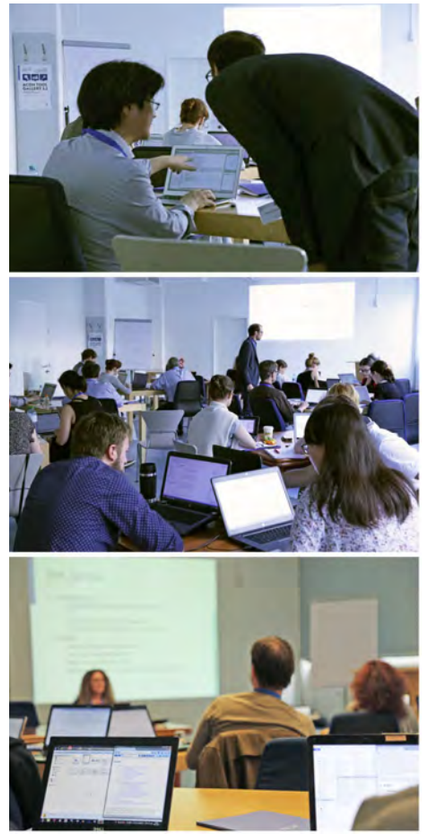
The ACDH Tool Gallery

Written by Darja Fišer and Jakob Lenardič

The training of new and experienced researchers plays a very important role in the digital humanities and social sciences, guaranteeing the far-reaching utilisation of computation tools and digital methods. In Austria, the ACDH Tool Galleries[15] are exemplary cases of such training endeavours. The Tool Galleries are user involvement events organised three times a year by the Austrian Centre for Digital Humanities (ACDH). At these, the developers of tools and experienced professionals share their theoretical and practical knowledge on tools that are designed for digital humanities users. The events take the form of morning lectures and hands-on sessions in the afternoons; the practical work done at these is particularly valuable, especially since it offers the attendees the chance to immediately consult with tool experts if they encounter a problem.

Eleven Tool Galleries have been organised so far, and each has been dedicated to presenting a different subfield in computational linguistics and related tools. For instance, the second Tool Gallery, which took place on 13 October 2015, focused on the importance and use of basic linguistic annotation and was intended for linguists and professionals from all disciplines. Annotation work on the Austrian Baroque Corpus (ABaC:us), which is presented on page 25, and the Austrian Media Corpus was presented, while Marie Hinrichs and Claus Zinn from the University of Tübingen gave a talk on Weblicht, which is a fully functional processing chain that brings together linguistic tools such as tokenisers, part-of-speech taggers, and parsers.

Although the Tool Galleries were originally intended as a service for employees of the Austrian Academy of Sciences, the format was soon extended to a much larger audience that now includes students and academics at all career stages. By organising Tool Galleries three times a year, the Austrian Academy hopes to achieve a regularity and continuity that will serve as a model of researcher training.



ACDH Tool Gallery (photo by Sandra Lehecka, CC-BY 4.0).

[15] https://www.oeaw.ac.at/en/acdh/events/event-series/

Stephan Procházka

Stephan Procházka is a linguist working at the Department of Oriental Studies at the University of Vienna who has collaborated with the Austrian CLARIN consortium in the interdisciplinary TuniCo project, which focused both on researching the linguistic dynamics of the greater Tunis area as well as producing a dictionary of Tunis Arabic and a corpus of transcribed texts. The interview was conducted by e-mail correspondence by Jakob Lenardič and edited by Darja Fišer.

1. Your main research interests lie in Arabic studies. What initially attracted you to the field and what excites you most today?

Initially I was mainly attracted by the fascinating Arab history and the rich material culture such as arts and architecture. For many years now, spoken Arabic varieties have become my main field of interest and research. The so-called dialects are not only interesting for linguists, but also vehicles of a multifarious oral culture ranging from traditional Bedouin poetry to hip-hop songs in the suburbs of Arab megacities.

2. How did your collaboration with the Austrian CLARIN begin and how has it influenced your own work and the way you perceive contemporary Arabic studies?

My collaboration with CLARIN began in 2011 when I was looking for a competent partner to build a kind of platform for Arabic dialectology. I found that in the then ICLTT, which was the fore-runner of the Austrian Centre for Digital Humanities of the Austrian Academy of Sciences (ACDH-OeAW). From this cooperation many projects such as VICAV emerged.

3. Your most recent project was the interdisciplinary TuniCo[16] project in which you and your team investigated the linguistic dynamics in the greater Tunis area. Could you briefly describe the methodological framework of the project, and highlight its impact for digital humanities and social sciences?

The project was based on the analysis of data gathered during two longer fieldwork campaigns in Tunis. Texts that had been transcribed from the recordings of conversations among young people were the core of our analysis. These texts formed the basis of both lexical and grammatical research, the latter mainly in the field of syntax.

4. Has your analysis of contemporary Arabic spoken by young speakers from different backgrounds revealed any interesting societal trends or culturally specific characteristics?

Yes, we found out that remarkable changes have happened during the last few decades. Young men in particular increasingly show features in their speech which are stigmatised and mostly connected with low-class people from the countryside. They deliberately choose these features to set themselves apart from the mainstream culture. Young educated women, on the other hand, have a preference for using many French words and phrases to show that they are modern and open-minded.

[16] https://tunico.acdh.oeaw.ac.at/

5. Can you describe the two main resources that were developed in this project? What kind of advantages do they bring to your fellow researchers in the field?

We produced a dictionary of Tunis Arabic that comes in a digitally reusable form and lives up to modern IT standards. It contains a very wide range of lexical data, from “historical” vocabulary taken from previous studies to up-to-date youth language taken from our interviews and rap songs, ca. 8,500 entries. It is currently the largest and technically most advanced online dictionary of a spoken Arabic variety worldwide. Together with the other VICAV dictionaries, it is the only such product that is freely available for future research and at the disposition of all researchers. The second resource is a corpus that consists of 24 transcribed texts with ca. 100,000 words. This corpus is linked to the dictionary and thus gives users direct access to the relevant dictionary articles and allows them to understand the Tunisian original. The inclusion of a large number of conversations is one of the innovative traits of our corpus approach, as there are extremely few corpora of spoken varieties of Arabic which include dialogues.

6. How does Arabic, or rather its varieties, fare in the digital context? Are language resources and tools for Arabic readily and widely available? Are there any difficulties specific to automatic processing of Arabic and its varieties? Is there any essential tool or resource that is still missing for Arabic?

Digital language resources for Arabic in general and its spoken varieties in particular, both data as well as tools, are, for several reasons, still under-represented in comparison to many other languages. A major problem is that automatic processing of Arabic, for instance part-of-speech tagging, is more complex because of the characteristic Arabic script that does not indicate short vowels. Arabic varieties are only written in informal settings and lack any standard orthography, which further complicates automatic processing.

7. Have your fellow researchers in the field embraced language technologies in their research frameworks? What is the potential of using language technologies for Arabic studies?

Many scholars in the field are still sceptical about language technologies. However, I see very high potential for my field of research, particularly in the fields of lexicography and syntax. While several treebanks have become available for Modern Standard Arabic, there remains much to be done for the spoken varieties of Arabic.

8. How is the available infrastructure provided by the Austrian consortium or CLARIN ERIC beneficial for your research? Could you highlight a CLARIN tool or resource that has been especially helpful for your work? Would you like to point out anything that could be improved in the future?

The cooperation has been excellent and the available infrastructure very satisfying. The main CLARIN tool for me is the Viennese Lexicographic Editor which from the very beginning facilitated the work in the project. The Vienna CLARIN Centre takes care of the entire resource publication side of our projects, provides both for hosting and preservation of research data, and has always been very helpful in setting up web-interfaces. Especially in our work on the corpus-dictionary interface, the infrastructure of CCV and ACDH-OeAW proved to be very useful.

9. What do you see as the biggest strength of Austrian CLARIN?

They are really interested in cooperation with the humanities and very user orientated. Their interest in further development of their infrastructures in concrete research projects opens up unprecedented synergies, and allows us to move our research in entirely new directions.

10. Where would you like to see CLARIN ERIC 10 years from now?

I think we all would like to have more freely available resources, data and tools that can be used by all researchers, can easily be adapted to the needs of a wide range of fields and projects. While many tools have become available, we still have a long way to go in terms of usability. Finally, I would like to say that I regard CLARIN's user involvement activities as a very important part of our activities. While much has already been achieved, there are still many in various fields of the humanities who are not aware of recent developments. My vision of CLARIN in 10 years from now is that all young researchers are sufficiently aware of the possibilities the pan-European infrastructure consortia provide, and that the new digital methods are taught in introductory seminar courses on a regular basis, which will eventually lead to wholly new research questions and results.

Vienna, Austria | photo by Andreas N. | Pixabay

The Netherlands

Written by Darja Fišer and Jakob Lenardič

CLARIAH-NL[17] is a project in the Netherlands that is setting up a distributed research infrastructure that provides humanities researchers with access to large collections of digital data and user-friendly processing tools. The Netherlands is a member of both CLARIN ERIC and DARIAH ERIC, so CLARIAH-NL contributes to both CLARIN and DARIAH. CLARIAH-NL not only covers humanities disciplines that work with natural language (the defining characteristics of CLARIN), but also disciplines that work with structured quantitative data. Though CLARIAH aims to cover the humanities as a whole in the long run, it currently focusses on three core disciplines: linguistics, social-economic history, and media studies.

CLARIAH-NL is a collaborative project that involves around 50 partners from universities, knowledge institutions, cultural heritage organisations and several SME companies. Currently, the data and applications of CLARIAH-NL are managed and sustained at eight centres in the Netherlands: Huygens Ing, the Meertens Institute, DANS, the International Institute for Social History, the Max Planck Institute for Psycholinguistics, the Netherlands Institute for Sound and Vision, the National Library of the Netherlands, and Dutch Language Institute. Huygens Ing, the Meertens Institute, the Max Planck Institute for Psycholinguistics, and Dutch Language Institute are Certified CLARIN Type B-centres. The consortium is led by an ten-member board, and its director and national coordinator for CLARIN ERIC is Jan Odijk.

The research, development and outreach activities at CLARIAH-NL are distributed among five work packages: Dissemination and Education and Technology deal with user involvement and the technical design and construction of the infrastructure, respectively, whereas the remaining three work packages focus on three selected research areas: Linguistics, Social and Economic History and Media Studies.

Dissemination and Education work package

In the user involvement-focused Dissemination and Education package, CLARIAH-NL aims to facilitate knowledge sharing among digital humanities and social sciences scholars as well as provide services that cater to the needs of their research. In this respect, CLARIAH-NL has successfully organised a variety of user involvement activities, such as the CLARIAH Linked Data Workshop (described on page 37), which took place in June 2017 and was intended to introduce Linked Data to both novice and advanced researchers.

[17] http://www.clariah.nl/en/about/international

Linguistics work package

MIMORE

In the Linguistics work package, CLARIAH-NL focusses on developing and improving applications for enriching corpora and searching through them – one such tool is MIMORE, which is described in greater detail on page 34 . This enables researchers to investigate morphosyntactic variation in the Dutch dialects by searching three related databases with a common online search engine. The search results can be visualised on geographic maps and exported for statistical analysis. The three databases involved are DynaSAND, DiDDD and GTRP.

SoNaR

An important data set in this connection is the Dutch reference corpus SoNaR (described in greater detail on page 35), which was created in earlier projects for developing NLP software, but has been opened up for research by humanities scholars through the OpenSoNaR web application. State-of-the-art tools for the enrichment of textual corpora are also developed at the consortium. An example of such software is Frog, which is an NLP suite containing a tokeniser, PoS-tagger, lemmatiser, morphological analyser, and is thus an entity recogniser and dependency parser for Dutch.

Social and Economic History work package

In the Social and Economic History package, structured databases of social-economic history are being integrated into the Linked Data paradigm. The use of a uniform structure and explicit semantics ensures that relations and connections can be searched across different databases, which is of crucial importance for historical analysis and allows researchers to easily test hypotheses that could not be investigated before.

Media Studies work package

Finally, the Media Studies package focuses on providing special tools for viewing, browsing and searching through large collections of audio-visual data, such as films, radio broadcasts, and vlogs. It aims to provide a Media Suite, with access to relevant audio-visual collections by integrating tools developed in earlier projects, such as AVResearcherXL,[18] which is a tool for exploring radio and television programme descriptions, television subtitles and general newspaper articles through a user-friendly graphic interface.

The board of CLARIAH-NL and National Coordinator Jan Odijk.

[18] http://www.clariah.nl/en/about/international

MIMORE

Written by Darja Fišer and Jakob Lenardič

MIMORE[19] is a tool developed at the Meertens Institute by means of which researchers can investigate morphosyntactic variation in Dutch dialects. Using this online environment, three different databases can be queried:

- the Dynamic Syntactic Atlas of the Dutch Dialects, which contains recordings and transcriptions of Dutch spoken in over 300 locations across the Netherlands, Belgium and a part of north-western France. It focuses on the syntactic variation among the dialects in these areas, such as differences in question formation and word order;
- the Diversity in Dutch DP Design, which contains oral and written interviews from about 200 locations in the Dutch language area and focuses on morphosyntactic variation in nominal structures; and
- the Goeman, Taeldeman, van Reenen Project, which focuses on morphological variation (such as the differences in verbal inflection) among roughly 600 locations in the Dutch language area.

MIMORE is very versatile in that it allows researchers to narrow down their search according to parameters relevant for the linguistic study of dialects, such as specific geographic locations or syntactic phenomena in which the dialects might differ from one another. The search engine also allows its users to export the data for statistical analysis, as well as to visualise them on a geographic map of the Dutch language area. It is accompanied by a comprehensive Educational Module, which provides an example of use on how to find sentences where the grammatical subject is realised twice within the same clause, which is a syntactic phenomenon that is limited to the south-western and western-central dialects of Dutch. There is also a case study available that demonstrates the combined use of MIMORE and the treebank search engine GrETEL.

The data that can be accessed through MIMORE point to the fact that many local Dutch dialects have begun to disappear or change in the direction of the standard language, due to increasing mobility and changes in the communication of its speakers. Additionally, the tool is also of great importance for formal linguistics, as it allows researchers to conduct micro-comparative studies of the Dutch language on the basis of data from dialects that show variation in terms of highly specific syntactic phenomena, such as the existence of complementiser agreement in certain Dutch dialects.

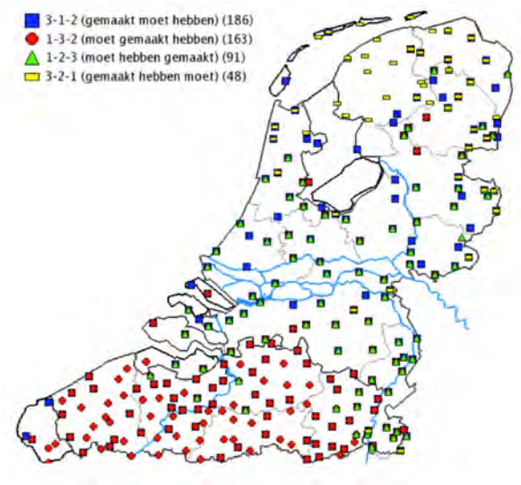


Figure 12: The visualisation tool offered by MIMORE which shows how Dutch dialects differ from one another on the basis of different word orders in the verbal phrase.

[19] http://www.meertens.knaw.nl/mimore/

The SoNaR Reference Corpus of Dutch

Written by Darja Fišer and Jakob Lenardič

SoNaR[20] is a reference corpus of standard written Dutch. It comprises contemporary texts ranging from printed media such as books and periodicals to computer-mediated communication such as chats and tweets from the Netherlands and the Dutch-speaking area in Flanders (SoNaR New Media). It is the result of the STEVIN project, which involved major universities in the Netherlands and the Dutch-speaking part of Belgium, Flanders. The aim was to create a corpus of the contemporary written language, originally primarily intended for use by language and speech technology researchers and developers. It was made accessible and usable for humanities researchers in the CLARIN-NL and CLARIAH-NL projects by providing a web application with an interface for humanities researchers.

SoNaR consists of two main subcorpora – SoNaR-1 and SoNaR-500. In addition, there is the SoNaR New Media Corpus.

SoNaR-1 contains one million tokens and is very richly annotated, especially in relation to the semantic layers, which consist of named-entity labelling, annotation of co-reference relations, semantic role labelling and annotation of spatial and temporal relations. Additionally, all its annotations have been manually verified. As one of its pivotal subparts, SoNaR-1 includes the Dutch Parallel Corpus, a sentence-aligned parallel corpus of English, Dutch and French. The larger subcorpus, SoNaR-500, contains 500 million tokens of full texts. The texts in SoNaR-500 have been tokenised, tagged for part-of-speech and lemmatised, but without manual verification.

The SoNaR New Media corpus contains approximately 35 million words and consists of tweets, chats and SMS. All texts have been automatically tokenised, tagged for part of speech and lemmatised.

In order to provide easy access to the corpus, CLARIN-NL and CLARIAH-NL have developed the OpenSoNaR search environment. OpenSoNaR, with a frontend called WhiteLab and backend named BlackLab, is a state-of-the-art concordancer which provides two primary interfaces of user-driven functionality that can be used by both laymen and specialist researchers alike. In the Exploration interface (Figure 13), a researcher can investigate the corpus distribution, see the statistical information of the subcorpora and retrieve n-grams. Through the Search interface, four search options are available:

- simple, which limits the search to words only;
- extended, which enables the researcher to query the corpus by either word form or lemma, set the part of speech and choose among semantic metadata filters (Figure 14);
- advanced, which allows users to further specify the lemma or word forms that they're interested in; and
- expert, which provides an input for CQL commands.

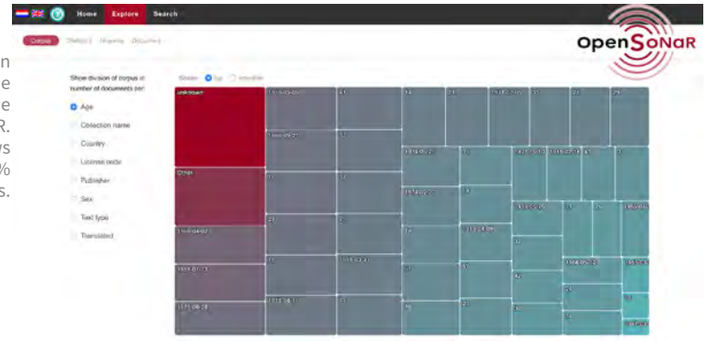


Figure 13: The Exploration interface showing the distribution of the subcorpora within SoNaR. The highlighted box shows that tweets make up 0.03% of the corpus.

[20] https://dev.clarin.nl/node/4195

The OpenSoNaR environment also stores previous search results, allowing researchers a great degree of flexibility and room for comparison between the temporary subcorpora that they have created during a single search session (Figure 15).

The successor of OpenSoNaR, called OpenSoNaR+, was developed in 2015.

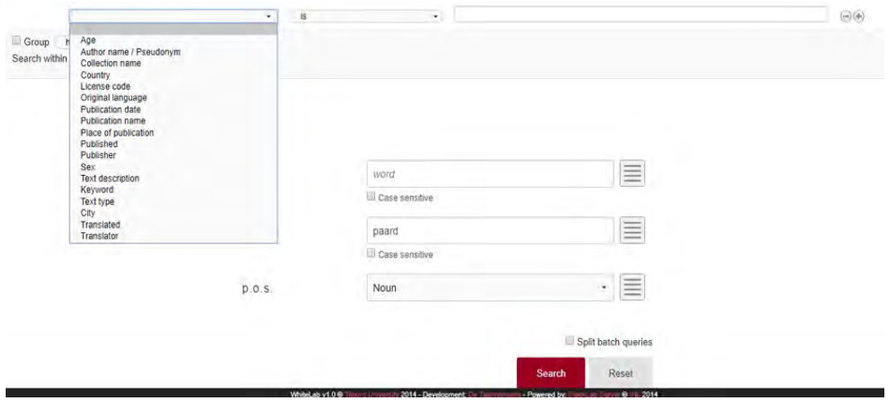


Figure 14: The ”extended” search interface — a search is being performed for the lemma "paard" (“horse”), while the drop menu shows metadata filters.

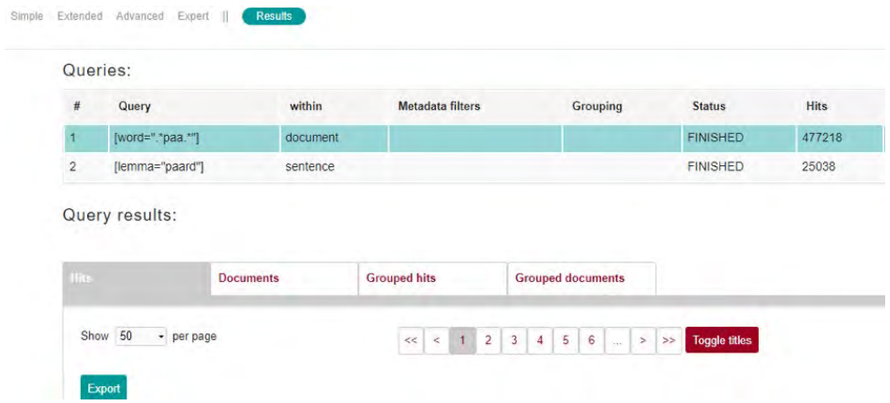


Figure 15: The “Results” tab shows that OpenSoNaR stores previous the results of queries.

CLARIAH-NL Workshops on Linked Data

Written by Darja Fišer and Jakob Lenardič

On 12 September 2016, CLARIAH-NL organised the first in a series of workshops on Linked Data,[21] which is a technological initiative that aims to ensure a greater degree of dynamic interoperability between language resources provided by the infrastructure. For instance, when Linked Data is applied to parliamentary corpora, as was the case in the CLARIN project Talk of Europe, the debates are enriched with extra layers of information in the sense that they are linked to their respective speakers, who are in turn annotated with biographic information, such as gender and political affiliation.

The workshop was attended by 40 researchers and primarily served two roles. On the one hand, it was an introduction to Linked Data and how it fares with respect to the CLARIAH-NL infrastructure, both in terms of successful applications and future challenges primarily related to bridging the gap between technological experts and humanities users who are not necessarily technologically savvy. On the other hand, the workshop directly involved arts and humanities researchers so that they could present their experiences of using Linked Data. For instance, Kaspar Beelen and Liliana Malgar from the University of Amsterdam gave a talk on the Digging into Parliamentary Data project, the aim of which was to enrich the parliamentary records of the Netherlands, United Kingdom and Canada with Linked Data so as to give researchers the opportunity to easily investigate the complex socio-historical aspects of politics.

CLARAH-NL Workshop on Linked Data, 12 September 2016. Image: CLARIAH-NL website

On 6 and 7 February 2017, CLARIAH-NL organised a follow-up international workshop that focused primarily on the application of Linked Data to linguistic research. It involved a number of international experts on Linked Data as well as prominent Dutch linguists who presented their current research topics in the fields of lexicology, phonetics and syntax in relation to using Linked Data. Among the speakers was Sjef Barbiers from the University of Leiden, who gave a talk on how Linked Data can benefit comparative syntax by ensuring interoperability between various databases of Dutch dialects that can be accessed through the tool MIMORE (read the presentation of the MIMORE on page 34). Jan Odijk together with Sjef Barbiers concluded the workshop by envisioning that the next step for CLARIAH-NL is to work on further stimulating the use of Linked Data in the fields of linguistic research where it is not yet widely applied, such as general lexica, and thereby reach out to a broader group of linguists.

[21] https://www.clariah.nl/en/new/blogs/clariah-linked-data-workshop

Melvin Wevers

Melvin Wevers is a digital humanities researcher focusing on the study of cultural-historical phenomena with the use of computational means. The following interview took place via Skype on 14 November 2018 and was conducted and transcribed by Jakob Lenardič and edited by Darja Fišer.

1. Can you please briefly describe your research background and tell us how you became a digital humanist who uses computational approaches to studying cultural phenomena?

I have a pretty diverse research background. I started out in the social sciences studying psychology and after I received my degree in 2006 I actually discovered that I didn't want to be a psychologist, but would rather do something that is much more based in the humanities, such as researching culture. Since I've always been interested in American culture, specifically, I applied for a Master's track in American Studies at Utrecht University. What I really liked about this field is its methodological variety in the sense that it combines elements from historical studies, media studies and literature. This kind of multidisciplinary approach made me become very interested in research and I decided to pursue a PhD after receiving my MA in 2009. However, I couldn't find a suitable PhD programme at first so I started studying cultural analysis at the University of Amsterdam, which was based more on quantitative methods, and at the same time began working at a software company that also used text mining, which sparked my interest for language technologies. Then I saw an advertisement for a PhD position at the University of Utrecht for using computational methods to research how American culture was represented in Dutch newspapers throughout the 20th century. I thought that this was a perfect opportunity for me, so I applied and began my career as a digital humanist!

2. How does your research benefit from the CLARIAH-NL infrastructure?

My PhD was funded by the Dutch Science Organisation, and the data that we used were provided by the National Library of the Netherlands. Though the project itself wasn't directly linked to CLARIAH-NL, I met a lot of people affiliated with CLARIAH-NL, like Arjan van Hessen and Franciska de Jong, at various conferences that I attended during the course of my studies. They pointed me to events organised by CLARIN consortia. These tutorials made it much easier for me to learn programming languages like R and Python, and I met a lot of my future colleagues with whom I could discuss my work in relation to source criticism or tool criticism. For instance, through the CLARIAH-NL consortium I learned that the German DARIAH was organising a tutorial on topic modelling. I learned a great deal about specific algorithms related to topic modelling that I would later use for my PhD. I think that CLARIAH-NL serves as an essential network that makes it significantly easier for researchers working in different fields to collaborate, especially since people like Arjan and Franciska put so much effort into helping novice researchers build the much needed connections.

3. In your PhD thesis, "Consuming America", you've applied a quantitative approach to a socio-historical study of how American consumer culture was depicted in the Netherlands throughout the 20th century. What inspired you to start researching this topic? Could you briefly describe your approach as well as the main findings of your research?

My PhD was part of a very large project funded by the Dutch government called "Translantis", which focused on determining how the United States was perceived in Dutch public discourse throughout the 20th century. My role was to focus on consumer goods such as Coca Cola and cigarettes in order to determine how American cultural values were portrayed in Dutch newspapers. This kind of research allowed me to gain a very multifaceted understanding of how Dutch people reacted to notions such as modernisation and globalisation through their perception of consumer goods. Since I've had a lifelong interest in all aspects of American culture—especially its international impact—I felt that this kind of research was a perfect opportunity for me.

In my approach, I combined the close reading of a more traditional historian with data-driven computational methodology. I first looked at a number of specific newspaper articles to get a very general feel of what the Dutch people thought about American culture at different times throughout the 20th century. Then I used quantitative methods like topic modelling on millions of newspaper articles to see whether such perceptions, as reported by these newspapers, constituted broader trends in Dutch history. In American Studies there is a deeply-entrenched idea that the 1950s and 1960s were a turning point during which American influences started becoming pervasive in the Netherlands—in other words, there is an idea of an American cultural invasion after the 1950s. However, by focusing on the depiction of consumer goods in newspapers I was able to show that the Dutch were already very much interested in and directly involved with American culture even before World War I. That is to say, the American influence in the Netherlands was relatively stable throughout the 20th century, so there was no specific mid-century turning point, as the Dutch had started to perceive themselves as modern consumers in the American sense from very early on.

In relation to a specific finding, Coca Cola was one of my case studies, and there I uncovered a very interesting dichotomy. In international advertisements, the Coca Cola company strove to advertise its product as global by omitting references to its American origin; however, in spite of this attempt, Dutch newspapers continued to overwhelmingly associate Coca Cola as something distinctively American. This in turn led me to uncover a major trend in Dutch public discourse, which is that the notion of globalisation became associated with Americanisation.

4. How does such a data-driven approach complement the traditional methods of a historian? Are there any specific advantages to such an approach?

After I finished my psychology degree, I lost interest in the quantitative methods of social sciences like statistics for a time, and instead wanted to solely focus on the traditional methods in the humanities, such as close reading and reading against the grain. However, I soon became critical of the lack of empirical evidence in the humanities, and I again became interested in data-driven methods. Ultimately what I learned is that these two approaches need to be combined, since this greatly increases the breadth of the research questions that a researcher is able to ask. That is, I think that computational methodologies can greatly assist a historian, especially since they make it much easier to adopt a bird's eye view of the periods that are being researched and thereby contextualise them properly as parts of the overarching historical trends.

5. Have historians working in your field generally embraced such quantitative methodology? Are there any changes that you would personally like to see take place within the field?

Unfortunately, in my field—that is, cultural history—using computational approaches is still a very new endeavour, so there are many researchers who outright refuse to use anything other than the traditional non-quantitative methods. This is understandable to an extent, since a senior researcher probably won't find the time to learn how to program late in his or her career. However, I feel that if you want to train a future generation of humanities scholars you should include courses on programming in the curriculum. Of course, this is far easier said than done, since I think this would require a kind of paradigm shift where entire syllabi would have to be revised in order to explicitly define, for instance, how a programming language like Python can be used to tackle research questions in fields where it is not immediately obvious how to apply quantitative methodologies. Because what often happens in practice is that a humanities department has a course on Python, but there are no other related courses that would help students apply their programming knowledge to research problems directly applicable to humanities questions. In general, my opinion is that there should be a marriage between distant reading and close reading in the humanities, so I would like to see a greater degree of collaboration between scientists from different fields, such as between historians and computational linguists. I've written some papers with people who have a better understanding of mathematics than I do. If I had been left solely to my own devices, I would have had to spend a lot of time learning advanced mathematics, which would in turn probably make me neglect the humanities part of my research question. However, since I know some programming and some mathematics, it is easier for me to communicate with people that are experts in these fields. Such communication has already resulted in some very worthwhile interdisciplinary collaborations.

6. In your opinion, what could CLARIN do to become more widely used by historians? What activities, resources or tools would be needed to achieve this?

In the Netherlands, I think that CLARIN is still associated almost exclusively with computational linguistics even though CLARIAH-NL tries very hard to branch out into other humanities disciplines. So I think that they should continue to organise tutorials and especially focus on showcasing how the various datasets that are already out there are relevant for various disciplines. For instance, there is a plethora of historical sources that have been digitised, but many historians aren't aware of the various exciting ways in which they could direct their research on the basis of the wide availability of these datasets.

7. You've been involved in the development of ShiCo, a tool for the analysis of how words denoting a certain concept change diachronically. Could you briefly describe how this project came about? What are the main advantages of ShiCo?

I have been interested in figuring out how words denoting a certain concept change over time, but found that approaches such as topic modelling were too rigid to do this efficiently. I approached another PhD student, Tom Kenter, who specialises in Natural Language Processing and information retrieval with this problem, and he came up with the idea to use a relatively novel technique to chart how these changes happen. We involved some other researchers working in the history department at the University of Amsterdam so that we could test whether the results of our first prototype were in accordance with their expert knowledge of the domains. Since the prototype was successful, we were encouraged by some of the professors to apply for a grant and turn the prototype into an interactive tool. By working with programmers from the eScience Centre, we eventually managed to turn the tool into ShiCo.[22]

[22] https://github.com/NLeSC/ShiCo

8. Can you highlight any other project that you're currently working on?

After finishing my PhD, I became a post-doc researcher at the National Library of the Netherlands, where I applied techniques from the field of computer vision to gain insights into non-textual trends in the Dutch advertisements landscape. The research produced a dataset of advertisements as well as a tool called SIAMESE to find visually similar images in a large corpus of advertisements. Currently I'm working on applying computational methods to analyse Dutch academic historical journals and thereby determine the trends related to the understanding of history on the part of Dutch historians – for instance, how notions like progress and modernity are discussed and which countries were in focus over different periods of time.

9. What is your vision for the future of CLARIAH-NL and digital humanities in the Netherlands?

I believe that computational methodology should become part and parcel of all kinds of disciplines and that digital humanities should, at a certain point, lose the modifier digital and become the standard way of doing humanities research. I think that CLARIAH-NL can play an important role in bridging the gap between these different fields, especially by offering interactive tutorials and ensuring interoperability between repositories and tools. Like I said previously, one of the problems is that researchers do not know what to do with the available data so CLARIAH-NL could offer these much-needed guidelines and training, as well as educate researchers on concepts like open science so as to ensure that their work is as transparent as possible.

Amsterdam, the Netherlands | photo by Sabina Fratila | Unsplash

Poland

Written by Jan Wieczorek and Ewa Rudnicka, edited by Darja Fišer and Jakob Lenardič

The Polish consortium CLARIN-PL[23] is a founding member of CLARIN ERIC and has been actively involved in its operations since the very beginning in 2005. It comprises six member institutions:

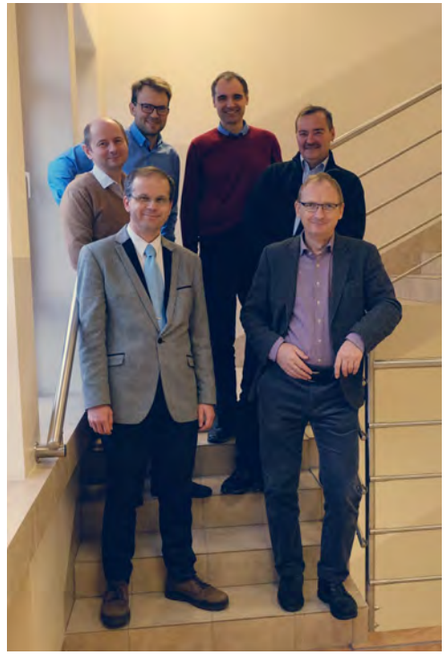
- Wroclaw University of Technology (Language Technology Centre);
- Institute of Computer Science (Polish Academy of Sciences);
- Institute of Slavic Studies (Polish Academy of Sciences);
- Polish - Japanese Academy of Information Technology;
- Lodz University; and
- Wrocław University.

The leader of the consortium is the Language Technology Centre at the Wroclaw University of Technology, which is a CLARIN B-Centre. The Polish National Coordinator is Maciej Piasecki. The team in the consortium includes a very diverse group of specialists: IT specialists, linguists, literary scholars, and specialists in library and information science.

The main goal of CLARIN-PL is to construct a technical infrastructure, tools and resources for natural language processing – especially for Polish language processing. The technical infrastructure (that is, the servers) is located at Wroclaw University of Technology at CLARIN-PL Language Technology Centre. The flagship tools and resources are:

- plWordNet, which is the biggest wordnet in the world available through the open licence together with its mapping to Princeton WordNet. It includes emotive annotation and was built in close collaboration with the valency dictionary Walenty. Read a more detailed presentation of plWordNet on page 47;
- DSpace repository, which is a large library of linguistic data and tools;
- SPOKES, which is a corpus of conversational data;
- Chronopress, which is a chronological corpus of Polish newspaper texts;
- Websty, which is a tool for the extraction of stylometric data. Most tools and resources work in the user-friendly web service technology (it does not require any software installation on the user's computer). A detailed presentation of WebSty can be read on page 44; and
- various speech recognition tools, such as Align.

[23] http://clarin-pl.eu/en/home-page/



CLARIN-PL Partners | First row (L-R): Maciej Piasecki (Wrocław University of Technology), Adam Pawłowski (Wrocław University). Second row: Roman Roszko (Institute of Slavic Studies, Polish Academy of Sciences), Krzysztof Marasek (Polish-Japanese Academy of Information Technology). Third row: Piotr Pęzik (Lodz University), Maciej Ogrodniczuk (Institute of Computer Science, Polish Academy of Sciences).

The second goal of CLARIN-PL is to raise awareness and popularise knowledge about NLP among the Polish digital humanities scholars. To this end, the Language Technology Centre has been organising a series of workshops called "CLARIN in research practice" (see page 49 for a detailed description). The consortium is also a strategic partner in many large research projects: employees of the consortium advice on the optimal use of the existing NLP tools and resources and help plan research, which gives them the opportunity to collect opinions and information about researchers' needs. In November 2017 at Wroclaw University of Technology, PolLinguaTec, a CLARIN Knowledge Centre for Polish Language Technology (Clarin K-Centre), was established. Its task is the continuation of user involvement activities.

Some of the CLARIN-PL team from the Language Technology Centre (Wrocław University of Technology).

WebSty, an Open Web-Based System for Stylometric Analysis

Written by Jan Wieczorek and Jakob Lenardič, edited by Darja Fišer

WebSty[24] is a powerful web-based system for stylometric, semantic and comparative analysis of texts. In its current implementation, the system is suited for the quantitative analysis of German, Polish, English, Hungarian, Russian and Spanish texts and is presented as an easy-to-use web interface that enables researchers to simply drag and drop the documents they want to analyse or provide links to uploaded .zip files containing the documents (Figure 16). WebSty is also integrated with the Polish D-Space based repository provided by CLARIN-PL. To ensure fast processing of the documents, WebSty is designed as service-oriented software in which each language tool runs as a separate process with pre-loaded data models. The English version of WebSty makes use of the following tools:

- SpaCy, an NLP suite that prepares texts for deep learning and features advanced annotation like Named Entity recognition;
- Fextor, a tool for the extraction of features from text collections;
- CLUTO, a tool for the clustering of datasets; and
- D3.js and D3-tip, which are the visualisation components.

After uploading the file to be analysed, researchers can use the Choice of features tab (Figure 17) to specify which linguistic features WebSty takes into account when performing the analysis. Among others, these include the specification of various grammatical classes and a host of features related to named entities. The results of the clustering are primarily visualised in the form of a dynamic dendrogram (Figure 18), which is generated on the basis of the D3.js library and involves an interactive binary tree where each subtree can be collapsed. In addition, WebSty allows researchers to download the results in the .xslx format and also to visualise them with other user-friendly methods, like a heat map, radar chart and multidimensional scaling.

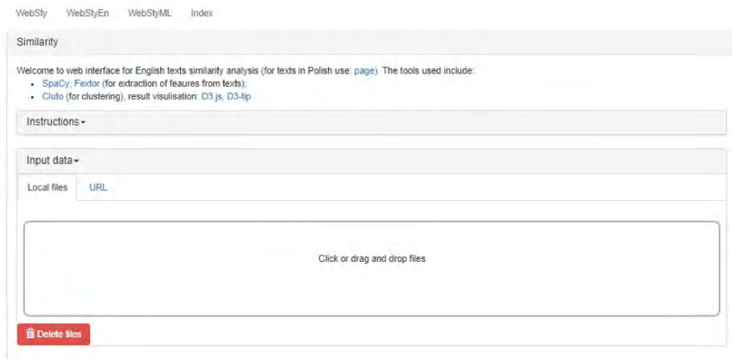


Figure 16: Uploading datasets in WebSty, where in the case of the English version researchers can either upload their own local documents or provide links to online resources. The Polish version is also integrated with the D-Space repository provided by CLARIN-PL.

[24] http://websty.clarin-pl.eu/

Since WebSty does not require in-depth computational knowledge, it is a crucial tool for fields in the social sciences and digital humanities in that it allows researchers to conduct massive-scale analyses of numerous resources, thus revealing characteristics that have been overlooked by traditional approaches. As an example of a successful application in literary studies, Maciej Maryl, who is Deputy Director at the Institute of Literary Research of the Polish Academy of Sciences, used WebSty to analyse a large collection of blogs with anonymous authorship and thereby detected subtle similarities between documents on the basis of the provided clustering options (the interview with Maryl can be read on page 50). As a successful application in sociology, Marek Troszyński from Collegium Civitas has used the tool in a project for monitoring and documenting manifestations of discrimination against the Ukrainian minority in Poland. In relation to languages other than Polish, WebSty has successfully been used by Palkó Gábor from the Petőfi Museum of Literature to analyse texts in Hungarian (Figure 19). Through cooperation with partners from the same museum, a new version of WebSty will be created with a dedicated interface in Hungarian.

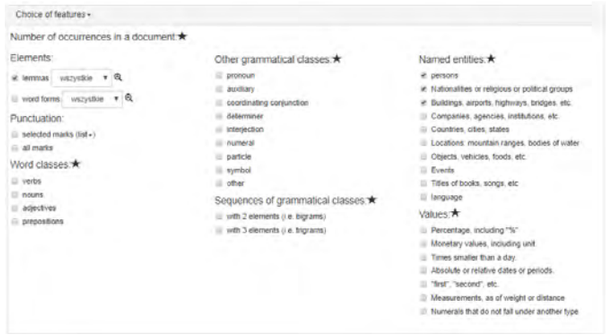


Figure 17: Choosing linguistic features for analysis.

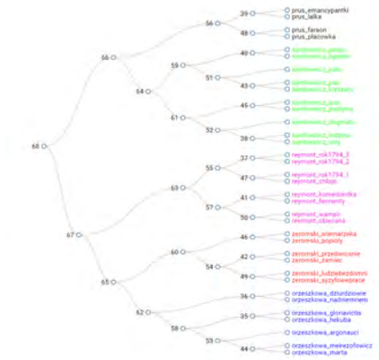


Figure 18: Clustering results (dendrogram and cluster membership) in a form of interactive dendrograms for corpus of Polish books.

Figure 19: Using WebSty to analyse Hungarian text. The visualisation shows clusters of similar texts scaled to 2D space.

Figure 20: Clustering results (dendrogram and cluster membership) in a form of circle (in the presented results two clusters were selected in contrast to results in Fig: 18 where five clusters were selected).

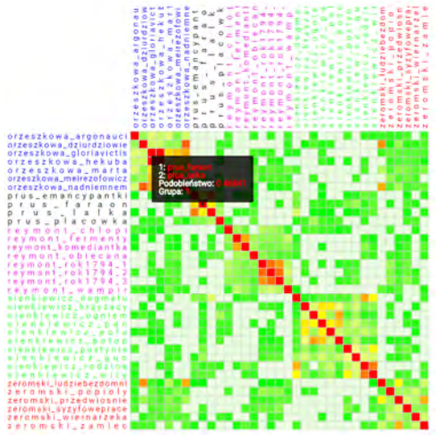


Figure 21: Similarity results in the form of a heatmap.

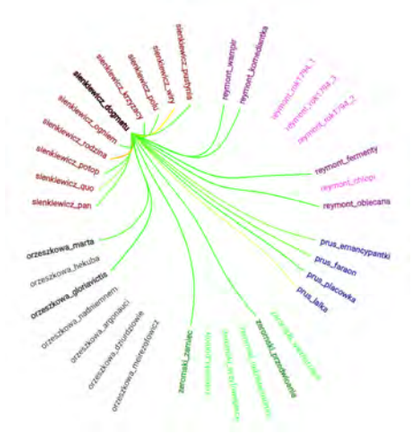


Figure 22: Similarity results in the form of a schemaball.

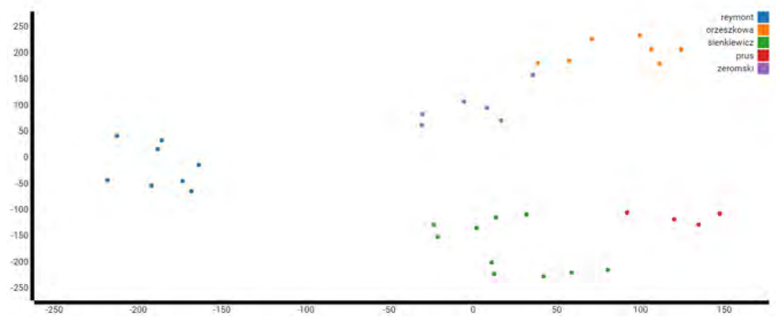


Figure 23: Distance results in the form of 2D plot.

plWordNet

Written by Jan Wieczorek, Ewa Rudnicka and Agnieszka Dziob, edited by Darja Fišer and Jakob Lenardič

plWordNet (Polish Słowosieć)[25] is a (large) lexico-semantic network reflecting (the current content and structure of) the Polish lexical system. It is a kind of dictionary in which word senses are represented by lexical units, linked by relations to create synonym sets – synsets. It is inspired by the Princeton University WordNet – the very first wordnet, which has been in development since the 1980s. Both wordnets are linked via inter-lingual relations, effectively creating a bilingual semantic network. plWordNet has been developed at Wrocław University of Technology by a team of linguists and programmers since 2006.

The meanings of lexical units and synsets are defined by relations; however, more and more units also contain a gloss and usage example that further describe their meaning. In version 3.0, certain units in plWordNet (a number of which grows progressively) are marked with sentiment values – positive, negative, ambiguous, or neutral. Version 3.1 of plWordNet, published in December 2017, includes:

- around 191,000 words (lemmas);
- around 290,000 senses (lexical units);
- around 600,000 relations that describe words and their meanings within plWordNet and around 239,000 inter-lingual relations;
- around 160,000 glosses and 70,000 usage examples; and
- around 80,000 units which contain emotive annotation.

plWordNet encompasses four parts of speech: nouns (around 177,000 senses), adjectives (around 54,000 senses), adverbs (around 14,000 senses), and verbs (around 40,000 senses) – and is being progressively expanded. In contrast with the Princeton WordNet, plWordNet is characterised by a wide range of relations both on the level of synsets and lexical units that are largely the result of the morphological richness of the Polish language.

plWordNet can be browsed online (Figure 24), via a mobile app available on Google Play, or via the WNLoomViewer application. It can be used for linguistic analyses both in Polish as well as in comparative and translation studies. Due to its open licence, which is based on the Princeton WordNet, it can also be used for data mining both in research and commercial projects.

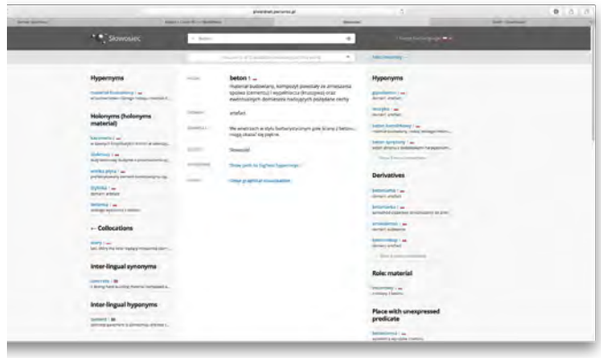


Figure 24: The interface of plWordNet.

[25] http://plwordnet.pwr.wroc.pl/wordnet/

The WordnetLoom Editor is a Java application that provides a visual, graph-based interactive presentation of the structures of plWordNet and thereby enables browsing and direct editing of lexico-semantic relations and synsets (Figure 25). It is remarkable for its flexibility and adjustability to the needs of individual users. It is currently being used by the Portuguese Wordnet team, and in a project led by Professor Ewa Geller from Warsaw University which aims to describe the Yiddish language and to map senses on the morphological and semantic level from Yiddish to corresponding senses in plWordNet, GermaNet, and the Princeton WordNet.

In 2014, plWordNet became one of the crucial parts of the semantic search engine for the Polish language called NEKST (the Natively Enhanced Knowledge Sharing Technologies), which is adapted to Polish syntax (especially flexible word order) and inflection. plWordNet served as the basis for word sense disambiguation (WSD) and the creation of links between words, and was also used to develop an anti-plagiarism system, based on the NEKST search engine.

For further reading, here is a list of the key publications on plWordNet:

Piasecki, M., Szpakowicz, S., and Broda, B. (2009). A Wordnet from the Ground Up. Wrocław: Oficyna Wydawnicza Politechniki Wrocławskiej.

Maziarz, M., Piasecki, M., Rudnicka, E., Szpakowicz, S., and Kędzia, P. (2016). plWordNet 3.0 – a Comprehensive Lexical-Semantic Resource. In Calzolari, N., Matsumoto, Y. & Prasad, R. (editors), COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, 2259-2268.

Figure 25: The interface of the WordnetLoom Editor, visualising the wordnet structure of the lemma “złodziej” (“thief”).

“CLARIN-PL in Research Practice” - a Lecture and Workshop Series

Written by Darja Fišer and Jakob Lenardič

Since April 2015, CLARIN-PL has organised a series of workshops and lectures titled “CLARIN-PL in Research Practice”. Eight editions of the series have taken place thus far, all of which were well attended, with around 40 participants per event. The latest event in this series took place in Wrocław between 19 and 20 June 2018.[26]

The goal of the workshops is to present as well as demonstrate the use of the tools and resources developed by CLARIN-PL. The participants, who come from various social sciences and humanities backgrounds, such as literary theory, sociology, psychology and history, get to learn how to create and analyse their own corpora and dictionaries as well as learn the foundations of statistical analysis at the workshops. CLARIN-PL thereby raises awareness of what their infrastructure has to offer and promotes novel research that can only be afforded by the use of NLP tools. Moreover, through the workshops the members of CLARIN-PL have themselves learnt what the expectations and needs of potential users are in relation to the presented tools. Every workshop has involved a large number of participants who were always eager to provide their own important perspectives and offer potential solutions.

Some participants have become regular users of CLARIN-PL services, and the consortium has thus become a technological partner in numerous successful research projects. To give a notable example, the first workshop in April 2017 led to the successful collaboration between CLARIN-PL and the Digital Humanities Centre at the Institute of Literary Research, which lasts to this day and has resulted in important empirical results in the field of literary studies, such as the creation of an interactive literary map of Warsaw, and work on state-of-the-art digital services dedicated to researching literature, such as the Literary Exploration Machine, which brings together various computational tools for literary analysis and exploration in the form of a single user-friendly online environment.

[26] http://clarin-pl.eu/pl/viii-cykl-warsztatow-wroclaw/

Maciej Maryl

Maciej Maryl is the Deputy Director of the Institute of Literary Research of the Polish Academy of Sciences. The following interview took place via Skype on 16 January 2018 and was conducted and transcribed by Jakob Lenardič and edited by Darja Fišer.

1. Could you please briefly introduce yourself? What inspired you to start studying literature and to take an empirical approach toward it?

I became interested in applying empirical methodologies of social sciences to literary studies as an MA student at the University of Warsaw. I tried to quantify the way people read and approach texts, which eventually led me to computational methods. My PhD, which I defended in 2013, was dedicated to the influence of electronic media on literary communication.

2. How did you get involved with the Polish CLARIN consortium? Are you currently collaborating with them?

In 2013, when I was in the process of setting up the Digital Humanities Centre at the Institute of Literary Research at the Polish Academy of Science, I was introduced to Maciej Piasecki, who is the coordinator for CLARIN-PL. At the same time, the Institute was organising the first THATCamp (The Humanities and Technology Camp) in Warsaw, so I invited him to present the tools developed by the Polish consortium. I was inspired by his talk and wanted to use the tools in my analyses of weblogs that I was conducting at the time. This in turn led to a very fruitful collaboration between CLARIN-PL and our institute, which goes on to this day. We have successfully cooperated on quite a number of projects. To name a few, one ongoing project involves the creation of the Literary Map,[27] in which geographical information that appears in Polish literary texts is mapped onto Google Maps. Another project is LEM (Literary Exploration Machine),[28] an online system that brings together various tools dedicated to processing and analysing literary texts. We have also started two lexicographical projects. One aims at creating a dictionary of Polish Romantic poets, using CLARIN-PL tools and WorldNet, while the other, in cooperation with many institutions, is dedicated to linking together various historical dictionaries of Polish on a single platform. In most cases, we help develop the tools that CLARIN-PL had already created, providing the expertise and needs of our field. This helps to establish a productive feedback loop between developers and users.

3. Which CLARIN services would you recommend to your colleagues working in Literary Studies?

I would especially recommend LEM. One of the biggest problems of novice literary scholars who want to conduct computational research is the lack of expertise in using linguistic tools. In other words, novice researchers are faced with elaborate and sophisticated tools which are simply overwhelming, especially for researchers without a computational background. LEM helps researchers to overcome this problem because it pools together a variety of tools into a single workflow and supplements them with detailed descriptions, which makes them user-friendly, even for beginners. Work on LEM is an ongoing process, and we are currently planning new features like topic modelling and description of case-studies which will enable a better understanding of tools.

[27] http://clarin-pl.eu/en/literary-map/
[28] http://ws.clarin-pl.eu/lem.shtml?en

4. Your website says that you are involved in the following projects – "The Polish Literary Bibliography" and "Blogs as a new form of multimedia writing". Could you describe them? How do they benefit from the CLARIN infrastructure?

The Polish Literary Bibliography is an ongoing project we run in cooperation with Poznań Supercomputing and Networking Centre. We use CLARIN-PL's INFOREX to extract structured information from scanned volumes of bibliographical records and incorporate them into a multipurpose online research platform. We are aiming to extract bibliographical data from printed volumes ranging between 1945 and 1988, and we are currently trying to work around some problems, such as the low quality of print, which makes parsing more difficult.

"Blogs as a new form of multimedia writing" is actually the project that marks the beginning of my collaboration with CLARIN-PL. Together with the Polish consortium we worked on the tools used to classify weblogs on the basis of their genre. To give some background, what we did at first at the Institute – that is, before involving CLARIN-PL – was to draft a typology of weblog genres based on a systematic, qualitative analysis of actual texts. We then started the cooperation with Maciej Piasecki in order to corroborate our findings with computational methods. We applied various clustering methods, using tools like CLUTO and CLARIN-PL's stylometric system WebSty (read more about this tool on page 44) to see whether they would group the weblogs together in accordance with our proposed typology. Together with Maciej Piasecki and Ksenia Młynarczyk, we have written an article dedicated to combining close reading with distant reading on the basis of CLARIN-PL tools.

However, weblogs are tricky when it comes to the application of computational methods. The main obstacle is that individual blogs are far from homogenous in terms of style and other linguistic characteristics, as they consist of many different posts. So the classification of genres did not yield satisfactory results – we were most successful with cooking blogs, which are characterised by very specific language. That is why we currently work on shifting the unit of analysis from entire blogs to individual posts, in order to get more accurate results.

5. What are the main advantages of taking a digital humanities approach to literary history? Can a quantitative approach help uncover answers to more traditional questions that are at the heart of literature, such as the political and sociological aspects of writing, the value of the literary canon, etc.?

There seems to be a consensus in the field that the application of computational methods actually involves a two-fold approach. First, computational tools and methodologies may be used to corroborate existing claims in the field, i.e. to see if we can arrive at similar results with empirical methodologies. And this is what we are doing in the blog project right now. Second, once we establish that our computational approach yields significant results, we may use it to uncover aspects of writing which are too difficult to assess by means of traditional non-computational methodologies, such as the problems of authorship or language change in literary history.

For me personally, a computational approach is important because it allows me to see a wider picture of the research field. However, we should not take computational results for granted. What I believe is crucial in using DH tools is that at some point we should return to actual texts in order to understand the computational results fully. In other words, the main advantage of working in digital humanities is the multifaceted approach that combines distant reading via the computational tools with the close reading. I think that both methodologies should be intertwined in a research workflow.

6. Can you discuss how the Internet has shaped the contemporary literary scene, especially that of Poland? How do literary historians and critics, particularly in your country, evaluate new forms of writing, such as fiction published through non-traditional media like blogs and forums, in relation to the older, more traditional printed forms?

There is of course a division between researchers who are dedicated to solely working with traditional texts and a relatively smaller group which also focuses on digital writing. However, I do think that more and more studies are beginning to focus on new textual phenomena, and sooner or later we just have to research them together as it is hard to talk about contemporary literature if you disregard digital writing. For instance, weblogs became popular in Poland around 2006, so slightly more than 10 years ago, and at first they served almost exclusively as a social medium through which people tried to connect with friends or write about their lives. However, blogs evolved over the years – partly thanks to social media which took over the function of the main platform for personal communication – and to some extent they now resemble print media like magazines, newspapers or books. In this process weblog genres have crystallised and now serve as a very interesting research object, especially given their accessibility for computational analyses, as one does not have to digitise them beforehand.

As for actual fiction writers, there has also been some change in the way writers make use of digital communication. When I started doing research for my PhD thesis around 10 years ago, a rule of thumb was that the more popular a writer was the more limited online presence he or she maintained. Popularity meant access to mainstream media, and that used to be enough not so long ago. Nowadays, however, there are many very successful writers who use Facebook or run their own blogs to cultivate relationships with readers. When it comes to actual experiments with literary form – such as electronic poems or interactive novels – there are many examples of interesting texts, but the majority of writers remains quite conservative and tends to stick to traditional forms. As popular interest in interactive narratives is captured by computer games, in literature there seems to be greater demand for traditional, stable, linear and finite narratives. Perhaps this is, as Umberto Eco observed 20 years ago, a real power and value of literature in the times of interactivity – it provides narratives that cannot be manipulated according to the readers' will.

7. How do your students and fellow researchers embrace the digital humanist approach? How are digital humanities in general represented in the Polish academic environment?

There are still quite a few scholars who think that using computational approaches shifts your attention from the actual texts to the linguistic surface. I actually believe that this kind of scepticism in Polish academia is quite a widespread phenomenon due to the idea that digital approaches are reductionist and ill-suited for addressing the "big", critical questions of literary studies. But we shouldn't forget that similar reservations have been formulated against empirical approaches in the humanities, probably since the birth of anti-positivism. So, we should have probably got used to it by now. However, in the last five years, digital humanities have begun to flourish in Poland, entering the phase of institutionalisation. Related research centres were established, and researchers established the CLARIN-PL and DARIAH-PL consortia. CLARIN-PL is especially eager to bridge the gap between computational experts and humanities users, organising hands-on workshops for researchers and translators. So, I expect the body of digital humanities research to grow, but let's not fool ourselves – it is not going to be mainstream. What we need is more digital humanities courses at universities, so the base of practitioners could steadily grow. Obviously, there are many courses in linguistic departments, but we should also reach out to students of history, literature, and cultural studies. My institute has just received a grant to start a graduate program on digital literary studies to enable the study of literature with the help of digital methods and technologies at the PhD level. This program will be carried out in cooperation with the Polish-Japanese Academy of Information Technology, which is also a member of CLARIN-PL.

8. What would you recommend CLARIN do in order to attract more researchers from your community? How do you envision the future of the Polish CLARIN consortium?

CLARIN-PL is very active in terms of attracting new researchers. It has already organised a series of workshops and we are proud to have hosted the first CLARIN-PL workshop in 2015. However, I believe a more structured approach to outreach is needed – that is, a long-term involvement of users through a more established educational program that would complement the workshops with something like additional online courses. Such a program could maintain researchers' interest after the events. We also need to continue making the interfaces of the tools more user-friendly, with better documentation guiding users through the research process. What could also help is a presentation of successful case studies from a variety of fields that could serve as a guidance for further research. The future that I envision for CLARIN-PL is one where more and more new researchers join its user network. One of the best things about the consortium is that it always addresses the needs of the end user. I think it can only be a good thing if more institutions and individual researchers who want to perform computational analyses but currently lack the tools or expertise needed reach out to CLARIN-PL.

Warsaw, Poland | photo by Jacqueline Macou | Pixabay

DLU/Flanders

Written by Catia Cucchiarini, Ineke Schuurman and Griet Depoorter, and edited by Darja Fišer and Jakob Lenardič

CLARIN DLU/Flanders[29] is a founding member of CLARIN and represents Flanders, the Dutch-speaking part of Belgium. The consortium consists of

- the Dutch Language Union (Nederlandse Taalunie - DLU);
- the Dutch Language Institute (Instituut voor de Nederlandse Taal – INT);
- the Centre for Computational Linguistics (Centrum voor Computerlinguïstiek – CCL – University of Leuven);
- the Language and Translation Technology Team (LT_3 - University of Ghent);
- the Computational Linguistics & Psycholinguistics (CLiPS) research group (University of Antwerp);
- the PSI Speech Group (ESAT-PSI – University of Leuven); and
- the Language Intelligence & Information Retrieval research lab (LIIR – University of Leuven).

Since Flanders is not a country but a region, it did not qualify as a member in CLARIN and is therefore represented by the Dutch Language Union (DLU), an international language policy organisation. The consortium is coordinated by the Dutch Language Institute (INT) and the national coordinator Griet Depoorter. The INT is a starting point for anyone who wants to know anything about the Dutch and Flemish languages through the centuries. The institute takes a central position in the Dutch-speaking world as a developer, manager and distributor of sustainable language resources, using reliable scholarly methods, and is a certified CLARIN B-Centre. For instance, the INT produced the Woordenboek der Nederlandsche Taal (Dictionary of the Dutch language), an enormous historical dictionary which describes Dutch words from 1500 to 1976. A few other examples of resources and tools that are available at the INT:

- the Dictionary of Contemporary Dutch (Algemeen Nederlands Woordenboek), which is online and corpus-based;
- the Reference Lexicon Dutch (Referentiebestand Nederlands), which contains 50,000 frequent Dutch words, enriched with linguistic information;
- the Dutch Parallel Corpus, a high quality sentence-aligned parallel corpus of 10 million words for the language pairs Dutch-English and Dutch-French;
- the Word list of the Dutch Language (Woordenlijst Nederlands), a list of words in the correct official spelling; and
- Blacklab, an open source corpus search engine built on top of Apache Lucene.

Apart from providing state-of-the-art language resources and tools, the consortium is also active in involving both students and researchers in its activities. For instance, a workshop was held at the Dutch Language Institute in October 2017, aimed at familiarising digital humanities researchers with the resources and tools the consortium offers. You can read more about the workshop on page 59.

Griet Depoorter, National Coordinator of CLARIN DLU/Flanders, and Vincent Vandeghinste, National User Involvement Representative.

[29] http://www.ivdnt.org/

Text2Picto and Picto2Text tools

Written by Griet Depoorter, Darja Fišer, Jakob Lenardič, Ineke Schuurman and Leen Sevens

Text2Picto and Picto2Text[30] are two complementary translation tools aimed at enhancing communication for people with reading disabilities. Both tools have been developed by the Centre for Computational Linguistics at the University of Leuven. Text2Picto translates sentences into pictographs – that is, graphic symbols that serve as stand-ins for verbal communication – while Picto2Text does the reverse by allowing users to select the pictographs that they want to translate into written text. Two different sets of pictographs are available for both tools – the Beta set and Sclera set. The symbols in each set are designed to be very concrete and easy to interpret, so their use reduces the cognitive complexity of reading e-mails, web pages, chats and work documents. Currently, the tools are available for Dutch, Spanish and English, but other languages can be used as well if there is a Wordnet available for them. Figure 26 shows the use of Text2Pico based on the Beta set and Figure 27 on the Sclera set.

Work is now being done by Leen Sevens from the University of Leuven to add additional features to the tools. Demos of these features are already available for Dutch:

- Spelling Correction, which is a crucially important feature since the users of Text2Picto often spell phonetically;
- Word-Sense Disambiguation, which identifies the correct sense of polysemous words and retrieves the correct pictograph for that sense; and
- Text2Picto + Simplification and Temporal Detection for Dutch (2017), which adds pictographs depicting temporal relations between the other symbols and simplifies syntactic structure (Figure 28).

The tools have generated a lot of interest and are already being used by their target audience. They were awarded the prestigious Language Industry Award in 2016 and have been implemented into the WAI-NOT website,[31] which allows people with mental disabilities to use the Internet within an accessible online environment. By using the pictographs, they can play games and chat even if they aren't able to read. There is a YouTube video available in Dutch that demonstrates the implementation in the WAI-NOT website. Additionally, the software is also part of the ABLE social services app, which is available through Google Play.

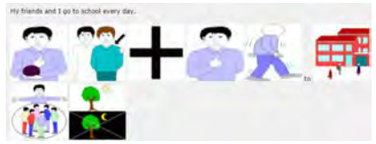


Figure 26: Translation of the sentence "My friends and I go to school every day" into the Beta pictograph set.

Figure 27: Translation of the same sentence into the Sclera set.

Figure 28: Translation of the Dutch sentence "Ik zal de rode wijn drinken die mijn moeder gekocht heeft" ("I will drink the wine that my mother has bought") into the Sclera set. Note the second pictograph in the upper row indicating future time reference corresponding to "zal" ("will"), and the middle pictograph in the lower row indicating the past time reference of buying the wine.

[30] http://picto.ccl.kuleuven.be/
[31] https://www.wai-not.be/

The Corpus of Contemporary Dutch

Written by Griet Depoorter, Katrien Depuydt and Hans Westgeest, edited by Darja Fišer and Jakob Lenardič

The Corpus of Contemporary Dutch (Corpus Hedendaags Nederlands - CHN)[32] is a collection of more than 800,000 texts taken from various sources, such as newspapers, magazines, news broadcasts, legal writings, and books, for the period between 1814 and 2013.

Since 1994, the Institute for Dutch Lexicology (which transformed itself into the Dutch Language Institute) made several corpora of contemporary Dutch available online: the 5, 27 and 38 Million Word Corpora and the Dutch Parole Corpus 2004. These older corpora were merged and a considerable amount of more recent material was added from the NRC Handelsblad, which is a Dutch newspaper, and De Standaard, which is a Flemish newspaper. Other sources that were added came from Suriname and the Netherlands Antilles (where Dutch is also an official language), such as newspapers, material published on internet (blog, website) and books written by Surinam authors. This collection of data became the Corpus of Contemporary Dutch, which serves as the first step towards a monitor corpus for contemporary Dutch.

The corpus contains approximately 440 million tokens:

- 224 million Dutch Dutch;
- 185 million Belgian Dutch (Flemish);
- 14.4 million Dutch as spoken in the Antilles; and
- 18.3 million Surinamese Dutch.

The corpus has been lemmatised and PoS-tagged. The CHN can be searched via a simple search interface and via CQL, and users can search for or filter by five criteria: title, author, year of publication, medium and language variety. The possible values for the last criterion are NN (Dutch from the Netherlands), BN (Dutch from Belgium), SN (Dutch from Suriname) and AN (Dutch from Netherlands Antilles). The software powering the CHN website was developed at the Dutch Language Institute during the course of the IMPACT and CLARIN projects, and the corpus search is powered by BlackLab. Another great advantage of the Corpus of Contemporary Dutch is that it continues to grow, and during 2018 a significant amount of new data will be added (among others, newspaper data from the period between 2014 and 2017).

Figure 29: Searching for all word forms of the lemma "gezellig" ("cosy") in Flemish newspapers from 2000 to 2010.

[32] http://chn.inl.nl/

The corpus data have already been successfully used in linguistic research. Jaspers et al. (2015) used CHN data when researching the syntactic and semantic characteristics of Dutch scalar modifiers denoting small degrees (like few in English), while Devos (2016) used the corpus to investigate a special category of Dutch infinitival phrases that act as obligatory modifiers of nominal predicates and add a causative meaning to the clause. In addition, the corpus has also served as the main source of data for a number of students' theses. As examples of using the corpus in student work, Saskia Lubrun at the University of Leiden researched the collocational properties of the Dutch subjunctive, while Wanda Polak at the University of Amsterdam investigated the phonological contexts of several Dutch suffixes.

References:

Devos, F. (2016). 'Joost was het gewoon om dergelijke zinnen te analyseren'. Over beknopte bijzinnen als oorzakelijk object in het Nederlands. In Een Sextant Voor Een Taalspecialist : Bijdragen Tot Joost Buysschaert in Profiel, 39–44.

Jaspers, D., van Craenenbroeck, J., and Vanden Wyngaerd, G. (2015). De syntaxis en semantiek van diminishers in het Nederlands en het Duits. https://static1.squarespace.com/static/5217e223e4b090faa01f8f2d/t/5637cab6e4b0e17c588d8538/1446496950714/paper_ewa_finaal.pdf

Lubrun, S. (2015). De collocaties van de Nederlandse subjunctief. https://openaccess.leidenuniv.nl/handle/1887/35053

Polak, W. (2015). In welke fonologische context komt afleiding met de achtervoegsels -ig, -erig en -achtig voor? http://www.fon.hum.uva.nl/archive/2015/2015-BA-WandaPolak.pdf

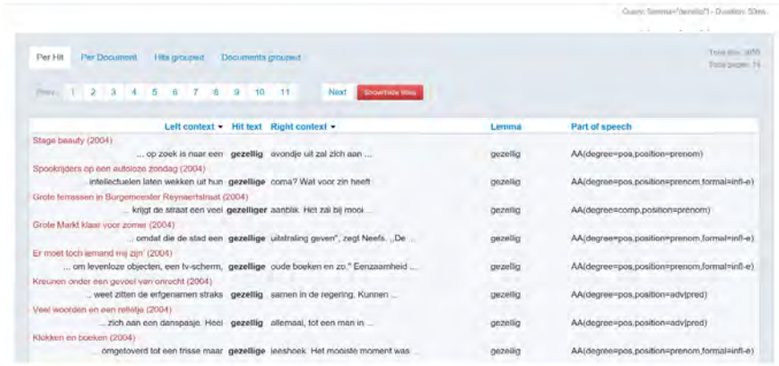


Figure 30: Some occurrences of the lemma "gezellig".

Workshops for Introducing Digital Humanities Researchers to the CLARIN Services & Resources

Written by Griet Depoorter and Katrien Depuydt, edited by Darja Fišer and Jakob Lenardič

In 2016, the Institute for Dutch Language (INT) sent out an information sheet and a questionnaire to Flemish research groups in the field of humanities. The goal was to promote CLARIN and the INT (as a CLARIN Centre), to get an overview of the usage of the CLARIN infrastructure and services, and to receive input concerning the expectations of the user community about CLARIN and the INT. After receiving the feedback, the director of the INT, Frieda Steurs, conducted talks with a number of these groups (for instance, the Language Group Flemish Sign Language at the University of Leuven, the Ghent Centre for Digital Humanities at Ghent University and the Department of Linguistics at the University of Antwerp) to receive further in-depth information. What the centre learnt this way is that the linguistic community in Flanders would like the CLARIN consortium to expand existing datasets (e.g. the Dutch Parallel Corpus) or create new ones, like a corpus of spoken (or written) Dutch and videos of the corresponding Flemish sign language. In addition, the INT CLARIN Centre should serve as a knowledge centre for standards and annotation protocols, as well as offer expertise and support for researchers who have little or no experience with digital research.

As a result of this knowledge sharing and gathering initiative, the INT organised the first training workshop[33] in October 2017 in Antwerp in cooperation with Digital Humanities Flanders (DHuF) and the Faculty of Arts of the University of Antwerp. An invitation was sent out to all DHuF members, and the event announcement was also published on the websites of DHuF and the INT. Eventually 32 researchers, primarily historians, from Antwerp, Leuven, Ghent and the Netherlands attended the workshop.

The workshop showcased what the INT, which had then been just established, can offer digital humanities researchers, especially those researchers who use historical language material. The event started with a presentation of CLARIN in general and the INT as a CLARIN centre, and continued with an introduction into corpus building, focusing on historical corpus building in particular. The best practices with regards to building historical resources were discussed and exemplified with a concrete use case; namely, with the Nederlab text collection. Nederlab is a web environment for researchers and students who study the evolution of the Dutch language, literature and culture. The website offers millions of pages of (historical) Dutch texts that can be researched and analysed with user-friendly text analysis software.

Another topic of the workshop was the enrichment of corpus material. The specific challenges of annotating historical texts were discussed; for instance, the fact that standard (modern) tag sets cannot be applied to historical texts, and that tokenisers cannot handle clitics well. Furthermore, the webservice INL Labs was demonstrated, which linguistically annotates (historical) texts (in various input formats). INL Labs uses two annotation tools: the Stanford NE tagger and the INT-developed tagger-lemmatiser for historical Dutch.

The workshop also demonstrated how historical data can be searched with Blacklab, a corpus retrieval engine which is available as a webservice and as a Java library. The tool allows fast, complex searches with accurate hit highlighting on large, tagged and annotated bodies of text and it can be used to search historical corpora like the Corpus Gysseling and Letters as Loot. Users were also shown how to search their own data by means of Autosearch, which is powered by the Blacklab Engine.

Finally, there was a presentation on the work on the diachronic computational lexica of the INT, GiGaNT and DiaMaNT, and what the benefits would be of having the data available as Linked Open Data. This was the first of several workshops that the INT CLARIN Centre plans to give in Flanders. On March 20, the consortium organised an information session that involved humanities researchers from the University of Leuven. During the session, the consortium gatheredinformation on what the specific needs of the attending researchers are in relation to the services provided by the INT and CLARIN in general. Based on their feedback, the DLU then plans to organise follow-up workshops that will be specifically tailored to the needs of focused research groups with shared interests.

[33] http://uahost.uantwerpen.be/platformdh/index.php/event/int-workshop-antwerp/

Cora Pots

Cora Pots is a PhD student in the Quantity and Quality in Linguistics project at the University of Leuven. The following interview took place via Skype and was conducted on 19 March 2018 and transcribed by Jakob Lenardič, edited by Darja Fišer.

1. Could you briefly describe your academic background and your current position?

My interest in linguistics started when I was studying for my Bachelor's Degree at Utrecht University. It was a two-track programme in modern literature and linguistics. In the linguistics track, I was working in the generative framework, primarily child language acquisition and syntax. I did an internship where I researched speech perception and language development in younger children, which really sparked my interest in research and inspired me to pursue it as a career choice. In my Master's degree studies I slightly changed the focus of my research and began investigating syntactic variation in Germanic languages. When I became a Research Assistant at the Meertens Institute, under the supervision of Professor Sjef Barbiers, I began combining my formal background with a computational approach to linguistics and co-wrote the Educational Module for the MIMORE tool (described in detail on page 34), which is used to investigate the morphosyntactic variation of Dutch dialects. After obtaining my Master's degree, I worked on various projects; for instance, I was a part-time lab manager at the Babylab at Utrecht Institute of Linguistics (Utrecht University), and I also worked for the AnnCor project – a project to make CHILDES, a large collection of corpora of child language, syntactically searchable. In 2016, I started my current position, which is a four-year PhD track at the University of Leuven, so I'm halfway through now.

2. How did you start collaborating with CLARIN DLU? Could you briefly describe the project you're currently involved in, called Quantity and Quality in Linguistics: Reverse Dialectometry?

I started using the tools and services of CLARIN DLU when I became a PhD student at the KU Leuven (Catholic University of Leuven). In the project that you mention, I investigate the formal properties of Dutch dialects/regiolects as spoken both in Flanders and the Netherlands, for which I use tools and resources provided by the Dutch and the Flemish consortia (OpenSoNaR, MIMORE).

3. Why is investigating non-standard language valuable for linguistic theory?

I'd like to answer this question by exemplifying a syntactic phenomenon related to infinitives in verb clusters. In Dutch dialects, the position of the infinitival marker "te", which is the equivalent of English "to", varies in the sense that in some dialects it gets doubled in a verbal cluster (for instance, "te zitten te werken", literally "to sit to work"), while in other dialects one of the markers gets dropped either in the first ("_ zitten te werken") or in the second position ("te zitten _ werken"). This is a linguistic fact that you wouldn't be able to observe if you studied only the official variant of Dutch, which is kept in check by the prescriptive rules, so such empirical data from the dialects actually give you a far more complex insight into the grammatical structure of Dutch infinitives. Additionally, MIMORE also shows the geographical distribution of the variation, which then allows you to investigate other possible grammatical phenomena that are also tied to the same pattern. Needless to say, without a tool like MIMORE it would be impossible to attain such insights into the linguistic structure of the Dutch language.

4. Given that you are an early-stage researcher, could you share your experience of how CLARIN can support researchers who are just starting their research and career? Do you have any advice for your fellow novice researchers?

What is great about CLARIN is that it allows you to explore a wealth of data that are already collected. Whenever you start working on a research project, you normally don't have any idea what's actually going on in the linguistic data. However, using a tool like MIMORE, you can quickly start working on a topic without having to do the field work yourself, which would of course be extremely time consuming. What is more, such resources have already been parsed and annotated by experts, so this is another aspect of CLARIN that I find amazing; it allows you to start applying its tools and resources fairly quickly, even if you don't have a lot of technical skills or a computational background.

My advice is to simply start using the available tools and resources, no matter whether you're a student or a more advanced researcher. The main problem, I think, is that not many people are aware of the research possibilities that CLARIN tools and resources afford. I know that there are many young researchers who study Dutch dialects, for example, and who would greatly benefit from using the CLARIN resources like the MIMORE databases, which are a goldmine of data. In this respect, these databases are very valuable since they not only consist of Dutch data as spoken in the Netherlands, but also all the Flemish dialects.

5. What can the Flemish consortium offer researchers working in the generative tradition?

I believe that GrETEL,[34] which is a tool that is developed by the Flemish consortium, has proven itself to be a very valuable service for a generative grammarian. What GreTEL primarily does is it allows you to efficiently search for specific syntactic constructions in the MIMORE databases without having to rely on technical knowledge about complex query languages. Normally, you would have to spend a lot of time searching a database for all the variants of a specific syntactic construction, but with GrETEL you only input an example of your own that conforms to the syntactic pattern you're interested in, and you immediately get all the relevant data.

6. Is working with a research infrastructure an established practice in your research community?

Well, it depends on what you consider my research community. As far as my fellow PhD students are concerned, a lot of them indeed make use of research infrastructures. Within the formal framework, however, researchers use it infrequently. Though I don't think it should be obligatory for generative grammarians to use corpus data in all their work, I still believe that many non-empirical researchers would still find it very helpful if they checked their claims in corpora. For one thing, corpus data can show that your intuition about a linguistic phenomenon isn't really all that representative across dialects. On the other hand, using a corpus-based approach can provide a very good stepping stone for a beginner, since it quickly shows you what the relevant linguistic situation looks like, from which you can then move on to making formal claims.

[34] https://www.clarin.eu/showcase/gretel-search-engine-querying-syntactic-constructions-treebanks

7. Since you are also involved in teaching, which is a major priority of CLARIN's user involvement initiative, can you tell us how you integrate the resources and tools provided by the Flemish consortium in your courses? Do you have any suggestions how the link between CLARIN and university curricula could be strengthened?

I teach a Master's course on syntactic variation in Dutch and Flemish dialects with my supervisor. We usually spend one class showing the students how to use MIMORE. We ask them to pick a specific formal topic or problem and then investigate how the formal claims correspond to the data in MIMORE. We then show the students how to use such data in their writing assignments. The main goal of the course, which I think is an important one, especially in terms of bridging the gap between the empirical and formal worlds, is how to work with a large dataset (MIMORE is comprised of 267 dialects) and apply the empirical data to a formal analysis, which is a far from trivial problem.

In the end, students really like this approach because it often allows them to get fairly novel results, even at the beginning stages. And many syntactic topics would be impossible to tackle were it not for these tools. We also use GrETEL to help students write their Bachelor's and Master's theses.

As for university curricula in general, I think the main problem is that only a few teachers have experience of combining both worlds – that is, formal analyses with empirical research. I think the first step that must be made is to encourage professors, post-docs and PhD students who also teach to become aware of these tools and resources and show them how to implement them in their courses. I think that guidelines like the Educational Module could really help in this regard.

8. Given that you have experience with two CLARIN consortia (apart from CLARIN DLU, you previously worked with CLARIN-NL), could you describe how the two complement each other?

What must be understood is that the division between the Dutch speaking part of Belgium and the Netherlands is a political state of affairs that does not correspond to the division of dialects. That is, dialects do not know political borders. However, tools and resources like MIMORE and GrETEL, which were often developed in collaboration by researchers working with both consortia, also contain data from Flemish dialects along with the data from Dutch as spoken in the Netherlands.

Consequently, such tools which transcend borders in this sense are really the only way to get an accurate linguistic representation of our language, and it's for this very reason that in our courses/thesis supervision at the KU Leuven we use both GrETEL, which is "our" tool, and MIMORE, which was developed by CLARIAH-NL. Additionally, the Educational Module that showcases MIMORE and GrETEL is still being updated by Sjef Barbiers, who is from the University of Leiden, and Ineke Schuurman and Liesbeth Augustinus, who work at the KU Leuven Leuven, which I think is a great cross-border collaboration. Consequently, I see no reason why other consortia should not also collaborate in a similar manner, especially if the languages in question are similar.

9. What would you say is the first thing CLARIN should do to be even more useful for researchers in your field?

I would find it really wonderful if researchers could use tools like GrETEL and MIMORE to search for historical variants of Dutch (dialects). I also know that there is a lot of dialect material that cannot be accessed yet, which is something that I would like to see available through CLARIN one day, but I understand that this is often related to copyright problems.

Bruges, Belgium | photo by Waldo Miguez | Pixabay

Czech Republic

Written by Darja Fišer and Jakob Lenardič

The Czech consortium LINDAT[35] is a founding member of CLARIN ERIC. It is a B-certified centre that involves four Czech research institutions – the Department of Cybernetics at the University of West Bohemia, the Institute of Formal and Applied Linguistics at Charles University, the Czech Language Institute at the Czech Academy of Science, and the NLP Centre at Masaryk University. The consortium is led by Professor Eva Hajičová.

The consortium offers a pioneering repository for language resources, whose architecture serves as the backbone of several other CLARIN repositories. The repository rigorously follows best practices on metadata presentation, so it is ensured that all language data are safely stored with clear documentation as well as outfitted with guidelines on proper citation. Many of the monolingual, parallel and speech corpora within the repository can be accessed through the concordancer KonText, which is a flexible search environment that allows users to perform queries of various complexities – from simple searches by lemma or word form to using CQL – as well as save search results for future research.

LINDAT also offers an integrated environment for storing, building, searching and visualising treebanks, which are databases of syntactically annotated sentences. As a pivotal tool for treebanks, LINDAT offers PLM Tree Query, through which researchers can browse a great variety of treebanks in 61 languages. For the novice researcher, the Tree Query is accompanied by a step-by-step tutorial that shows how to execute searches in the query language. Together with the Norwegian INESS, LINDAT is a CLARIN Knowledge Centre that specialises in the creation and maintenance of treebanks.

[35] https://lindat.mff.cuni.cz/en

LINDAT Team | Back row: Pavel Straňák, Jaroslava Hlaváčová, Pavel Pecina, David Mareček, Ondřej Bojar, Jan Hajič, Milan Fučík. Front row: Barbora Hladká, Anna Nedoluzhko, Vendula Kettnerová, Eva Hajičová (national coordinator), Anna Vernerová, Veronika Kolářová, Magda Ševčíková, Marie Křížková.

LINDAT actively works on introducing its state-of-the-art language technologies to researchers both within computational fields like NLP and within the digital humanities and social sciences. To this end, LINDAT organised a user involvement workshop on 24 April 2018 in Prague, which aimed to showcase how technological infrastructures are also relevant beyond the computational framework. You can read more about the workshop on page 70.

Prague, Czech Republic | photo by Studio Reasons | Unsplash

UDPipe

Written by Barbora Hladka and Jakob Lenardič , edited by Darja Fišer

UDPipe[36] is a state-of-the-art tool pipeline which performs several complex annotation tasks: tokenisation, Part-of-Speech tagging, lemmatisation, sentence segmentation and dependency parsing, all to a high degree of precision. The architecture of UDPipe employs a deep neural network and is trained on language models from the Universal Dependency treebanks provided by LINDAT (see page 68 for a presentation of the Universal Dependencies). UDPipe can be used to annotate and parse texts from over 50 languages, many of which are non-Indo-European, such as Arabic, Irish, Indonesian and Tamil. It was (and is being) developed at the Institute of Formal and Applied Linguistics at Charles University, and can be freely used for non-commercial purposes.

UDPipe is available both as a downloadable program that is compatible with Linux, Windows and OS X, as a library in programming languages such as C++, Python, Perl, R, Java, C#, and as an easy-to-use web application. Researchers who wish to run UDPipe as a standalone program on their own computers must also download one of the Universal Dependencies language models, which are described in detail in the UDPipe User's Manual:

- the Universal Dependencies 1.2 models, which contain cross-linguistically consistent treebank annotation models for 33 languages;
- the Universal Dependencies 2.0 models, which are an updated version of the former and contain annotation models for over 50 languages; and
- the CoNLL17 Shared Task Baseline UD 2.0 models, which contain a different version of the Universal Dependencies 2.0 models.

The UDPipe Web Application is provided through the LINDAT architecture. It is very easy to use in the sense that researchers need only select one of the many languages in one of the three training models and input the text (or upload whole files) they wish to have annotated. The results can either be visualised in the form of a tree structure, which shows the syntactic dependencies (Figure 31), or in table form, where each individual word is accompanied by its Part-of-Speech label as well as more complex set of grammatical features, such as case, person, gender, and tense (Figure 32).

Figure 31: The tree structure of a complex English raising construction. Apart from visualising the sentential structure, the tree structure also shows the parts of speech and syntactic features of the constituents.

[36] http://lindat.mff.cuni.cz/services/udpipe/

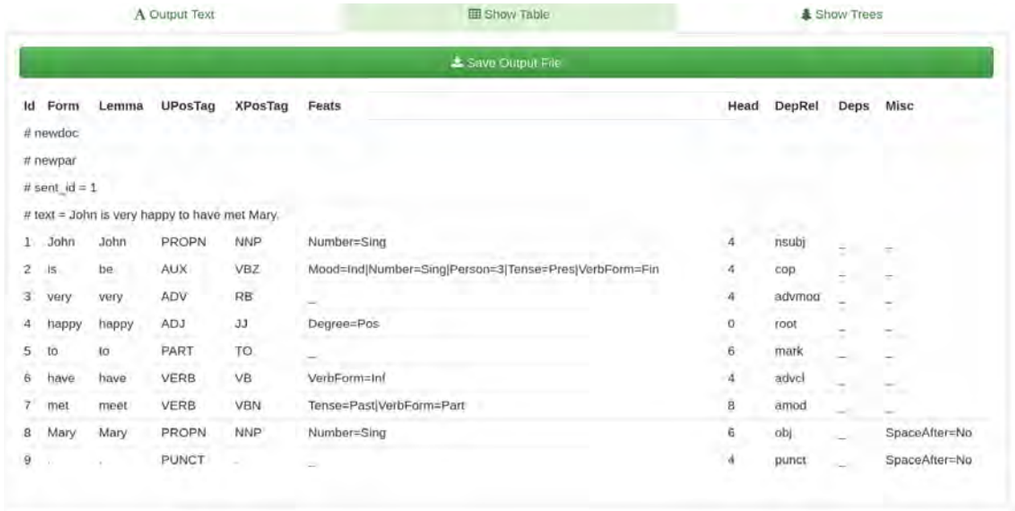


Figure 32: UDPipe shows the grammatical features of the sentence "John is very happy to have met Mary" in table form. Note that it can detect very complex features, such as the perfect (i.e. past tense) use of the infinitive in the subordinate clause.

The powerful flexibility of UDPipe was demonstrated in the CoNLL 2017 shared task, which was of crucial importance for the development and research of dependency parsing. In the shared tasks, UDPipe was used to process raw text in 40+ languages based on the Universal Dependency models with very high precision, which shows that UDPipe can also be easily adapted to annotate and parse new languages. The CoNLL 2018 is a follow-up of CoNLL 2017 and UDPipe is used as a baseline system.

For more details on UDPipe see Straka and Straková (2017) and Straka et al. (2016):

Straka, M. and Straková, J. (2017). Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with Udpipe. In Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Vancouver, Canada, August 2017. http://ufal.mff.cuni.cz/~straka/papers/2017-conll_udpipe.pdf

Straka, M., Hajič, J., and Straková, J. (2016). UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia, May 2016. http://ufal.mff.cuni.cz/~straka/papers/2016-lrec_udpipe.pdf

Universal Dependencies (UD)

Written by Barbora Hladka, edited by Darja Fišer and Jakob Lenardič

Universal Dependencies (UD)[37] is an open collaboration project in the field of Natural Language Processing (NLP). Its motivation comes from multi- and cross-lingual research, and its goal is to develop a universal approach to grammatical annotation, applicable to as many languages as possible. UD is administered by an international team under supervision of Joakim Nivre. The UD project has been up and running since the spring of 2014.

UD provides a universal inventory of part-of-speech categories and syntactic relations for consistent cross-linguistic annotation, as well as several existing treebanks that are richly annotated with the grammatical features. The following picture shows a UD tree structure for the sentence “Mary loves John”. Three part-of-speech categories – PROPN (proper name), VERB (verb), and PUNCT (punctuation) – and four syntactic relations – root (predicate), nsubj (nominal subject), obj (object), and punct (punctuation) – occur in the tree.

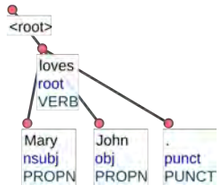


Figure 33: A syntactic tree for "Mary loves John".

UD is also accompanied by detailed guidelines for carrying out the annotation, with examples from numerous languages. The following figure illustrates the complex criteria UD uses to recognise nominal modifiers, which often also take into account complex grammatical interdependencies from formal grammar, such as case assignment/checking.

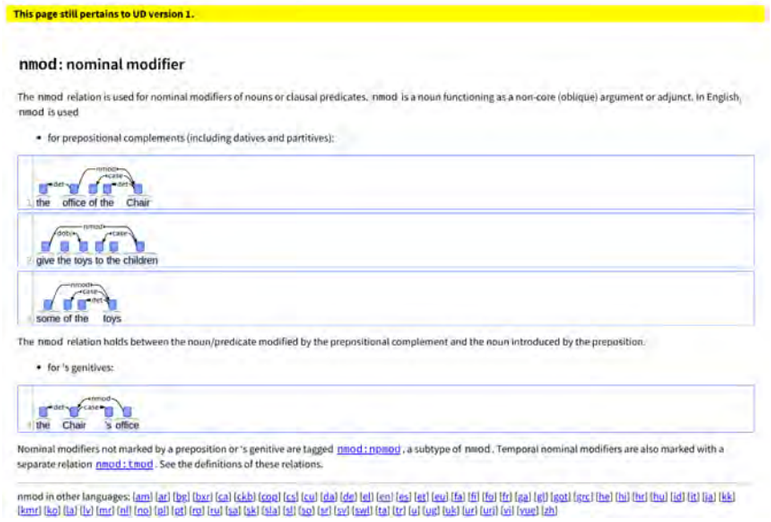


Figure 34: The definition of a nominal modifier in UD.

[37] http://universaldependencies.org/

To search the UD treebanks, researchers can use the online PML-TQ (PML Tree Query) service and UDPipe (presented on page 66), which is an automatic UD annotation pipeline that uses models trained for nearly all the treebanks, so it offers an easy access point to the Universal Dependencies. A number of graphical user interfaces for manual UD annotation are also available. One of them is TrEd, which is a fully customisable and programmable editor and viewer of tree structures developed at the Institute of Formal and Applied Linguistics. The editor, which offers an extension for UD annotation illustrated in the following picture, has been successfully used to annotate thousands of sentences in the Prague Dependency Treebanks.

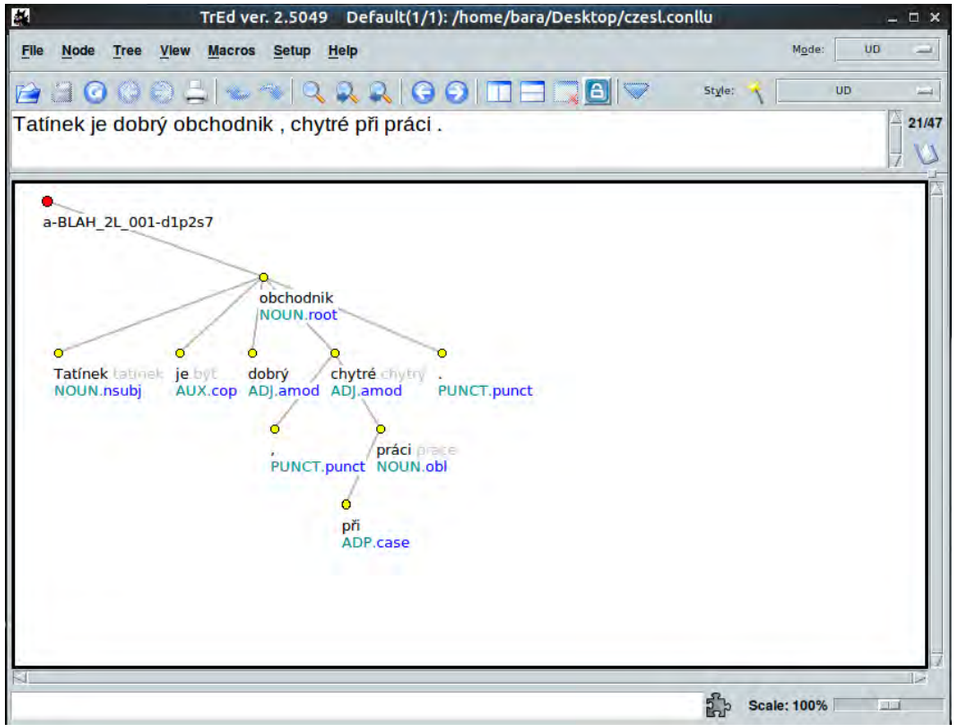


Figure 35: Using the editor TrEd to parse a Czech sentence.

A new version of the UD treebanks is released every six months. The latest version (2.1) came out at the end of 2017 and consists of an impressive number of treebanks, 102, for an equally impressive number of languages, 60. This version offers a ten times greater number of treebanks for six times more languages than the very first release in 2014, which shows how the inclusion of new language data is growing exponentially. All the versions are downloadable from the LINDAT/CLARIN repository.

After a period of rapid growth in 2014–2017, LINDAT organised a series of events dedicated to training and conducting parsing experiments with UD treebanks, as well as discussions of UD-related topics. Among these was a tutorial on UD at the EACL 2017 conference in Valencia in Spain, the first workshop on UD in Gothenburg in Sweden in May 2017, and the CoNLL 2017 and 2018 Shared Tasks, in which the UD treebanks were successfully used as models for the development of advanced dependency parsers.

DARIAH-CZ Workshop on Digital Humanities 2018

Written by Barbora Hladka, edited by Darja Fišer and Jakob Lenardič

The members and partners of the Czech CLARIN consortium recently submitted a proposal to establish DARIAH-CZ, a Czech node of the DARIAH European researcher infrastructure for arts and humanities. In light of the proposal, a one-day international workshop titled DARIAH-CZ Workshop on Digital Humanities 2018[38] was held in Prague on 24 April 2018 at the Academy of Sciences of the Czech Republic in order to introduce the project and generally promote computational approaches within humanities and social sciences, both in the Czech Republic and internationally. During the workshop, sixteen lectures were given by prominent computational and digital humanities researchers working at leading Czech and European research institutions. The workshop was well attended. There were around 70 participants, most of whom were researchers from various Czech institutions while some also came from Slovakia, Poland, Hungary, and Germany.

The workshop began with the introduction of the European projects and, institutes related to the DARIAH-CZ project both structurally (DARIAH, DARIAH-PL, DESIR) and thematically (EADH, Austrian Centre for Digital Humanities). In the afternoon, the Czech projects that are planned to be integrated in DARIAH-CZ were presented. Perhaps most prominently, Pavel Straňák gave a comprehensive presentation of the LINDAT/CLARIN repository and Silvie Cinková introduced the recently established Czech Association for Digital Humanities, which will be a partner in the project. The afternoon session was concluded with a lecture on EHRI, which is a portal dedicated to the presentation and interpretation of Holocaust-related archival documents on the basis of digital tools.

In conclusion of the workshop, the following projects were presented to showcase the successful application of computational methods within the humanities and social sciences:

- the GEHIR project, which is an interdisciplinary research initiative that applies computational methods to the historiography of ancient Graeco-Roman religions;
- the Archaeological Information System of the Czech Republic, which is a tool used to integrate digital resources on Czech archaeology;
- the READ project and its main system, Transkribus, for transcribing and searching historical text collections; and
- Electronic Enlightenment, which is a wide-ranging online collection of edited correspondence from the early 17th to the mid-19th centuries.

The workshop successfully raised awareness of the proposed DARIAH-CZ and its related projects in the context of digital humanities. In addition, it strengthened the ties among the members of the Czech CLARIN consortium, its related partners and other national and international institutions, opening new research avenues for further collaboration.

Figure 36: Boosting research in Digital Humanities with Research Infrastructures.

[38] https://www.lib.cas.cz/en/dariah-cz-workshop-2018/

Radim Hladík

Radim Hladík is a postdoctoral researcher at the Institute of Philosophy at Academy of Sciences of the Czech Republic in Prague and at the National Institute of Informatics in Japan. The following interview took place via Skype on 16 May 2018 and was conducted and transcribed by Jakob Lenardič, edited by Darja Fišer.

1. Please describe your academic background and your current position(s).

I received my PhD in sociology at the Faculty of Social Sciences in Prague. Currently, I'm a JSPS post-doctoral fellow at the National Institute of Informatics in Japan, where I am representing my sending organisation, the Institute of Philosophy of the Czech Academy of Sciences. Many of my colleagues in Japan are computer scientists, so this is a wonderful opportunity for me to improve my coding skills and be inspired about how to combine computational methodologies with social science research topics.

2. How did you get involved with Czech CLARIN consortium? Could you describe your collaboration with the consortium?

Two years ago, as a delegate of the Institute of Philosophy, I was one of the coordinators in a fairly large digital humanities project by the Library of the Czech Academy of Sciences. The proposal was ambitious, since we wanted to strike up a collaboration between many Czech institutes relevant for digital humanities, such as libraries, universities, and various institutes for linguistics and social sciences. Sadly, the project never left the planning stages, but it nevertheless brought together proponents of digital humanities, including me and the colleagues from Czech CLARIN. I was very inspired by their work and soon started learning how to code and apply computational approaches to my own research, which is otherwise rooted in sociology and media studies. Since then, I've been using tools and resources that Czech CLARIN provides and am in contact with their experts like Pavel Straňák, with whom I discuss my work and who has often helped me with technical issues.

As for concrete collaborations, we've recently established the Czech Association of Digital Humanities, for which I currently serve as the Chair. Several people from Czech CLARIN are very active in this association, like Eva Hajičová and Silvie Cinková. We've also submitted a project under the Czech DARIAH node last year with Czech CLARIN as the principal investigator. Its goal is to conduct an extensive corpus-based analysis of modern Czech texts from various domains (e.g., 20th century philosophy). I'll be involved as a representative of the Institute of Philosophy, which aims to contribute its historical and philosophical corpora and texts collections to Czech CLARIN. I believe that such a collaboration is of great importance for both sides. On the one hand, Czech CLARIN will give us an invaluable platform for the curation and sustainability of our resources, while on the other they'll be able to expand the applicability of their tools to new domains and across historical language variations based on our resources.

3. Which are the tools and resources provided by Czech CLARIN that you use in your research? Could you discuss how you use them in your own work?

If you work with texts in a language that is as morphologically complex as Czech, lemmatisation and morphosyntactic annotation of texts is needed even for the simplest analyses. In this sense, the tools that Czech CLARIN provides are essential for my current work.

I'd like to point out MorphoDiTa,[39] which is a tool for tokenisation, lemmatisation and morphological analysis. What I especially appreciate about MorphoDiTa is its flexibility, in that you don't need to install it as a stand-alone program on your computer, but you can use it as an API service which you easily integrate in your own code. This way, I don't need to worry about having additional components installed and their dependencies. I often come across tools that require a complicated installation processes, which dissuades me from using them.

What I also appreciate is that the Czech CLARIN repository keeps track of all the versions of a resource you upload. I believe this takes a lot of pressure off the whole publishing process, since I know that I can always publish a newer version of a specific dataset in case I do some additional work on it, making me more confident in releasing a dataset sooner, since the repository also welcomes non-final versions, which are then automatically linked to newer ones.

4. Your research scope is very broad; among others, you apply a digital humanist approach to the study of scientific writing in social sciences. Could you briefly describe how you conduct your research in connection with this topic?

In my postgraduate work I have been interested in how historical events are represented through mediated communication, and why only certain statements about the past are regarded as truthful representations. Currently, I've been tackling similar questions in connection with scientific writing, where I'm mostly interested in how scientists establish the validity of their claims. However, most sociological research on this topic has been purely qualitative or conducted on a handful of sampled texts. I find such an approach limited, since you can't really make general claims about whole decades of scientific writing in a particular domain based on a few dozen papers.

Consequently, I soon started wondering what a proper digital approach would reveal about this topic, and I began working on creating a corpus of Czech sociological articles from scratch. Currently, my corpus is fairly small – after the clean-up it consists of around 500 articles, but will hopefully grow with time.

5. Have there been any significant results yet?

I've obtained some interesting results by combining my corpus with a corpus of literary texts that I downloaded from the repository of Czech CLARIN. I brought the two corpora together by creating a vector space model of the documents consisting of very low-level features – the most frequent verbs that are shared between the corpora. I then applied clustering methods to the combined corpus to see which specific sociological texts have the most in common with the literary texts. As an example, clustering showed that such sociological texts often give voice to their data, by providing quotes of the people who are the subjects of the study in question. But the clusters do not only differ in language use. What I found out is that such texts are also more likely to be written by female authors, and often tend to be cited less than those texts which have little in common with fiction. Both observations turned out to be statistically significant. I plan to release this sociological corpus through the Czech CLARIN repository once it's completed.

[39] http://ufal.mff.cuni.cz/morphodita

6. Why is an infrastructure like Czech CLARIN (or CLARIN ERIC in general) important for the general research community?

I've met quite a few researchers from non-technical disciplines who oppose the use of quantitative methods in what they perceive to be qualitative research questions. I understand their point of view, which I used to share to an extent. But now that I have some experience with using language tools and resources myself, I find that such opposition often isn't really justified, although researchers must be aware of potential limitations and make sure to use the right tools for their purposes. In other words, there are many misconceptions about quantitative research and I believe that Czech CLARIN can help a lot in this regard through its user involvement events. After my personal experience of auditing the CLARIN-PLUS workshop: "Working with Digital Collections of Newspapers",[40] I think that the workshops are especially important because they're a platform where CLARIN experts can show how their tools work and how they do not only answer specific research questions from various disciplines, but also open up many approaches to doing research. An event that directly involves its participants is definitely much more convincing than a dry lecture on digital humanities that does not provide any kind of concrete examples.

Additionally, such events are often the starting points of many fruitful cross-disciplinary collaborations in which social scientists or humanities researchers team up with computer science experts. Due to such collaborations, getting involved in digital humanities does not necessarily mean that you need become an expert programmer yourself; you often only need to get intuitively acquainted with the computational methodologies and learn the basic skills, just enough to find common ground for conversations with the specialists.

7. How do your students and fellow researchers embrace the digital humanist approach? How are digital humanities in general represented in the Czech academic environment?

At the Institute of Philosophy, there is quite a lot of enthusiasm for digital humanities, since the management and many researchers see it as a step forward in scientific research. At universities, it depends a lot on the particular department. For instance, I once attended a course on programming in R that was given by Silvie Cinková from Czech CLARIN. Many students who also attended this course were from various humanities disciplines. They were very enthusiastic about learning how to programme and potentially applying programming skills to research questions within their own domains. Consequently, I think there are more students who are interested in such quantitative approaches than the management of humanities departments might realise. The problem, of course, is that the faculty at such departments doesn't usually have the required skills to teach a digital humanities course, so they often invite external teachers from the industry to teach a course or two. However, fully embracing the digital humanities would probably require a revamping of the curriculum with a greater number of digital courses tailored to topics that are directly relevant to humanities research interests.

8. What is your vision for the future of Czech CLARIN?

What I really appreciate about Czech CLARIN is that they have managed to develop tools for Czech that can easily compete with state-of-the-art language technologies developed for larger languages, like English. At LREC 2018, it was obvious to me that language technologies are rapidly becoming more and more advanced worldwide. I'm confident that Czech CLARIN will continue to keep up and make sure that their tools are always in touch with the state-of-the-art. If there's one thing that I'd like to see improved, it's the documentation of the tools and resources, which could be made more user-friendly and contain more examples of use because learning a new tool can be very intimidating.

[40] https://www.clarin.eu/event/2016/clarin-plus-workshop-working-digital-collections-newspapers

Greece

Written by Darja Fišer, Maria Gavriilidou and Jakob Lenardič

The Greek network clarin:el[41] has been a member of CLARIN ERIC since February 2015. It was founded by three Greek research institutions:

- the Athena Research and Innovation Centre;
- the National Centre for Scientific Research Demokritos; and
- the Greek Research and Technology Network (GRNET S.A.).

It has since expanded to a nation-wide network currently including five universities and two research centres:

- University of Athens,
- Aristotle University of Thessaloniki,
- Ionian University,
- University of the Aegean,
- Panteion University,
- Centre for the Greek Language, and
- National Centre of Social Research.

The Greek consortium is coordinated by Stelios Piperidis, Head of the Department of Natural Language Processing and Language Infrastructures of the Institute for Language and Speech Processing/Athena Research and Innovation Centre.

Clarin:el primarily functions as a secure and stable national research infrastructure, offering a network of dynamic repositories devoted to and enabling the sustainable storage and dissemination of language tools and resources. Researchers can access the tools and resources of the consortium via the clarin:el inventory, which acts as a single access point to a number of local, institutional, repositories that are part of the clarin:el network. The clarin:el inventory provides user-friendly browsing and search functionalities, offering a customisable faceted search interface that allows researchers to narrow down their search queries on the basis of metadata-based features such as resource type, language, thematic domain, temporal or geographic coverage, access terms and conditions, etc.

Clarin:el currently contains around 500 language resources and 35 tools, which can be accessed by registered and non-registered users, in full compliance with the licence terms defined by the resource providers. Many of the language tools in the inventory are offered as web services for processing content in Greek as well as other languages, which means that researchers can use them to process their data directly through the inventory: users can either select resources from the clarin:el inventory to process, or they can upload their own data for processing. The outcome of the processing constitutes a new resource which can directly be added to the inventory,

[41] http://www.clarin.gr/en

accompanied by automatically created metadata. Statistics related to the use of the resources (such as number of views and downloads) as well as dynamic recommendations of related resources and services (such as similar resources viewed by other users) are available to all users. Organisations that are members of the clarin:el network have the ability to set up their own repository within the infrastructure, which can then be accessed through the central inventory. Individuals who join the clarin:el network may store their resources at a dedicated repository, the so-called Hosted Resources Repository, which is also available for the storage of resources provided by organisations which do not wish to maintain their own repository.

The Clarin:el Team

The Greek consortium also actively promotes user involvement. Recently, on 27 June 2018, the Greek consortium organised an event intended to deepen the dialogue with digital humanities and social sciences researchers, better understand their requirements and familiarise them with the clarin:el infrastructure. On the one hand, the event featured an interactive session where the researchers had the opportunity to present their own research questions and experiences with using language technologies to clarin:el experts, while on the other, a hands-on session was organised, where the researchers were able to familiarise themselves with the clarin:el inventory and the use of its resources and tools. You can read more about the event on page 79.

Kos, Greece | photo by Mico59 | Pixabay

GrNE-Tagger

Written by Maria Gavrilidou, edited by Darja Fišer and Jakob Lenardič

The GrNE-Tagger[42] is a tool available through clarin:el that automatically recognises proper names (Named Entities) in Greek texts and classifies them into one of the following five category types:

- PERSON: person names, family names;
- LOCATION: political or geographical names such as continents, countries, cities, etc.;
- ORGANISATION: names of entities such as companies, institutions, organisations, etc.;
- FACILITY: names of buildings and other human-created structures, such as streets, bridges, etc.;
- GPE (Geo-political entity): entities whose names coincide with a location name, but whose semantic content actually refers to its government or administration.

The GrNE tagger is not a single tool, but rather a pre-defined pipeline of tools seamlessly integrated, in the sense that the output of one tool constitutes the input for the next:

Tokenisation > Sentence Segmentation > Part-of-Speech Tagging > Lemmatisation > Chunking > Named Entity Recognition

The annotation processes before Named Entity Recognition constitute the pre-processing of the text. After the pre-processing stage is completed, the Named Entity Recognition algorithm is applied to the text in two stages: it first uses linguistic rules to identify a set of candidate NEs and subsequently checks them against manually created wordlists of existing proper names. If a proper name in the pre-processed text is not identified in this manner, the tool tags it as UNKNOWN.

To consolidate a candidate NE or a proper name labelled as UNKNOWN, and to finally place it into the correct category, GrNE-Tagger applies another round of linguistic rules that search for specific keywords in the context of the ambiguous expression. The keywords used for such disambiguation are, for example, professional titles, words denoting nationality or kinship terms such as father of, sister of etc. (in the case of PERSON); prefixes or suffixes denoting company types, such as Corp., Ltd. etc. (for ORGANISATION); words such as street, bridge etc. (for LOCATION) and so on. Based on shallow syntactic parsing, the system also disambiguates between LOCATION and GPE (geo-political entity).

GrNE-tagger has been integrated in the clarin:el infrastructure as a web service, which means that the users do not need to install the tool locally; they simply select a resource from the clarin:el inventory (or upload their own resource) and they process it. After the completion of the processing, the users receive an email with a link to the results of the processing. Furthermore, the tool has already been successfully applied to annotate several resources; for instance, one such resource enriched with GrNE-tagger is a corpus of interviews conducted with female entrepreneurs in Athens.

GrNE-tagger has been developed and is maintained by the Institute for Language and Speech Processing / Athena RC, and is available under a licence that permits Academic – Non Commercial Use.

Figure 37: The output of GrNE-tagger (using GATE as a visualisation tool), in which different NEs are marked with different colours.

[42] http://hdl.grnet.gr/11500/ATHENA-0000-0000-23F2-7

The Hellenic Parliament Sittings and Hellenic Parliamentary Corpus H-ParCo

Written by Katerina T. Frantzi and edited by Maria Gavriilidou, Darja Fišer and Jakob Lenardič

The corpus Hellenic Parliament Sittings,[43] developed by the Laboratory of Informatics, Department of Mediterranean Studies of the Aegean University, includes minutes of meetings of the Greek Parliament and speeches of Parliament members, spanning the years 2011–2015. The resource has a total size of approximately 28.7 million words.
The corpus forms part of the dynamic Hellenic Parliamentary Corpus, H-ParCo, whose development was actually inspired by the participation of the University in the clarin:el network. The latest version of H-ParCo consists of language materials from all Plenary Sessions Minutes published by the Hellenic Parliament from 3 July 1989 to 31 April 2018; so in total, 29 years of Plenary Sessions Minutes. This version will soon be available through clarin:el, while the current published version, namely the Hellenic Parliament Sittings corpus, can already be downloaded under the CC-BY-NC licence.

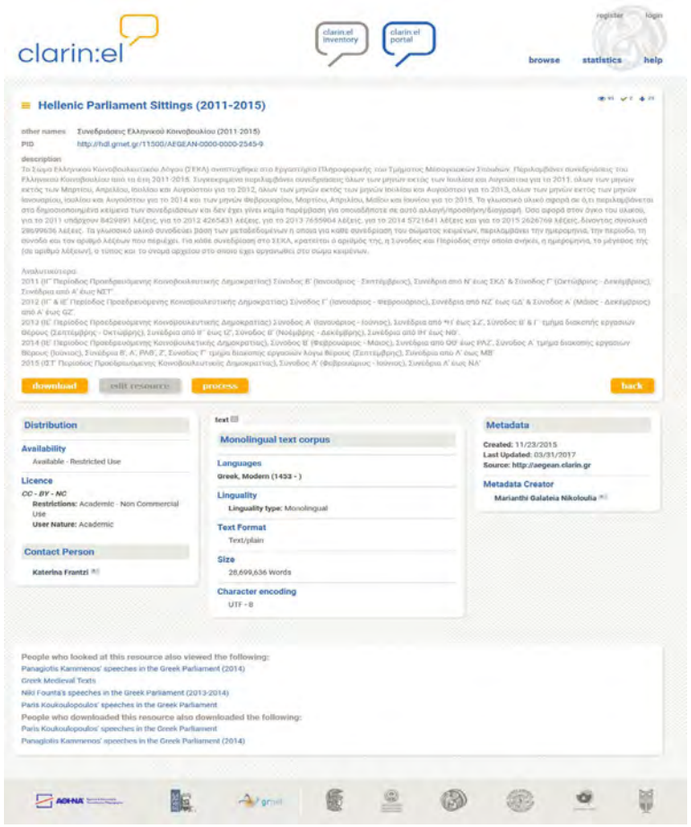


Figure 38: Greek parliamentary corpus in the clarin:el repository.

[43] http://hdl.grnet.gr/11500/AEGEAN-0000-0000-2545-9

The collection process has not been an easy task: the Hellenic Parliament publishes data in three formats: as .pdf files, as .doc files, and as .txt files. The data have been retrieved manually and classified according to the year and month they pertain to. All .pdf and .doc files have been converted into .txt files, so that they can be processed by existing clarin:el tools. The original files have also been kept and organised so that a one-to-one correlation with the corresponding .txt files is maintained. The corpus also contains rich metadata that specify the date, the parliamentary term and session, the meeting, the original file name, the corresponding .txt file name and the size in terms of number of words for each file.

The actual language material contained in the resource is exactly what is included in the publicised texts of the meetings; no manual or automatic interventions have taken place to alter the recorded language (for instance, to correct errors or "sanitise" the language used).

Given that the development of H-ParCo is an ongoing process, future work involves:

- The continuous addition of minutes of recent Parliament Plenary Sessions (from 31 April 2018 onwards). These files are expected to be of the same formats as the previous ones, so the retrieval and processing procedures are also expected to be the same.
- The addition of minutes of older Parliament Plenary Sessions (before 3 July 1989). These are mostly image files (scanned images); this is expected to hamper the data retrieval and processing procedures, which is expected to be a lot more time-consuming, as the task of their conversion to .txt files is not a straightforward process.
- The development of a similar corpus consisting of the Parliament Plenary Sessions Minutes of the Democracy of Cyprus. In this case, the files are in .pdf form. Therefore, the retrieval procedure is expected to be the same as that of H-ParCo.

Generally, H-ParCo is aimed at researchers of various domains and disciplines, such as Linguistics, Political Discourse Analysis, Critical Discourse Analysis, Digital Humanities, Communicational Techniques, Political Sciences, Sociology, Gender Studies and more. It has been already successfully used for Political/Critical Discourse Analysis purposes (Georgalidou et al. 2017 and 2018 [in Greek]).

References:

Γεωργαλίδου, Μ., Φραντζή, Κ. Τ., and Γιακουμάκης, Γ. (2018). Κοινοβουλευτικός λόγος, ευγένεια και επιθετικότητα στο ελληνικό κοινοβούλιο. Book of Abstracts of the 39th Annual Meeting of the Department of Linguistics, School of Philology, Aristotle University of Thessaloniki, Thessaloniki 19-21 April 2018.

Georgalidou, M., Frantzi, K., and Giakoumakis, G. (2017) Addressing adversaries in the Greek Parliament: a corpus-based approach. Book of Abstracts of the 13th International Conference on Greek Linguistics, Westminster 7-9 September 2017.

Language Data and Technologies in Social and Political Sciences

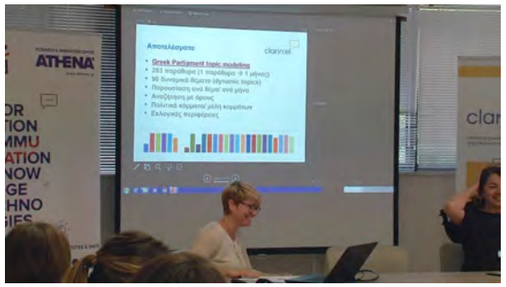
Written by Maria Gavriilidou, edited by Darja Fišer and Jakob Lenardič

On 27 June 2018, the Greek Infrastructure for language resources, technologies and services clarin:el organised an event on Language Data and Technologies in Social and Political Sciences, which took place at the Institute for Language and Speech Processing (ILSP/"Athena" R.C.).

ILSP, which serves as the coordinator of clarin:el, invited prominent Social and Political Scientists from the National Centre for Social Research (EKKE) and the Department of Political Science and History of the Panteion University (which are the two new member institutions of the Greek CLARIN network) for a focus-group workshop.

Aiming at the mutual acquaintance of the two scientific areas (social and political sciences on the one hand, and language technology on the other), the 15 researchers of EKKE and Panteion University were given the opportunity to present their research questions, the way they work, the current methodologies they follow, and the tools and data processing techniques they are familiar with. This took place in a lively interactive session, which has been recorded on video.

In turn, ILSP researchers presented specific language technology applications developed in the framework of multidisciplinary research projects, and showed how they can be used to tackle qualitative research questions. These presentations reinforced the interconnection between the two scientific areas, highlighting the key role of language technology in facilitating research in the domain of social and political sciences.



The second half of the event familiarised the social and political scientists with the clarin:el infrastructure. A detailed presentation of clarin:el (which focused on the portal, the repositories, the inventory of resources, and the tools and web services) served both as an introduction for the two new member institutions and as a glimpse of their tasks as nodes of the network. Additionally, the researchers were shown how to set up their repositories, prepare relevant documentation and upload their resources.

The presentation was followed by a hands-on training session, where the researchers were given a "guided tour" of all the features of the infrastructure and, through guided exercises, were able to familiarise themselves with the clarin:el inventory and the use of its resources and tools.

The discussion session that closed the event revealed some interesting points as regards the use of language resource infrastructures:

- The most crucial factor for the uptake of the clarin:el infrastructure (as expressed by the participants) was access to tools and web services, followed by access to other resources. Lower in their motivation for using the clarin:el infrastructure was the incentive to share their own resources, or to store resources in the repository.
- The role of language resource infrastructures in terms of certain legal issues (clearance of IPR, copyright issues, distribution issues, standardisation of licensing procedures, etc.) was considered to be very significant as regards the protection of language resource contributors and consumers from illegal use of their data.
- Finally, the promotion of a "sharing culture" and of open language data and tools was also highlighted as a crucial activity for language resource infrastructures.

Vassiliki Georgiadou

Vassiliki Georgiadou is Associate Professor of Political Science at the Panteion University of Social and Political Sciences. The following interview took place via e-mail and was conducted by Maria Gavriilidou, edited by Darja Fišer and Jakob Lenardič.

1. Could you tell us a little bit about yourself, your background and your current work?

I am Associate Professor of Political Science at the Department of Political Science and History of Panteion University of Social and Political Sciences. Our university is one of the oldest Greek universities, and the first school of political sciences in Greece. I studied political science in Athens (Panteion) and in Münster, Germany (Institute of Political Science) and I hold a PhD from the Faculty of Philosophy of the Westphalian Wilhelms University of Münster. I am a member of the Steering Committee of the Centre for Political Research of Panteion University, which operates as a laboratory of political sociology and comparative politics. Since April 2016, I have been a member of the National Council against Racism and Intolerance in Greece, which has already started planning a national strategy to combat discrimination and racism.

My research interests focus on political behaviour, far right parties, populism, radicalism and political extremism. I was principal investigator of the XENO@GR research project,[44] which examines xenophobia in Greece during the economic crisis, based on a computational social sciences approach; currently, I am co-investigator of a research project that examines the different expressions of violence in Greece from 2008 onwards (London School of Economics (LSE)-Hellenic Observatory Grants).

2. How did you hear about CLARIN, and how did you get involved?

I knew about this European research network, since CLARIN is one of the most relevant infrastructures for researchers of the humanities and the social sciences working with language related material. I am also involved in So.Da.Net network, which is another research infrastructure that brings together social science data archives across Europe, and I was aware of the facilities that research infrastructures provide to the scientific community in order to conduct top-level research in their respective fields. I became involved in clarin:el during our project XENO@GR and the collaboration with the research team of ILSP/ATHENA, which coordinates the clarin:el consortium and was also our research partner in the XENO@GR project.

3. Could you describe the XENO@GR project in more detail?

The basic aim of this research effort was to examine the phenomenon of xenophobia in Greece through a large-scale multi-source study based on the use of advanced computational social science approaches. There is a common perception that xenophobia is a deep-rooted social phenomenon that escalates under circumstances of severe economic crisis. In line with this perception, xenophobia should have increased in Greece after the outburst of the economic crisis in 2009. Drawing on a vast amount of data from a rich variety of sources and

[44] http://xenophobia.ilsp.gr/?lang=el

exploiting a wealth of research instruments, we tried to test the validity of this , addressing the following research goals:

a. to study the historical evolution of the phenomenon of xenophobia in Greece from the 1990s onwards;
b. to examine whether the recent economic crisis has raised the xenophobic sentiments and behaviour of Greeks against any kind of "others"; and
c. to decompose the effect of the economic crisis on the behaviour of the Greek people against the "others", in order to examine the expressions of continuity as well as the possibility of change, with reference to xenophobia as a social phenomenon deeply rooted in the perceptions and consciousness of Greeks.

Figure 39: The workflow for creating the event database in the project.

To achieve the above goals, we created a large event database capturing events which were related to the phenomenon under study and happened in the timespan of the last twenty years. All entities (people, organisations and locations) involved in these events, as well as the sentiments and emotions expressed, were captured and coded in a knowledge network facilitating the exploration and further analysis of such social interaction in Greek society.

4. On the basis of your work within this project, could you explain how social sciences and humanities researchers collaborate with experts from a research infrastructure offering language technologies (in your case, the Greek CLARIN consortium)?

In our first contact with language technologies we were impressed by the potentialities we had in our hands. We understood that every text, sound or video in the world is a new data source. We were confronted with different and rich data sources and we selected those that could help us answer our research questions. Our next step was to decide how to analyse them. Language technology experts explained all the possibilities and we decided to use event analysis for the newspaper data we had and sentiment analysis for the social media data. We collaborated on building a codebook for the events' description and a similar first-step categorisation of Twitter data that resulted in different sentiment categories of verbal aggressiveness. This procedure was a step by step collaboration of both teams (social scientists and language technology experts) at both the conceptual and the analytical levels. The findings were coded and then stored in a knowledge network so as to promote the examination and analysis of the focal social interactions in the Greek society.

Figure 40: The main targets of verbal aggressiveness in Greece in terms of national/ ethnic background, based on the analysis of Greek newspapers during the XENO@GR project.

5. How has CLARIN influenced your way of working? Would you like to single out any tools and resources provided by the Greek consortium that you used in your work?

CLARIN (and every other repository infrastructure) must be considered as an innovative tool of communication between researchers and a discussion platform for the academic community. With CLARIN, every researcher has the opportunity to use language resources and computational tools for analysing empirical data, while being affiliated with the ethical and technical rules and standards of data management and data protection. In addition, data collection and analysis must comply with methodological standards that will facilitate their replication or reproduction and corroboration. Clarin:el offers a number of data analysis tools, like sentiment analysis and event analysis. In particular, in the XENO@GR project we used Natural Language Processing (NLP) tools (offered by clarin:el as web services) for tokenisation, sentence splitting, part-of-speech tagging, and lemmatisation, before embarking on the more semantic, domain-specific event and sentiment analysis of the XENO@GR data. A detailed description of the tools used can be found on the project's website. All of our processed data were uploaded at clarin:el.

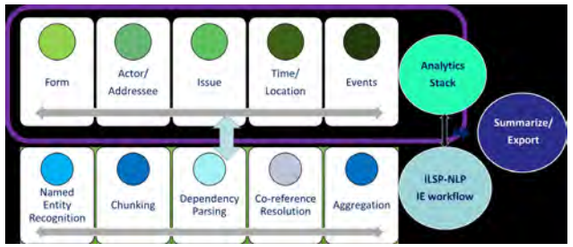


Figure 41: The NLP tasks performed on the XENO@GR data.

6. How easy was it for you to adapt to the changes that language technologies introduced in your research methodology? Do you think that using language technologies opens up new research opportunities for political scientists?

From my point of view as a social scientist, it is quite challenging to collaborate with colleagues from other disciplines. Through interdisciplinarity, I believe that academic research can become more robust and useful to society. However, it is not an easy task to find a common communication code with a discipline that is mostly based on technology. At the beginning we had to adjust to the terminology that the two teams used, in order to choose the appropriate operational definitions for our study; but as soon as this became common ground, we obtained a new perspective for our research and became familiar with the potential that this interdisciplinary collaboration brings. We live in a world with multiple data sources that can be valuable not only for science, but also for the stakeholders. Language technologies give us the opportunity to explore large amounts of data in the minimum amount of time, and thus make more concrete and grounded inferences. I believe that this is a very big step for the social sciences in general and it expands research opportunities as well.

7. Did this experience influence your decision to join the clarin:el network and set up your own LRs repository?

The clarin:el network facilitates access to language data sources that are extremely relevant in our research as social scientists. Having joined the network as a legal entity (Panteion University), we can upload and share our data through the university data repository, thus enabling other researchers to work with our resources, and we can have access to the resources of others. This provides synergies among researchers and improves data availability, accessibility and sustainability.

8. Could you share your experience with the recent clarin:el event for social sciences and humanities researchers (presented on page 79)? How are such events valuable to your research community?

The purpose of this event was to bring together clarin:el with members of Social Sciences research community in Greece, notably researchers from Panteion University and the National Centre for Social Research that joined clarin:el in 2017 and 2018, respectively. For us, as new members of the infrastructure, it was extremely important to be informed of what our participation in clarin:el could bring, how to get involved and take advantage of the opportunities offered by the infrastructure. The keynote talk (Maria Gavriilidou) and the presentations (Xaris Papageorgiou, Maria Pontiki, Stelios Piperidis) elaborated the clarin:el infrastructure, architecture, and the contents, as well as technical and legal issues regarding the use and the sharing of resources that are available in clarin:el. For us, as social scientists, it was particularly useful that the presentations were followed by an intensive hands-on session, where we were trained on the use of the clarin:el infrastructure.

9. How would you envisage future collaboration of your university with CLARIN?

I hope that more members of our research community at Panteion University will get involved in clarin:el, and that our students, researchers and colleagues will take advantage of the opportunities offered by the infrastructure.

Lithuania

Written by Darja Fišer and Jakob Lenardič

The Lithuanian consortium CLARIN-LT[45] has been a full member of CLARIN ERIC since 2015. The consortium consists of three partner universities – Vytautas Magnus University, Kaunas Technology University and Vilnius University – and is led by Assoc. Prof. Andrius Utka as the head of the consortium, Dr. Jurgita Vaičenonienė as the National Coordinator.

CLARIN-LT offers a C-certified repository, which provides a host of specialised and well-annotated language resources suitable for research within digital humanities disciplines. The consortium offers dedicated online access to the Corpus of the Contemporary Lithuanian Language. Additionally, a researcher is given access to some important Lithuanian language resources, such as ALKSNIS – the largest Lithuanian Treebank (presented on page 87), LITIS – a corpus of user-generated comments, and the Lithuanian Parliament Corpus for Authorship Attribution, which is especially tailored to authorship attribution tasks and has been successfully used in a variety of interdisciplinary research endeavours.

[45] http://clarin-lt.lt/?lang=en

To promote active user involvement the consortium has set up two help desks, whose experts can be contacted via e-mail or telephone:

- the Lithuanian Language Technology Helpdesk at Vytautas Magnus University provides information and consultations on corpus analysis, terminology extraction, the use of tools for part-of-speech tagging, syntactic parsing, and similar uses of language technology; and

- the Semantic and Conceptual Modelling Helpdesk at Kaunas University of Technology provides information and help on approaches related to database and information system engineering, ontology development methods, the implementation of semantic search processes and other relevant issues.

The consortium has also organised a series of user involvement events. In 2016, CLARIN-LT experts organised a seminar where they presented successful use cases on how the language resources and computational tools developed at the consortium can be applied within digital humanities and social sciences research. In 2017, the consortium organised their biggest event yet – a two-day CLARIN-PLUS workshop dedicated to the creation and use of social media resources, which was attended by some of the foremost computational and digital humanities experts on computer-mediated communication.

CLARIN-LT team (L-R): Andrius Utka, Agnė Bielinskienė, Jurgita Vaičenonienė, Erika Rimkutė, Rūta Petrauskaitė, Jolanta Kovalevskaitė, Loic Boizou.

Colloc

Written by Tomas Krilavičius, Jolanta Kovalevskaitė, Jakob Lenardič and Darja Fišer

Colloc[46] is an experimental tool aimed at the automatic identification of Multiword Expressions (MWEs). MWEs (or multiword units) are fixed word combinations that can be different in their nature: some of them are semantically non-compositional, i.e. their global meaning is different from the sum of their individual parts (idioms or phraseologisms), whereas others are transparent, but have usage-based co-occurrence restrictions (collocations). The tool, developed by a team of researchers working at the CLARIN-LT centre at Vytautas Magnus University and the Baltic Institute of Advanced Technology, covers the whole process of MWE identification, and can also be used for the development of new methods of MWE identification.

The experimental prototype includes all the steps of linguistic analysis, namely:

1. Text pre-processing
2. PoS tagging
3. N-gram generation and calculation of their statistical properties
4. Calculation of Lexical Association Measures (LAMs)
5. Word embedding generation
6. MWE identification using:
   - Filtering (gazetteers, dictionaries)
   - Application of LAMs
   - Application of machine Learning
   - Hybrid methods

The basic user version of Colloc, which is cloud-based and will be available soon, currently supports only Lithuanian and was trained on a 70 million-word corpus, collected from the Lithuanian news portal delfi.lt. The tool has been statistically trained on GloVe Word Vectors and employs artificial neural networks. It is designed to be user friendly, so researchers will only have to upload the text file whose multi-word expressions they want to have analysed (as in Figure 42), and the tool will simply return the annotated document.

Figure 42: The Colloc user interface.

It is important to have a tool that can extract MWE candidates from particular text(s), since this opens more possibilities not only for terminological, lexicographic and NLP perspectives on language analysis, but also for different areas in applied linguistics, like language learning. The tool will help linguists perform deeper analyses of language, investigate its compositionality, idiomaticity and dynamics. Language technology specialists will be able to use Colloc to improve automatic text analysis, machine translation, information extraction tools, and make chatbots more human.

The tool development is funded by Lithuanian Research Council, Pastovu project.

[46] http://mwe.lt/en_US/

ALKSNIS, the Lithuanian Dependency Treebank

Written by Agnė Bielinskienė, Jakob Lenardič and Darja Fišer

ALKSNIS is a syntactically annotated corpus of Lithuanian, and serves as a gold standard for the syntactic analysis of the language. ALKSNIS currently consists of 2,355 syntactically annotated sentences in the PML (Prague Mark-up Language) format. The format allows researchers to visualise and edit the syntactic trees with the editor TrED (see page 69).

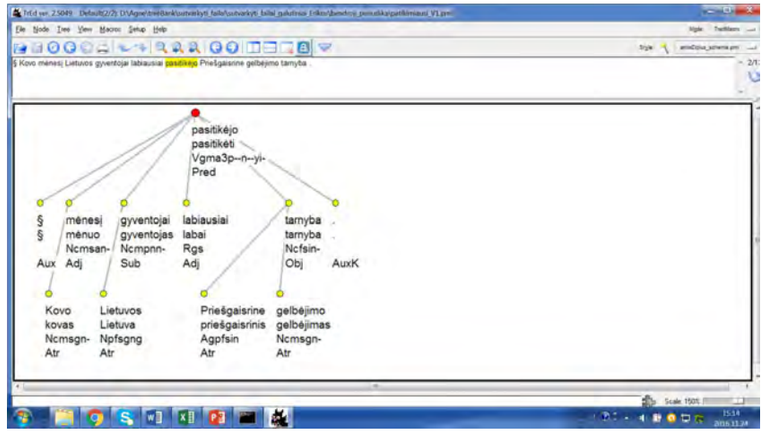


Figure 43: Using TrED to show the syntactic structure of a sentence from the ALKSNIS corpus.

Figure 43 shows the syntactic tree structure of the Lithuanian sentence “Kovo mėnesį Lietuvos gyventojai labiausiai pasitikėjo Priešgaisrine gelbėjimo tarnyba” (“In March, Lithuanian residents mostly trusted the Fire and Rescue Service”), as presented by TrEd. Each terminal node corresponds to a word, a punctuation mark or other text element (symbol, digit, etc.) within a sentence, while the links show the syntactic dependencies. The prepared list of abbreviations for syntactic labels and the presentation of the syntactic relations and dependences were based on the experience of Czech researchers (Hajič et al. 1999). The editor presents the following information for each node:

1. the form used in the sentence (e.g., “gyventojai”, “residents“ in the given example);
2. the corresponding lemma (e.g., “gyventojas”, “resident”, which is the singular form of the plural “gyventojai”),
3. the morphology tag (e.g., “gyventojai” “residents” has the tag Ncmpnn-, which stands for Noun, common, masculine, plural, nominative, non-reflexive, - indistinctive), and
4. the syntactic function (e.g., “gyventojai”, “residents” is the grammatical subject in the given example).

The corpus can also be searched via the ANNIS interface (Krause and Zeldes, 2016). The interface visualises the syntactic dependencies of a sentence and lists its morphosyntactic features, as shown in Figure 44: “Patalpos jau išnuomotos. Taip pat jau rezervuota pusė ploto kitais metais iškilsiančiame statinyje. Dauguma didmenine” (“The premises have already been leased. Also, half of the area of the building to be finished next year has already been reserved. Mostly wholesale”).

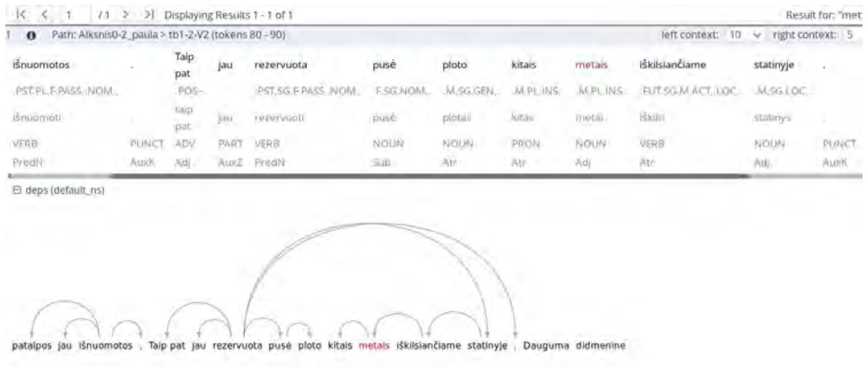


Figure 44: Using ANNIS to parse a sentence in ALKSNIS.

So far, the syntactically annotated corpus has been successfully used by different user groups. For example, at Vytautas Magnus University, students are taught to work with ALKSNIS as part of the curriculum and use corpus data to do various assignments or to develop their theses (for instance, Kristina Brokaitė's Master's thesis used the corpus to analyse grammatical forms of various complex and non-complex predicates in Lithuanian).

The corpus will be enriched with new texts and converted to the Universal Dependency (UD) format. The CoNLL-U format provided by the UD guidelines will serve as the core version of the ALKSNIS treebank. We also plan to annotate the corpus for multiword expressions (also see the description of Colloc, which is a tool for annotating MWEs, on page 86). This will help enhance the usability of the corpus in parsing and in data-driven applications of MWE processing models as well as provide linguists with the information about the syntactic behaviour of Lithuanian MWEs. Finally, a syntactic parser is going to be trained on the Alksnis corpus.

References:

Bielinskienė, A., Boizou, L., Kovalevskaitė, J., and Rimkutė, E. (2016). Lithuanian Dependency Treebank ALKSNIS. In Proceedings of the Seventh International Conference Baltic HLT 2016. Amsterdam: IOS Press, 107–114. http://ebooks.iospress.nl/volumearticle/45523.

Brokaitė, K. (2017). Tarinio raiška gramatinėmis formomis sintaksiškai anotuotame lietuvių kalbos tekstyne ALKSNIS. https://eltalpykla.vdu.lt/1/34649.

Hajič J., Panevová J., Buráňová E., Urešová Z., and Bémová A. Annotations at Analytical Level. (1999). Instructions for Annotators (11.10.1999), UK MFF ÚFAL Praha.

Krause, Th. and Zeldes, A. (2016.) ANNIS3: A New Architecture for Generic Corpus Query and Visualization. In Digital Scholarship in the Humanities 2016 (31). http://dsh.oxfordjournals.org/content/31/1/118.

The Annual CLARIN-LT Seminars

Written by Jurgita Vaičenonienė and Darja Fišer

Since the establishment of the CLARIN-LT centre in 2015, we have been organising different types of events to disseminate knowledge about language resources and language analysis tools deposited in the repository of the national consortium. We help lecturers, teachers and students of the humanities and social sciences to use language resources efficiently in their work and research, contribute their data to our repository, or get involved in various CLARIN related activities. Taking into consideration the needs of different audiences, we offer both recurring and single events with the focus on Lithuanian language resources.

An especially successful initiative which addresses researcher and lecturer communities is our annual seminar series[47] organised at the end of each year. The aims of the seminars are to introduce and give an update about the activities of CLARIN-LT and CLARIN ERIC in general; to present the language resources stored or soon to be added to our repository; and show their possible applications. Most importantly, we also use the opportunity to find out the expectations of the audience related to language resource creation and use.

Every year, we attract about 20 researchers from different Lithuanian universities from different disciplines, such as sociology, psychology, economics, politics, linguistics, journalism, cultural studies, informatics, and history.

Part of the seminar is devoted to the presentations of participants where they introduce their research in digital humanities.

Finally, the most important part of the event is a round table discussion when all participants are given the floor to share their thoughts which helps us to find out the needs of current and prospective CLARIN service users and, at the same time, offer our help or give explanations in response to some of the questions posed.

The outcomes of these discussions are especially helpful when scheduling our yearly activities and help to share our experiences with everyone interested in digital humanities within and outside academia. Each year we receive more and more inquiries to present the mission and goals of CLARIN ERIC and help with various research related questions, which we are always willing to do.

[47] http://clarin-lt.lt/?page_id=458

INTERVIEW: **Erika Rimkutė**

Erika Rimkutė is Senior Researcher at the Centre of Computational Linguistics at Vytautas Magnus University. The following interview took place via e-mail and was conducted by Jurgita Vaičenonienė, edited by Darja Fišer and Jakob Lenardič.

1. Could you briefly tell us about your academic background? What motivated you to apply a computational approach to linguistics?

I studied linguistics at Vytautas Magnus University (VMU). I was inspired to take up computational and corpus linguistics by Prof. Rūta Petrauskaitė, who is the founder of corpus linguistics in Lithuania and was my thesis advisor. The topic of my PhD was morphological ambiguity, which I analysed using a morphologically annotated corpus of Lithuanian. I defended my PhD in 2006 and am now a researcher at the Centre of Computational Linguistics and a lecturer at the Department of Lithuanian Studies at VMU.

I've been a member of the Centre of Computational Linguistics at VMU since my MA studies, which has given me a lot of valuable opportunities to get involved in computational research. I was able to get acquainted with specialists in corpus and computational linguistics working in Lithuania and other countries, to try out different language analysis software and corpora, as well as to observe other developments in language research. The experience gained in this way allowed me to specialise in automatic morphological analysis.

2. You've worked on quite a few important language projects with the Lithuanian CLARIN consortium. What have your contributions been when collaborating with the consortium?

I've had the opportunity to contribute to the creation of some of the key resources in the CLARIN-LT infrastructure. For example, the first version of MATAS[48] was compiled during my PhD studies to analyse the problem of morphological ambiguity, which had previously been very limitedly investigated in Lithuania and abroad. Manual annotation of semi-automatically annotated texts helped me to describe this phenomenon in detail in my dissertation, which contributed to the development of more accurate automatic morphological annotation tools for Lithuanian. The revised version of MATAS was added to the CLARIN-LT repository, so that it is now available for anyone interested in it.

[48] http://hdl.handle.net/20.500.11821/9

3. Would you like to recommend a language resource or tool developed at the consortium that you think is important for the study and analysis of the Lithuanian language?

Since 2016, I've been leading the project Automatic Identification of Lithuanian Multi-word Expressions financed by the Research Council of Lithuania. The project aims to develop a methodology for analysing Lithuanian MWEs by creating or adapting necessary tools and resources. Apart from the MWE identification methodology, we also aim to create MWE extraction tools, a database of Lithuanian MWEs with multifunctional search options, and a corpus-based dictionary of Lithuanian collocations.

CLARIN-LT is a partner on the project, which to me is an example of a successful collaboration between CLARIN-LT and linguists. CLARIN-LT provides me with the technical support and creates the tools necessary for the implementation of the project. In return for their support, we will upload all the results into the CLARIN-LT repository and make them easily accessible for other researchers. For example, at the end of the year, the first dictionary of Lithuanian collocations will be released. Users will be able to access a database of Lithuanian multiword units encompassing over 10,000 lemmas. I think that this is an important contribution both for the development of further lexicographic resources as well as language teaching, especially given that collocation dictionaries don't yet exist for most under-resourced languages like Lithuanian.

4. You are also a teacher at the Department of Lithuanian Studies at the Vytautas Magnus University. How do you integrate the computational approach into your course-work? Do you introduce the CLARIN infrastructure to your students?

I cannot imagine my classes without introducing students to the morphologically and syntactically annotated corpora. Naturally, before starting work with the resources I introduce the main principles of CLARIN and the role of national repositories. I always encourage my students to use the Corpus of Contemporary Lithuanian Language, which was developed by CLARIN-LT, for example. Although corpora do not provide ready-made information, in contrast to dictionaries, I believe it is vital to teach students the importance of making linguistic claims on the basis of authentic language use. The students of BA and MA study programmes of Lithuanian Philology and Modern Linguistics, where I teach, are taught to work with the Lithuanian Morphologically Annotated Corpus MATAS during the lectures on morphology and word formation. The students have to identify the missing node in the collocations extracted from the Corpus of Contemporary Lithuanian Language; identify parts of speech and grammatical categories in extracts from MATAS; analyse syntactic relations in ALKSNIS, etc. Apart from in-class activities, students also write seminar papers and BA and MA theses drawing on data extracted from the mentioned resources, some of them are even invited to work on our research projects. For example, Rūta Brinkutė's MA thesis analyses the distribution of grammatical categories in different genres. I believe that knowing how to work with annotated corpora and tools might be valuable for students in their future work as language editors or researchers.

5. You have been part of the team that created the LILA corpus,[49] which is a parallel corpus of Lithuanian and Latvian. The team included both Lithuanian and Latvian researchers involved with CLARIN. How does the Lithuanian CLARIN consortium benefit from such cross-border collaborations? Do you plan to upload the corpus into the consortium's repository?

The project was part of the EU Cross-Border Cooperation Programme and was conducted in 2011-2012 before either Lithuania or Latvia were CLARIN members. The nine million-word Lithuanian-Latvian-Lithuanian parallel corpus aligned on paragraph and sentence level was compiled by researchers of the Vytautas Magnus University's Centre of Computational Linguistics and the Latvian University's Mathematical and Informatics Institute's Laboratory of Artificial Intelligence (LU MII). It was a really interesting and mutually beneficial experience to work with my Latvian colleagues, as our teamwork not only resulted in the creation of the corpus itself, but also in several joint publications. I believe that if the project was implemented now, when both countries have CLARIN centres, the project aims and results could have been formulated on a much larger scale and more language pairs could have been included in the corpus. I see great value in such collaborative projects, as they allow us to combine a wide variety of research perspectives and approaches, which in turn enhances professional and personal cooperation between the research centres and scientists in different countries.

6. What would you recommend CLARIN to do in order to attract more researchers from the Lithuanian linguistics community?

In relation to my previous comment on the LILA corpus, I think that CLARIN could focus more on promoting joint scientific projects among the CLARIN centres of different countries to create comparable language resources and compatible processing tools. I also think that the fact that there is a consortium like CLARIN-LT which develops tools and resources specifically dedicated to Lithuanian can be very inspiring for new initiatives and research projects that also might want to start working with other less-resourced languages.

Also, I would like to see interoperable lexicographical databases to become available through CLARIN. At the very least, providing more information on the availability of such resources would be very helpful. For example, during the lexicographic project "Automatic Identification of Lithuanian Multi-Word Expressions", we were looking for a database we could reuse for our research. As we found none, we spent a lot of time as well as human and financial resources to create the database ourselves.

[49] http://tekstynas.vdu.lt/page.xhtml?id=parallelLILA

Klaipeda, Lithuania | photo by Åsa K | Pixabay

COLOPHON

Coordinated by
Darja Fišer, Jakob Lenardič and Karolina Badzmierowska

Edited by
Darja Fišer and Jakob Lenardič

Proofread by
Paul Steed

Designed by
Karolina Badzmierowska

Map on the cover: Designed by Freepik.

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

Publication number
CLARIN-CE-2018-1341
November 2018

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 2018