CLARIN

Common Language Resources and Technology Infrastructure

Tour de CLARIN

VOLUME **TWO**

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

Common Language Resources and Technology Infrastructure

Tour de CLARIN

VOLUME **TWO**

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

CONSORTIA

K-CENTRES

Foreword

Since 2016, the Tour de CLARIN initiative has been periodically highlighting prominent user involvement activities in the CLARIN network in order to increase the visibility of its members, reveal the richness of the CLARIN landscape, and display the full range of activities that show what CLARIN has to offer to researchers, teachers, students, professionals and the general public interested in using and processing language data in various forms. The initiative was initially conceived as a series of blog posts published on the CLARIN webpage and disseminated through the CLARIN newsflash and social media channels. In 2018, the results of the initiative were published in a printed volume. Gradually, Tour de CLARIN has proven to be one of the flagship user involvement initiatives by CLARIN ERIC, is highly valuable for our network and incredibly popular with our readers. That is why the initiative has since been expanded from presentations of the work carried out by national consortia to also feature the work of CLARIN Knowledge Centres, which provide a physical or virtual place where researchers, educators and developers alike can get cross-border access to knowledge and expertise in specific areas.

As a reflection of the double focus, this second volume of Tour de CLARIN is organized into two parts. In Part 1, we present the seven countries which have been featured since November 2018, when the first volume was published: Estonia, Latvia, Denmark, Italy, Slovenia, Hungary, and Bulgaria. In this part, each country is presented with five chapters: an introduction to the consortium, their members and their work; a description of one of their key resources; a presentation of an outstanding tool; an account of 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.

In Part 2, we present the work of the four Knowledge Centres that have been visited thus far: the Knowledge Centre for Treebanking, the Knowledge Centre for the Languages of Sweden, the TalkBank Knowledge Centre, and the Czech Knowledge Centre for Corpus Linguistics. In this part, each K-centre is presented with two chapters: a presentation of what the K-centre offers to researchers; and an interview with either a renowned researcher who has collaborated with the K-centre or a leading developer who is part of the centre itself and who provided as with valuable insight into what the centre offers.

This second volume would not have been possible without the contributions and dedication of the national user involvement representatives and national coordinators: : Olga Gerassimenko and Kadri Vider from Estonia, Ilze Auziņa and Inguna Skadiņa from Latvia, Valeria Quochi and Monica Monachini from Italy, Lene Offersgaard and Costanza Navarreta from Denmark, Nikola Ljubešić and Tomaž Erjavec from Slovenia, Réka Dodé and Tamás Váradi from Hungary, Petya Osenova and Kiril Simov from Bulgaria. We are equally grateful for the contributions by the Knowledge Centre coordinators: Koenraad De Smedt and Jan Hajič from the Knowledge Centre for Treebanking, Rickard Domeij from the Knowledge Centre for the Languages of Sweden, Brian MacWhinney from the TalkBank Knowledge Centre, and Michal Křen from the Czech Knowledge Centre for Corpus Linguistics. We would also like to thank all the researchers who have kindly agreed to be interviewed for their time and invaluable insights: Marin Laak, Sanita Reinsone, Beatrice Nava, Klaus Nielsen, Kaja Dobrovoljc, Noémi Vadász, Aneta Nedyalkova, Helge Dyvik, Susanne Nylund Skog, Nan Bernstein Ratner, and Ondřej Tichý.

Tour de CLARIN will continue to visit CLARIN member countries and K-centres and present their success stories online as well as in future printed volumes.

Darja Fišer and **Jakob Lenardič**

Ljubljana, Slovenia
November 2019

Consortia featured in this volume:

Estonia

Latvia

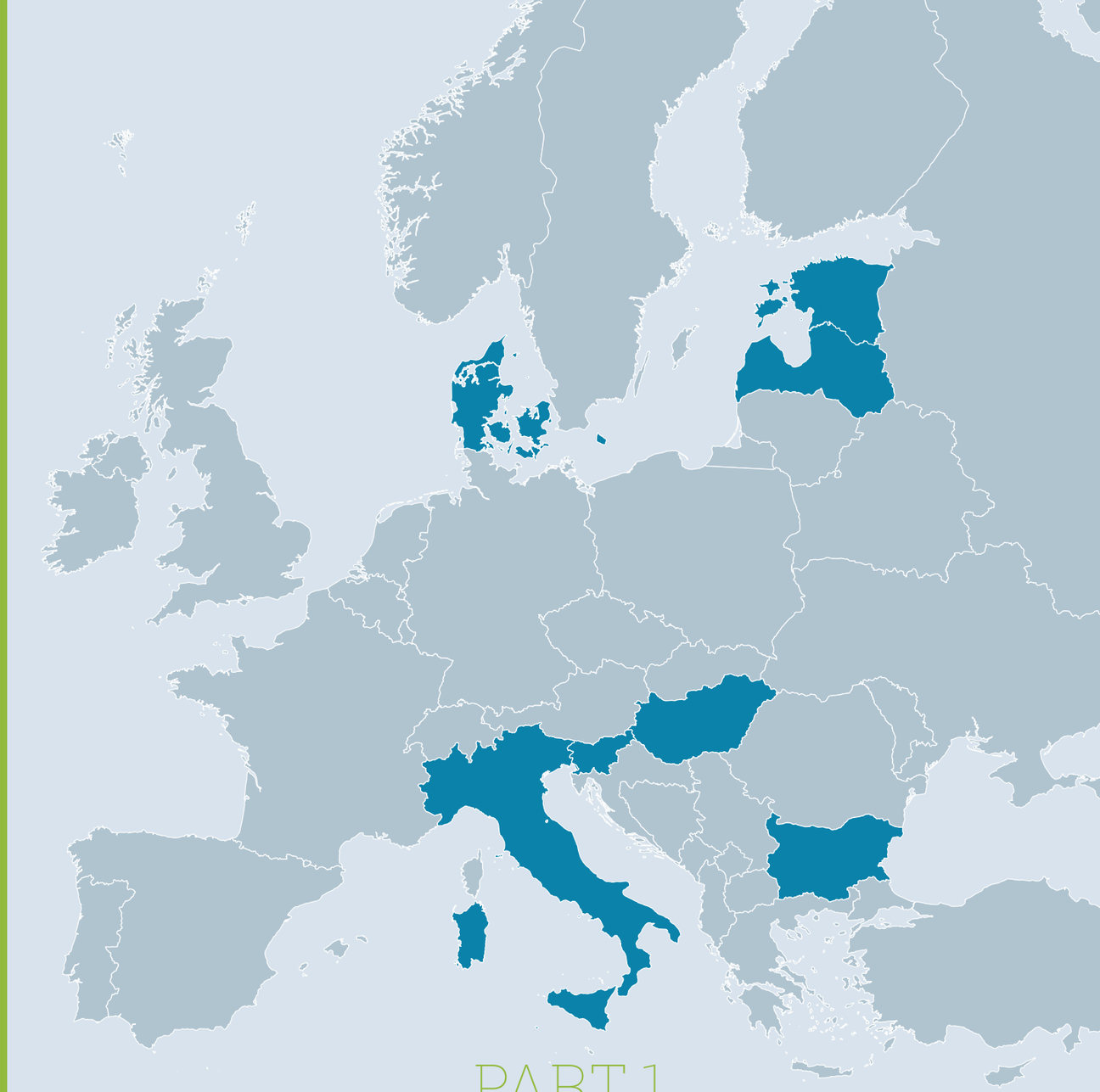
Italy

Denmark

Slovenia

Hungary

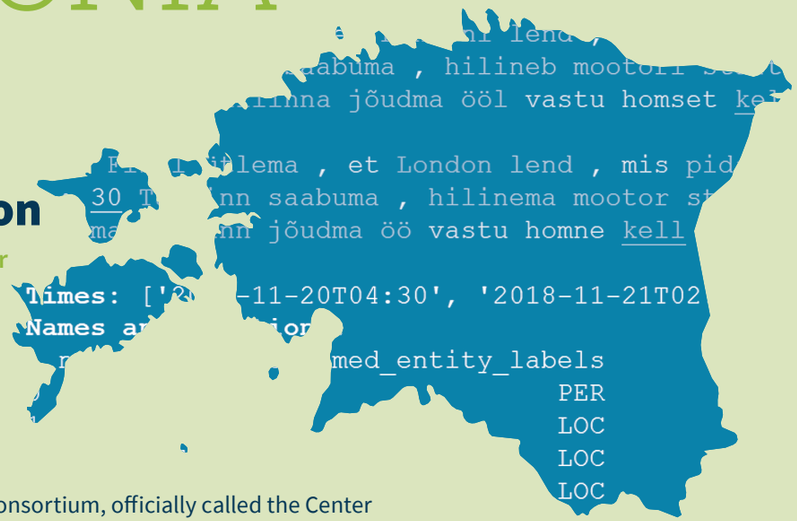
Bulgaria



PART 1

CONSORTIA

ESTONIA



Introduction

Written by **Kadri Vider**

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

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

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

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

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

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

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

The Estonian CLARIN team

Tool | **EstNLTK**

Written by **Krista Liin**

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

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

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

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

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

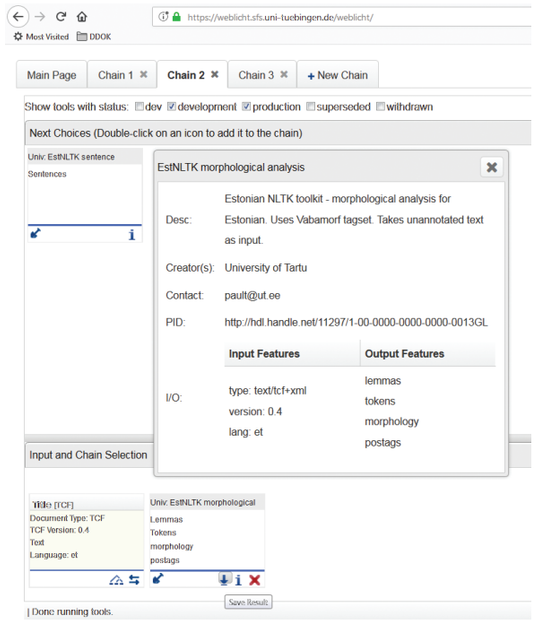


Figure 1: *Using EstNLTK in a WebLicht workflow for morphological analysis*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Reference:

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

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

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

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

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

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

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

Aleksei Kelli leading the workshop “Copyright and Cultural Heritage”

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

Interview | **Marin Laak**

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

Could you briefly tell us about your academic background?

<

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

>

How did you get involved with the Estonian CLARIN consortium?

<

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

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

>

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

<

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

>

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

<

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

>

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

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

<

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

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

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

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

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

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

>

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

<

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

>

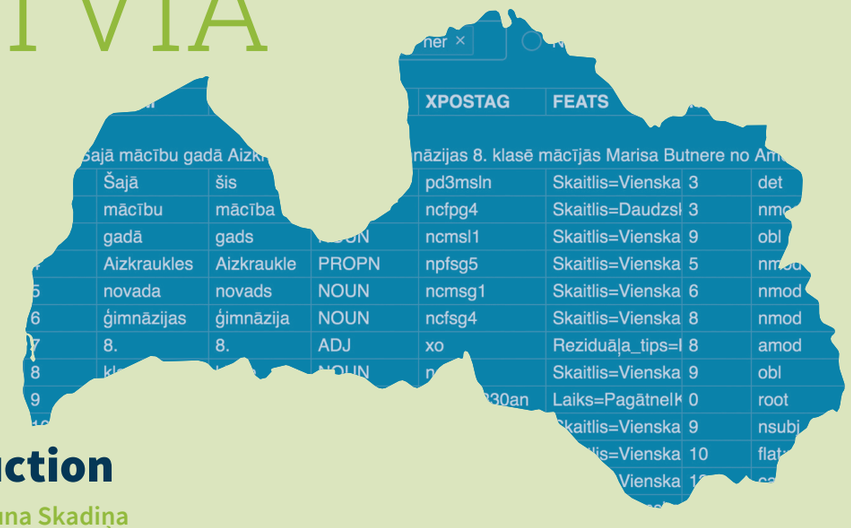
In your opinion, how can research infrastructures like CLARIN help museums (staff and visitors alike)?

<

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

>

LATVIA



Introduction

Written by **Inguna Skadiņa**

Latvia joined CLARIN ERIC in June 2016. The national coordinator of CLARIN Latvia is Inguna Skadiņa, the user involvement activities are led by Ilze Auziņa, while Roberts Darģis is involved in the Centre Committee.[6] The coordinating centre of CLARIN Latvia is the Artificial Intelligence Laboratory of the Institute of Mathematics and Computer Science, University of Latvia. The laboratory has been conducting research on natural language processing and has provided access to different language resources, including corpora and lexicons (e.g. tezaurs.lv), for almost 30 years. Prominent corpora offered by CLARIN Latvia, most of which are available through online concordancers like noSketchEngine, include:

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

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

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

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

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

Members of the Artificial Intelligence Laboratory at a brainstorming session

Tool | **NLP-PIPE**

Written by **Artūrs Znotiņš**

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

NLP-PIPE supports a wide range of annotation services for Latvian, including tokenization, morphological tagging, lemmatization, universal dependency parsing, and named entity recognition. The easiest way to start using the toolchain is via the on-line demo version. In the web based interface, a user simply selects the required processing tools and inputs the text they want to annotate. The results can then be viewed either directly on the website (see Figure 5) or exported in several formats.

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

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

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

References:

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

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

Resource | **Latvian FrameNet**

Written by **Normunds Grūzītis**

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

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

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

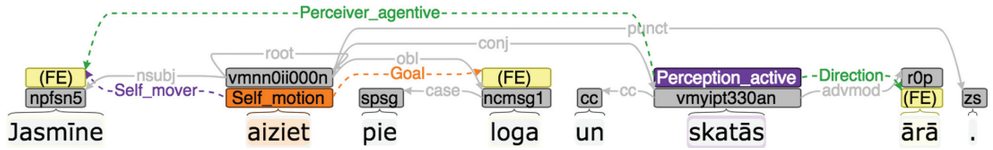


Figure 6: *Latvian FrameNet annotation*

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

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

Acknowledgements

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

References:

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

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

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

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

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

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

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

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

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

Written by **Ilze Auziņa**

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

Normunds Grūzītis opening the seminar

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

Inguna Skadiņa introducing CLARIN

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

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

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

Interview | **Sanita Reinsone**

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

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

<

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

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

>

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

<

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

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

>

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

<

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

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

>

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

<

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

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

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

>

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

<

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

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

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

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

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

>

How can research infrastructures such as CLARIN benefit from crowdsourcing?

<

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

>

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

<

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

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

>

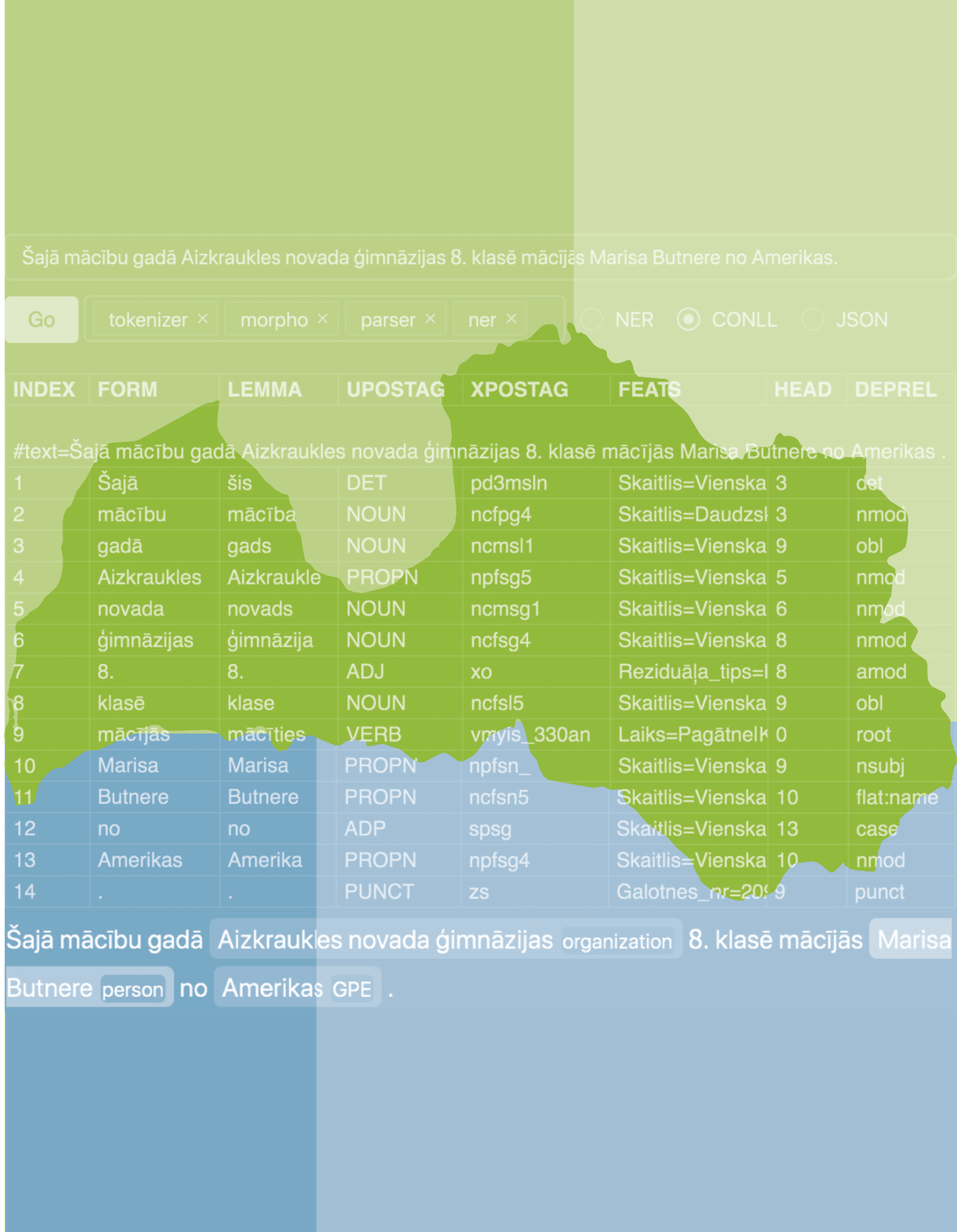
How in your opinion could CLARIN Latvia help promote computational methods and the use of research infrastructures in traditional fields such as your own (i.e., folklore studies)?

<

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

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

>



ITALY

Introduction

Written by **Monica Monachini** and **Valeria Quochi**

The Italian CLARIN consortium, CLARIN-IT, has been a member of CLARIN ERIC since October 2015. The Italian National Coordinator is Monica Monachini. The consortium comprises five full members:

- the Institute for Computational Linguistics “A. Zampolli”, which is the founding and coordinating node of CLARIN-IT (Monica Monachini);
- the Department of Education, Human Sciences and Intercultural Communication at the University of Siena (Silvia Calamai);
- the Centre for Comparative Studies “I Deug-Su” at the Department of Philology and Literary Criticism at the University of Siena (Francesco Vincenzo Stella);
- the Institute for Applied Linguistics at Eurac Research (Andrea Abel);
- Fondazione Bruno Kessler (FBK) (Sara Tonelli).

A number of other Italian institutions have expressed their interest in participating in the consortium in the future, including the University of Pisa, the Scuola Normale Superiore and the University of Parma. Professor Anika Nicolosi at the University of Parma is currently involved with CLARIN-IT as an expert of Classics and Philology. CLARIN-IT also closely cooperates with the Consortium GARR on technical issues, in particular with the IDEM-GARR office that supports federated authentication in CLARIN. Because of this cooperation, any member or participant of the IDEM-GARR federation has access to the resources and services hosted in any CLARIN centre via their institutional credentials.

CLARIN-IT has established two national centres: ILC4CLARIN, which is hosted by the Institute for Computational Linguistics “A. Zampolli” in Pisa and is a B-certified repository that has been active since 2016, and ERCC, which is hosted by EURAC Research in Bozen and is currently a C-certified repository that has been active since 2018 and aims to become a B-certified centre. Through the two repositories, CLARIN-IT offers a variety of resources and services, such as MERLIN, which is a multilingual learner corpus for German, Italian and Czech, and SIMPLE, which is a multi-layer lexicon of Italian based on the Generative Lexicon Theory. There are also several natural language processing and analysis tools, many of which are offered as web services and integrated into Weblicht.

The Italian consortium focuses on the field of Digital Classics, which still suffers from shortage or restricted availability of lexical resources for historical languages such as Ancient Greek, Latin or Sanskrit. To this end, the consortium aims to make some of the existing digitized resources for Ancient Greek and Latin available through its repositories (e.g. an LOD version of the TEI-dict Perseus Liddell-Scott Jones dictionary), as well as to create new ones by enriching existing corpora and lexical datasets with Linked Open Data. CLARIN-IT also specializes in the research of non-standard forms of language as found in learner corpora and computer-mediated communication. Moreover, CLARIN-IT focuses on oral archives which are at the crossroads of speech sciences, digital humanities and digital heritage.

Monica Monachini (second from the right) and some members of ILC4CLARIN: Francesca Frontini, Fahad Khan, Andrea Bellandi, and Federico Boschetti

Some consortium members are involved in international infrastructural projects aiming to strengthen the cohesion of research across a number of related fields associated with the humanities, such as PARTHENOS (language studies and cultural heritage), ELEXIS (e-lexicography) and the recently established SSHOC project (European open cloud ecosystem of data and tools for social sciences and humanities). They are also active in standardization initiatives, such as ISO, TEI, W3, and international academic organizations and networks, such as Learner Corpus Association, Special Interest Groups on CMC and the COST Action European Network for Combining Language Learning with Crowdsourcing Techniques.

Valeria Quochi (CLARIN-IT User Involvement Manager) and Alessandro Enea (ILC4CLARIN Technical Manager)

Paola Baroni (CLARIN-IT Membership, Web and Communication Manager | ILC4CLARIN Web and Communication Manager) and Riccardo Del Gratta (ILC4CLARIN Repository Manager)

ERCC staff: Egon W. Stemle, Lionel Nicolas, Andrea Abel, Verena Lyding, and Alexander König

Tool | **LexO: Where Lexicography Meets the Semantic Web**

Written by **Andrea Bellandi, Monica Monachini** and **Fahad Khan**

LexO is a collaborative web editor used for the creation and management of (multilingual) lexical and terminological resources as linked data resources. The editor makes use of Semantic Web technologies (which enrich web data with semantic information in order to make them machine-readable) and the linked data publishing paradigm in order to ensure that lexical resources can be more easily shared and reused by the scientific community. In particular, LexO offers the following functionalities:

- It hides all the technical complexities related to markup languages, language formalities and other technology issues, facilitating access to the Semantic Web technologies to non-expert users who have not yet mastered Semantic Web-based standards and technologies, such as the Resource Description Framework and the Web Ontology Language (OWL).
- It provides the possibility for a team of users, each one with his/her own role (lexicographers, domain experts, scholars, etc.) to work on the same resource collaboratively.
- It adheres to international standards for representing lexica and ontologies in the Semantic Web (such as OntoLex-Lemon and OWL), so that lexical resources can be shared easily or specific entities can be linked to existing datasets (it is based on the OntoLex-Lemon model, currently regarded as a de facto standard for the modelling and publication of lexical resources as linked data).
- It provides a set of services implemented by means of RESTful Web Services that allow software agents access to resources managed by LexO.

LexO: uses, lexical resources and communities

LexO has so far been used in several DH research projects, such as:

- DiTMAO, a born-digital multilingual medico-botanical terminology focused on Old Occitan developed by philologists;
- FdS, a multilingual diachronic lexicon of Saussurean terminology in the framework of a lexicographic project;
- Totus Mundus, a bilingual Chinese-Italian resource dealing with Matteo Ricci's Atlas. LexO has been used by historians to build the linguistic resources related to the map.

In addition, a demo of LexO with a subset of italwordnet adjectives is available as an online service through ILC4CLARIN.

The Interface

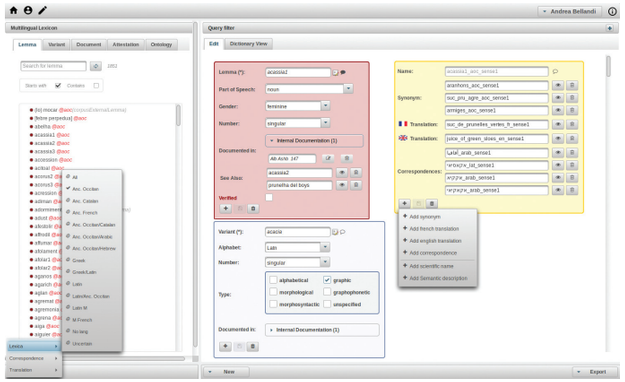


Figure 7: *The main interface of LexO*

The LexO interface is composed of two main sections (Figure 7). Depending on which tab is selected, the left-hand side column will either show the list of lemmas composing the resource, a list of word forms, a list of lexical senses, or a list of concepts belonging to a reference ontology. If the resource is multilingual, then users have the possibility of filtering lemmas, forms and senses by language. Information related to the selected entry is shown in the central panel where the lemma appears in the upper part of the leftmost column on at head of a list of related forms. On the right, the lexical senses are shown.

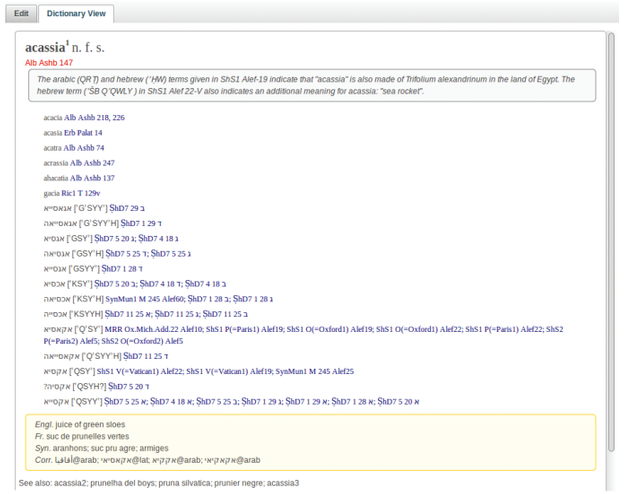


Figure 8: *Dictionary view of LexO. By selecting the "Dictionary View" tab, the central panel shows a dictionary-like rendering of all the information related to the selected entry*

When selecting the "Dictionary View" tab, the central panel will show a dictionary-like rendering of all the information related to the selected entry (Figure 8). At the top of the central panel, a section can be expanded to query the resource, either by filling a series of fields for advanced searching (Figure 9b) or by composing queries in a controlled natural language style interface (Figure 9c). A team of users can work simultaneously in LexO to create, modify or delete a lexical entry, form, or sense, or to connect an entity to another entity, such as a sense to another sense via the "synonymy" property or a sense to a concept via the "ontological reference" property. The ontology can be imported using a dedicated tab in the left column (Figure 9a). Finally, administrators can monitor the lexicon construction process, for example by adding/removing users to the team, monitoring their productivity, and access the basic statistics of the lexicon (Figure 10).

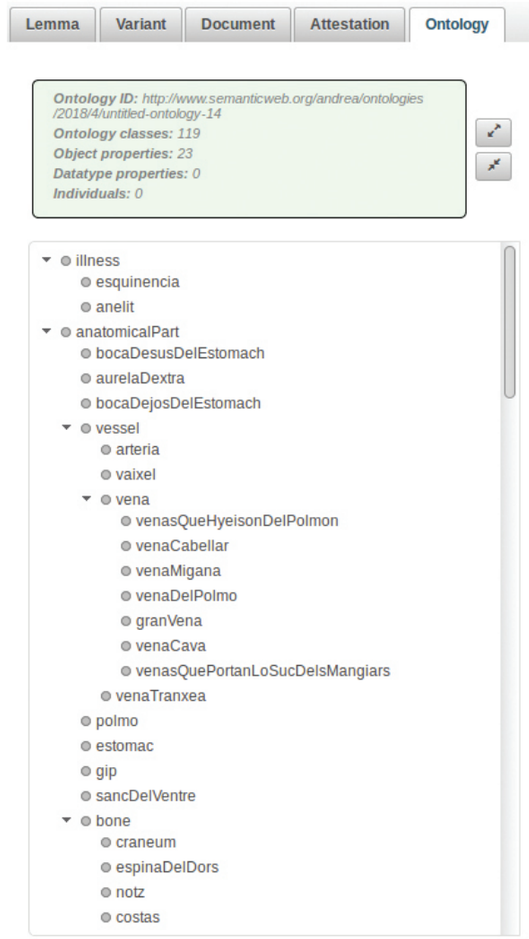


Figure 9a: *Imported ontology*

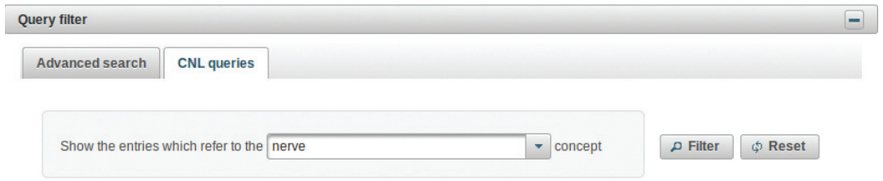


Figure 9b: *Querying the lexicon by means of ontology concepts*

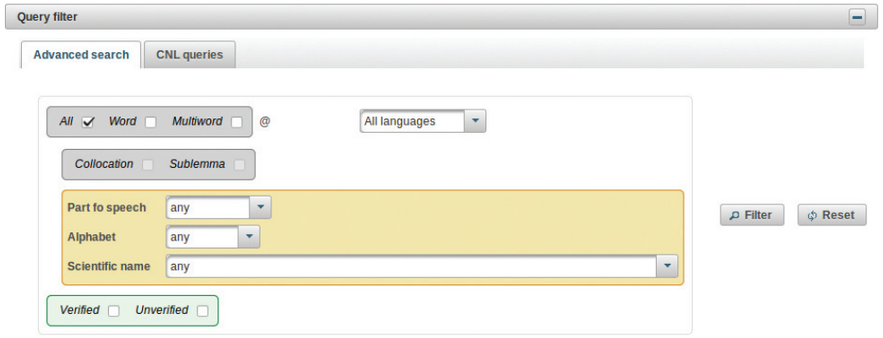


Figure 9c: *Lexicon search panel*

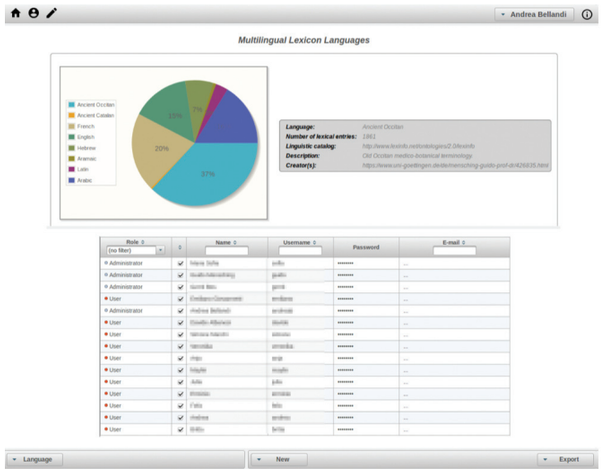


Figure 10: *Administration panel*

References:

Bellandi, A., Giovannetti, E., and Weingart, A. 2018. Multilingual and Multiword Phenomena in a lemon Old Occitan Medico-Botanical Lexicon. *Information* 9 (3): 1–52.

Khan, F., Bellandi, A., and Monachini, M. 2016. Tools and instruments for building and querying diachronic computational lexica. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities* (LT4DH), 164–171.

McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P., and Cimiano, P. 2017. The OntoLex-Lemon Model: Development and Applications. In *Proceedings of eLex 2017 Conference*, 19–21.

Resource | **MERLIN – A Written Learner Corpus for Czech, German and Italian**

Written by **Alexander König** and **Monica Monachini**

The MERLIN corpus is a written learner corpus for Czech, German, and Italian.[7] The corpus is composed of over 2,200 texts, about 1,000 in German, 800 in Czech and over 400 in Italian, and can be downloaded in various formats from the ERCC repository of the Italian CLARIN consortium. The corpus can also be browsed online via a multi-functional web interface that enables users to explore authentic written learner productions in relation to their CEFR classification and annotated learner features.

The corpus has been designed to illustrate the Common European Framework of Reference for Languages (CEFR) with richly annotated authentic learner data. Since its publication in 2001, the CEFR has become the leading instrument of reference for the teaching and certification of languages and for the development of curricula. At the same time, there is a growing concern that the CEFR reference levels are not sufficiently illustrated, leaving practitioners such as teachers, test and curriculum developers, and textbook authors without comprehensive empirical characterizations of the relevant distinctions between the proficiency levels. This is particularly true for languages other than English, where supplementary empirical tools are urgently needed.

The MERLIN corpus was designed to address this demand for the three languages of Czech, German and Italian, by annotating authentic written learner productions and relating them to CEFR in a methodologically sophisticated way. To create the corpus, the partners relied on existing corpus annotation and search tools as much as possible. As no single tool was able to fulfil all the annotation requirements, a combination of tools was required to support the wide range of manual and automatic annotation that had been designed to illustrate the CEFR scales.

The manual annotation, which includes error annotation and the linguistic characteristics of the learner language, was performed using the Falko add-on for Microsoft Excel, which provides an existing framework for annotating learners' errors, and the MMAX2 multi-level annotation tool, which is a flexible GUI-based tool for creating new annotations as well as visualizing them. Parallel to the manual annotation, the developers of CLARIN-IT created a custom UIMA toolchain in order to enrich the corpus with additional layers of linguistic annotation, such as part-of-speech tagging and syntactic parsing. All in all, the texts were annotated with about 70 different features, covering orthography, grammar and lexicon of the learner language as well as specific sociolinguistic or pragmatic characteristics. This regards features such as the appropriate use of formality/politeness, e.g. the T/V distinction in German, or of idiomatic expressions like greetings or closing formulae.

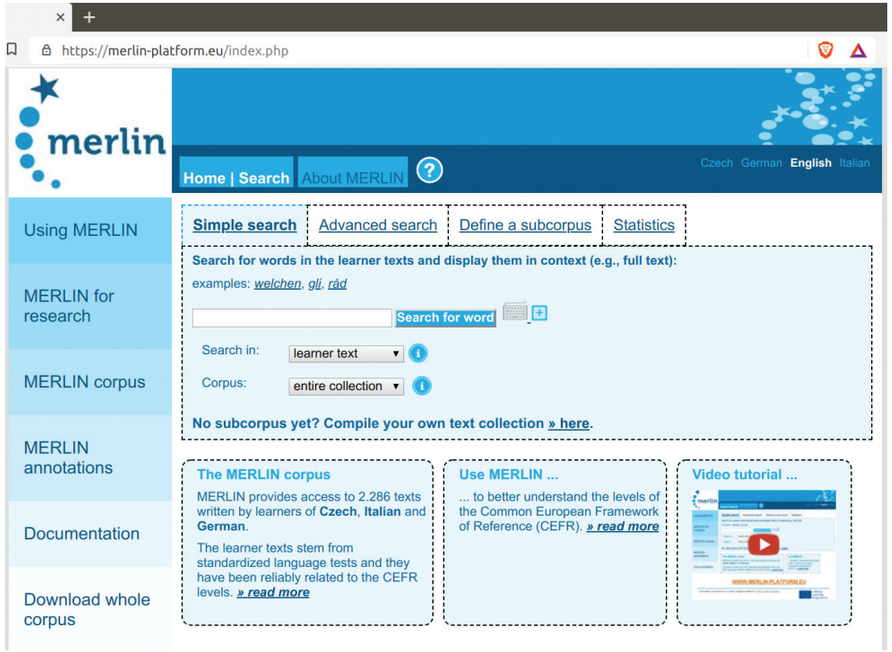


Figure 11: *The online search interface of the MERLIN corpus*

MERLIN is now mainly used by linguists specialized in learner language, but also teachers and language test developers who use richly annotated authentic examples to improve their methodology. The MERLIN online platform is especially crucial for language teachers, as it provides ready-made usage scenarios in Czech, German, and Italian which show how the corpus can be used for data-driven teaching in a classroom environment. In this respect, the online platform also gives access to several pre-prepared language learning tasks that students can solve by using the corpus. There is also a YouTube demonstration that is aimed at language teachers and shows how the corpus can be used as part of the syllabus.

[7] http://hdl.handle.net/20.500.12124/6

Tok	=	automatically tokenized and manually checked learner text
Correctly represented learner production		
Ctok	=	level for emergency corrections of tokenization/transcription
Perspective Ia (orthography & grammar errors)		
TH1(ZH1)		Target Hypothesis 1 (complete, corrected learner text)
ZH1Diff	=	level for marking differences between TH1 and ctok
ZH1spec	=	level 1-3 for marking speculative hypotheses
EA1_lev1,2,3		Error annotation 1, specified on level 1-3
EA1_tlm		target language modification
Perspective Ib (vocabulary, coherence, sociolinguistic, pragmatic errors)		
TH2(ZH2)		Target Hypothesis 2
EA2_lev1,2,3		Error annotation 2, specified on level 1-3
EA2_tlm		target language modification
Perspective II (learner language features that are not related to errors)		
LLF		Non-error related learner language features
LLF_lev1-3		Specification of not error-related phenomena according to Annotation Scheme

Figure 12: *A schema of the annotation levels in the corpus, which include the mark-up of both word/sentence- (e.g., orthography) and discourse-level (e.g., errors in achieving coherence) errors*

The corpus was collected from 2012 to 2014 within the project MERLIN "Multilingual Platform for the European Reference Levels: Interlanguage Exploration in Context". The project was funded by the EU Lifelong Learning Programme with a consortium of seven partners: Technische Universität Dresden (DE) as the Lead Partner, the European Academy Bolzano (IT), Charles University (CZ), telc GmbH (DE), Berufsförderungsinstitut Oberösterreich (AT), Eberhard-Karls-Universität Tübingen (DE), and finally the European Centre for Modern Languages of the Council of Europe (AT) as Associated Partners.

The corpus has also been successfully used in several master's theses:

- Tina Schönfelder 2014. *REQUESTS im Italienischen und Deutschen als Fremdsprache* ("REQUESTS in Italian and German as Foreign Languages").
- Tassja Weber. 2013. *Verbvalenz und Rektion im Bereich Deutsch als Fremdsprache. Eine korpusgestützte Analyse zweier Verbgruppen* ("Valency and Case in German for Special Purposes as a Foreign Language").
- Julia Hancke. 2013. *Automatic Prediction of CEFR Proficiency Levels Based on Linguistic Features of Learner Language.*

References:

Abel, A and Wisniewski, K. 2015. MERLIN - die mehrsprachige Plattform für die europäischen Referenzniveaus at the 6th (Österreichische Gesellschaft für Sprachendidaktik) ÖGSD Conference in Salzburg.

Wisniewski, K. 2015. Empirisch gestützte Arbeit mit dem GeRS: Zur Einschätzung schriftlicher Leistungen in Deutsch, Tschechisch und Italienisch als Fremdsprachen mit dem Lernerkorpus MERLIN. 26. Kongress der deutschen Gesellschaft für Fremdsprachenforschung in Ludwigsburg.

Event | **Roadshow Seminars**

Written by **Monica Monachini** and **Valeria Quochi**

Since CLARIN-IT was established in 2016, its members have been organizing a series of roadshow events aimed at the Italian Digital Humanities and Social Sciences community. At the roadshows, CLARIN-IT experts present tools and resources deposited in CLARIN repositories (such as the Italian ILC4CLARIN data centre), offer examples of how CLARIN helps to promote novel research with NLP tools, as well as provide guidelines on how to eradicate bottlenecks that hamper the growth of a newly established digital discipline and, ultimately, in which ways scholars can profit from research infrastructures.

One of the most prominent events that the roadshow visited was the 5th Annual Conference of the Association for Humanities Information Sciences and Digital Culture, where the CLARIN-IT Coordinator Monica Monachini gave an invited talk about CLARIN in which she outlined how the Digital Humanities research community in Italy can benefit from the language technologies offered by CLARIN as well as from being involved in a large international research network based on interdisciplinary collaborations in a digital framework.

In order to help train young researchers and thus guarantee a broader utilization of computational tools and methods, a series of roadshow seminars aimed at students was also organized. The first seminars took place in October 2016 at the University of Parma and were intended both to promote CLARIN among students of Ancient Greek and to motivate them to adopt methods and concepts of computational linguistics and Digital Humanities in their studies. A second round took place from November to December 2017 with lectures concentrating on the methods, resources and instruments of a digital approach to philology. This included the encoding of text variants, digital repertoires of multiple editions of the same text and tools for their automatic alignment. During the lessons, CLARIN-IT provided examples of application of the TEI markup and Semantic Web technologies, by annotating geographical and personal references and linking an Ancient Greek lexicon in the Linked Open Data paradigm with a TEI-encoded fragmentary text of the poet Archilochus.

Other roadshow events were aimed at the promotion of CLARIN among graduate students with lectures at the "Master in Digital Humanities" event at Ca' Foscari University on 3 November 2017 and as part of the "Digital Humanities, Web Resources, and Infrastructures" course at Venice International University on 4 December 2017. In addition, CLARIN-IT information events aimed at introducing the consortium, as well as CLARIN ERIC, to decision-making university figures and

to new potential infrastructure partners and providers, such as the meeting on digital research infrastructures on 29 June 2018, which also involved members from the Italian academic senate along with professors and researchers in Digital Humanities and Social Sciences. A series of working days in 2017 and 2019 were dedicated to critical issues related to audio archives, such as legal aspects involved in collecting and (re)using audio data and possible ways to promote collaboration among linguists, speech scientists, speech technologists, oral historians and infrastructures.

Monica Monachini giving an invited talk about CLARIN on 26 June 2017 at the University of Pisa

Interview | **Beatrice Nava**

Beatrice Nava is a PhD student who uses digital methodologies in Classical Studies.

Please describe your research background.

<

I have a bachelor's degree in Greek Philology and a master's degree in Modern Philology. I have always been interested in text reconstruction, and chose philology because I believe it is a crucial starting point for critical research into texts and the cultural contexts in which they are produced. My first hands-on experience in Digital Humanities was on the modelling of metadata of literary sources, which was made possible through a GARR grant (2017-2018). I am currently pursuing a PhD in Literary and Philological Cultures at the University of Bologna, where I am preparing a critical edition (i.e., a scholarly edition that includes a "critical apparatus" – annotations on the primary source material of a text) of a tragedy written by the famous Italian Romantic poet Manzoni and developing a model for digitizing the critical edition of the tragedy. I am also collaborating on a project funded by the Italian Ministry of Education, Universities and Research which aims to create a web portal dedicated to Manzoni. My principal role in this project is the digitization of Manzoni's works using XML/TEI encoding. Additionally, I am using the same approach in the DEA project that focuses on Greek philology.

>

How has getting to know CLARIN influenced your research directions?

<

CLARIN was the starting point for my interest in Digital Humanities, and has made the opportunity to work in this field more realistic. It has motivated me to apply for a GARR grant with a proposal to model metadata for the description of Alessandro Manzoni's manuscripts, which became the basis of my PhD project.

More recently, due to my involvement in the DEA project, my collaboration with CLARIN-IT has intensified. It has provided me with new research directions and interest in computational linguistics, and has shown me how Digital Humanities can facilitate a transversal and interdisciplinary approach to different fields. In particular, I think that CLARIN-IT's focus on Digital Classics has encouraged my methodological transition from traditional to digital philology.

>

Please describe the DEA project. How is it involved with CLARIN-IT?

<

DEA, which stands for *Digital Edition of Archilochus: New models and tools for authoring, editing and indexing an ancient Greek fragmentary author*, is a project led by principal investigator Anika Nicolosi (University of Parma) done in collaboration with ILC-CNR of Pisa and CLARIN-IT. The goal of the project is to create a complete digital edition of the fragments by the Greek lyric poet Archilochus. We have around 300 fragmentary poems by this important author who lived in the 7th century BC and was closely related to Homer. However, a complete critical edition of his works is still lacking. In fact, some fragments have only recently been published, so they are not yet included in the most widely used editions. The main objective of the project is to provide scientifically reliable texts, with critical apparatuses, commentaries and translations, and to make available an online and easily accessible augmented corpus of ancient Greek fragmentary literature.

DEA can be regarded as a case study in the framework of CLARIN-IT and its interests in specialization towards the Digital Classics. ILC4CLARIN offers the corpus in their repository, along with other existing digitized resources for Ancient Greek (e.g. a Linked Open Data (LOD) version of the TEI-dict Perseus Liddell-Scott Jones Greek-English dictionary[8]). This allows us to enrich our corpus with lexical datasets in LOD and integrate our data with other existing resources, with the final aim of obtaining a complete edition that is useful not only for scholars interested in Classical and Ancient Studies, but also for non-specialist users.

>

[8] http://lari-datasets.ilc.cnr.it/ml/

Why is the Digital Humanities approach important for classical philology? What kind of new research avenues does it open in the relatively traditional field?

<

The application of language technologies and methodologies to solve research questions in Classical Philology is very important in relation to the structural potential of the digital medium. For example, just to mention one of the well-known but essential aspects, which is the option of organizing, storing and managing a substantial amount of data. In our case, we can easily manage and store all the hypotheses of previous editors and the additional useful information linked to the edition in a single place. Therefore, by providing an edition that is richer and much more complete than a paper-based one, it is possible to facilitate new philological studies. In fact, offering all the interpretations of previous editors through a single resource, with the addition of new hypotheses formed by studying the whole corpus of fragments, reopens the debate on some critical points. What is also important is that the digital medium helps scholars to efficiently exchange their ideas and results, as well as accelerates the response to new interventions into the text.

In addition, linguistic annotation allows the development of new teaching methods of Ancient Greek that are aimed at beginners and include the use of language services, such as treebanks and tools like TüNDRA adapted for classics. The annotations also enable an interactive approach to texts that is more inviting and accessible to students.

In addition to making data accessible and interoperable, NLP approaches facilitate and allow for the systematic production of fragment-specific lexica. In our case, having an annotated corpus allows us to develop linguistic services for teaching (e.g., Hyper-Text Archilochus, which is a prototype that provides the learner with a set of resources and tools that ease the critical assessment of ancient texts). It also acts as a stable and immutable sample for automatic translation experiments.

>

What are the challenges of digitizing and applying NLP techniques to Ancient Greek poetry? How does fragmentary poetry differ from other literary texts and what does this entail for its processing?

<

Fragmentary ancient Greek poetry is very different from other literary texts. In fact, its tradition is more complex since it has different kinds of sources (manuscripts, papyrus, epigraphy) with variants and substantial lacunae (i.e., missing parts in a text). Applying NLP techniques to a fragmented tradition, which is complex and has many parts missing, can be particularly challenging, because gaps in the text call for multiple options for its reconstruction. Automatic

linguistic analyses of the whole corpus not only support new readings and interpretations, but also lead us to greater certainty as regards text corrections, integrations and authorship.

>

Which tools would you like to see CLARIN Italy develop next that would help researchers interested in classical philology?

<

I would mostly like to see CLARIN-IT introduce in its repositories an integrated online environment that would support the proof-reading, encoding and enrichment of classical texts. I would also like to see CLARIN-IT experts draft precise guidelines or propose a paradigmatic schema on how to provide metadata specific to digital classics, such as the physical description of the source (papyrus, epigraph, manuscript, etc.), information on the origin, history, publication, and so on. We also need to provide the concordances of different editions of the same text, with correspondences to the fragments that have a different identification number in each edition.
Moreover, I would like to see CLARIN-IT develop tools tailored to non-computational researchers that would help them perform linguistic and textual annotation (morpho-syntactic, semantic, etc.) without requiring them to possess a great deal of technical know-how. In addition, improving the performance of the existing parsers for Ancient Greek on fragmentary texts would offer very important upgrades for the study and teaching of Ancient Greek.

You have recently visited CLARIN-DK experts on the CLARIN Mobility Grant. How was your research visit beneficial for your work? What knowledge and expertise did you gain from CLARIN-DK experts?

<

At the Centre for Language Technology (Department of Nordic Studies and Linguistics, University of Copenhagen) I was given advice on how to better encode the different kinds of sources in which the digital classics are attested. Aside from the practical skills developed during my research stay, I found it inspiring to meet professors and researchers with different backgrounds and research objectives. Before my visit, I had mainly focused on the XML/TEI encoding, but my research stay allowed me to turn my attention to automatic linguistic analysis. I also gained a better understanding of existing tools and the potential of the CLARIN infrastructure as a network of not only language technologies, but also invaluable expertise.

>

Would you like to continue collaborating with CLARIN-IT after you finish your PhD? Do you have any wishes or plans already?

<

Yes, of course, I would like to continue working on the DEA project until the new edition of Archilochus' fragments is completed. I believe that further collaboration could improve my research methodology and allow me to gain better skills in applying linguistic analysis tools to my future research in Philological and Literary Studies.

Moreover, I think the greater availability of post-PhD research grants, also on a national level, would be useful, as it would support research and, at the same time, aid the development of the national consortium in specific fields of knowledge. In the CLARIN-IT repositories, there currently aren't many classical texts and resources, so I think that my philological knowledge in combination with my digital skills could make valuable contributions in this direction.

>

DENMARK

Introduction

Written by **Costanza Navarreta**

Denmark has been a member of CLARIN ERIC since February 2012 and is one of its founding members.[9] The Danish infrastructure CLARIN-DK was funded through two projects, the DK-CLARIN (2008-2010), and the DIGHUMLAB project (2011-2017). Since 2018, CLARIN-DK has been funded by the Faculty of Humanities and the Department of Nordic Studies and Linguistics, at the University of Copenhagen. The Danish national coordinator is Costanza Navarretta and the leading institution is the Centre for Language Technology, which is part of the Department of Nordic Studies and Linguistics.

CLARIN-DK involves the following institutions:

- The University of Copenhagen
- The Royal Danish Library

CLARIN-DK is a stable national research infrastructure where researchers can deposit, share and download language resources such as domain-specific corpora (e.g., The Danish Parliament Corpus 2009–2017 and the Johannes V. Jensen Corpus, which is a literary corpus collecting the works of the famous modernist poet Johannes Jensen from the early 20th century), as well as lexicons, word lists, speech transcriptions, and audio/video files in a secure way. CLARIN-DK also offers on-line language technology tools comprising e.g. a tokenizer, PoS tagger, a lemmatizer for Danish and English, a named entity recognizer for Danish, a keyword extractor, a TEI-to-text converter and a pipeline to linguistic annotation. Tools for performing basic frequency counts of words in textual data are also included, as well as visualization and corpus linguistics tools developed by other research groups, such as Korp and Voyant. Aside from being a certified B Centre, CLARIN-DK also runs a Knowledge Centre called DANSK, which provides expertise and help with using the language resources and technologies offered by the Danish consortium together with the Danish Language Council.

CLARIN-DK is involved in various Danish research projects and networks. For example, it is part of the Danish collaboration initiative DIGHUMLAB that involves various research communities, such as NetLAB, which is aimed at the cross-disciplinary study of internet materials, and LARM.fm, which is an online platform used for automatically locating the missing metadata of broadcast radio programmes. CLARIN-DK is also partner in an external funded research project Infrastrukturalisme with PI Henrik Jørgensen, of Aarhus University. The consortium is also involved in a research network, Multimodal Child Language Acquisition, with the University of Hong Kong and the Chinese Hong Kong University, (PI Costanza Navarretta), and contributes tools and guidance in a number of research activities comprising the linguistic annotation of medieval documents and TEI encoding of literary corpora, mainly at the University of Copenhagen. CLARIN-DK is also involved in research data management and the promotion of FAIR data in the Humanities.

The CLARIN-DK team participates in the following CLARIN committees: Standing Committee for CLARIN Technical Centres (Lene Offersgaard, Bart Jongejan), Legal and Ethical Issues Committee: Sussi Olsen, Assessment Committee (Lene Offersgaard as Chair).

The Clarin-DK group at the University of Copenhagen: Mitchell John Seaton, Costanza Navarretta, Dorte Haltrup Olsen, Bart Jongejan, Sussi Olsen and Lene Offersgaard

[9] https://clarin.dk/

Tool | **CST Lemmatizer**

Written by **Bart Jongejan** and **Costanza Navarretta**

Lemmatizers generalize over the different forms of a word used in free text and provide its lemma, which is the base or dictionary look-up form. They are therefore one of the basic NLP tools which are not only important for NLP, but also for lexicographic work and all text-based studies. They are especially indispensable in morphologically rich languages that have a large number of word forms for the same lemma, which severely hinders querying or processing all of them in running text.

The CST lemmatizer has been developed over many years and as part of various projects, especially the Danish STO (Jongejan and Haltrup 2005) and the Nordic Tvärsök (Jongejan and Dalianis 2009).[10] While it was initially used as a tool to support Danish lexicographic work, it has gradually been extended with a dynamic self-learning algorithm which learns new lemmatization rules from morphological lexica that contain the relations between word forms and their corresponding lemmas. The lemmatization rules are organized in a decision tree.

In comparison to other state-of-the-art stemmers and rule-based lemmatizers, the current version of the CST lemmatizer not only learns lemmatization rules from word endings, but also recognizes a wide variety of derivational patterns; e.g., prefixation, infixation, suffixation. Therefore, it can deal with languages with different morphological systems. Currently, the CST lemmatizer has been trained on 25 languages. The list of these language-trained versions of the CST lemmatizer available from the Center for Language Technology is shown in Figure 13.

The lemmatizer is available for download via GiTHub. Figure 14 shows the CLARIN-DK web service for the CST-lemmatizer, while Figure 15 shows a Danish example sentence that was lemmatized with the tool.

[10] https://cst.dk/online/lemmatiser/uk/

Figure 13: *The languages for which the trained CST-lemmatizer is available. Danish and English texts can be lemmatized online with the CST lemmatizer.*

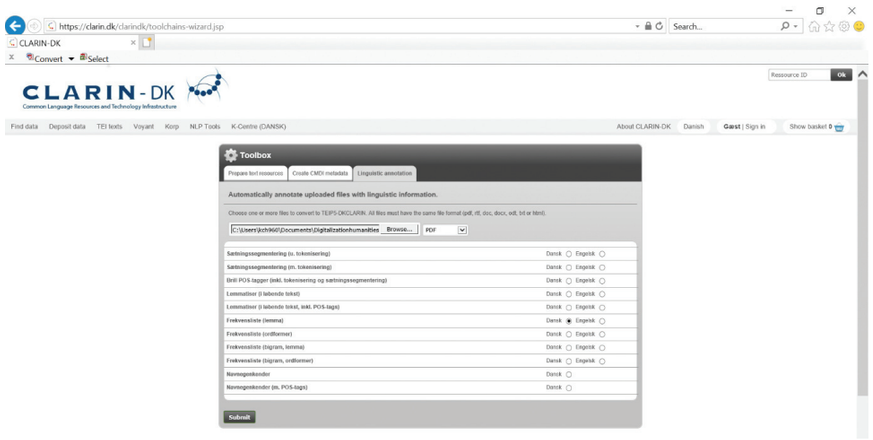


Figure 14: *The online CST lemmatizer on CLARIN-DK*

```
<sentence id ='8'>
Dog     dog
,       ,
året    år
der     der
er      være
gået    gå
,       ,
kan     kunne
også    også
have    have
budt    byde
på      på
tunge   tung
stunder stund
–       –
ikke    ikke
alt     al
er      være
glæde   glæde
for     for
os      vi
alle    al
.       .
</sentence>
```

Figure 15: *Lemmatization of the Danish sentence* Dog, året der er gået, kan også have budt på tunge stunder – ikke alt er glæde for os alle *("However, the past year can also have provided sad moments – not everything can give happiness to all of us "), which is taken from the 2017 New Year's Eve speech by the Danish Queen*

The CST lemmatizer trained for Danish has been used in many NLP projects, but also outside the NLP community. Frederik Hjorth, who is a political science researcher at the Department of Political Science, at the University of Copenhagen, has applied the CST lemmatizer to political speeches as one of the preprocessing steps in order to investigate how members of the existing political parties have addressed right-wing populists who have been challenging the order of the established political system (Hjorth 2018). The results of the study indicate that young politicians are often willing to engage with the populists as well as with other politicians across the political spectrum in the name of democratic freedom (which Hjorth calls the *strategy of engagement*), while older politicians often describe the populist challengers as morally illegitimate (which Hjorth calls the strategy of *disparagement*) and refuse to enter into discussions with them.

The CST lemmatizer was also used for many other languages in different linguistic projects. For example, it was trained on Russian (Sharoff and Nivre 2011) and then used e.g. for event identification (Solovyev and Ivanov 2016), and for anaphora and co-reference resolution (Toldova et al. 2014).

References:

Jongejan, B. and Haltrup, D. 2005. The CST *Lemmatiser*. Center for Sprogteknologi, University of Copenhagen version 2.7. http://cst.dk/online/lemmatiser/cstlemma.pdf.

Jongejan, B. and Dalianis, H. 2009. Automatic Training of Lemmatization Rules That Handle Morphological Changes in Pre-, in- and Suffixes Alike. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1* - ACL-IJCNLP '09. Vol. 1 Suntec, Singapore: Association for Computational Linguistics, 145.

Hjorth, F. 2018. Establishment Responses to Populist Challenges: Evidence from Legislative Speech. 2018 *Annual Meeting of the Danish Political Science Association*. http://fghjorth.github.io/papers/responses.pdf.

Sharoff, S. and Nivre, J. 2011. The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge. In *Proceedings of Computational Linguistics and Intelligent Technologies DIALOGUE2011*, Bekasovo, 591–604. https://pdfs.semanticscholar.org/36df/5fbe04f425e9b089437e979581d1f5375a94.pdf.

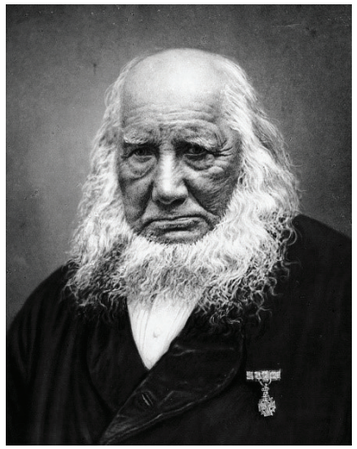
Solovyev, V. and Ivanov, V. 2016. Knowledge-driven event extraction in Russian: corpus-based linguistic resources. *Computational Intelligence and Neuroscience*. https://doi.org/10.1155%2f2016%2f4183760.

Toldova, S. et al. 2014. RU-EVAL-2014: Evaluating Anaphora and Coreference Resolution for Russian. *Computational Linguistics and Intellectual Technologies* 13 (20): 681–694.

Resource | **Grundtvig's Work Corpus**

Written by **Dorte Haltrup Hansen** and **Costanza Navarretta**

Nikolai Frederik Severin Grundtvig was a theologian, a priest, a philosopher, a poet, a writer, a teacher and a politician (member of the Rigsdagen, one of the two parts of the Parliament), who lived in Denmark between 1783 and 1872. He was a contemporary of Hans Christian Andersen and Søren Kierkegaard. Grundtvig's ideas have had a lasting impact on many areas of Danish culture such as education, politics and the church. For example, Grundtvig advocated for a reform of the school system, which also included educating adults to participate actively in society and cultural life. Therefore, Grundtvig is considered to be the mind behind the folk high school.
He was part of the national romantic movement, and contributed to the development of the Danish national awareness. Grundtvig's written works are thus an important key to the understanding of Danish culture and mentality.



Nikolai Frederik Severin Grundtvig

The collection Grundtvig's Works are published by the Grundtvig Centre at the University of Aarhus and will contain 1,000 critically annotated texts written by N.F.S. Grundtvig when finalized in 2030.[11] The works are available to the public through a searchable interface, including registers of persons, places and Bible citations.
The researchers at the Grundtvig Center wanted to a reliable and consistent way to cite the publication and a sustainable and interoperable environment in which they could share the work among other scholars and the public in general.
Since the Grundtvig Centre itself does not offer the possibility for downloading the underlying files, CLARIN-DK was approached as a repository provider.

[11] http://www.grundtvigsvaerker.dk/

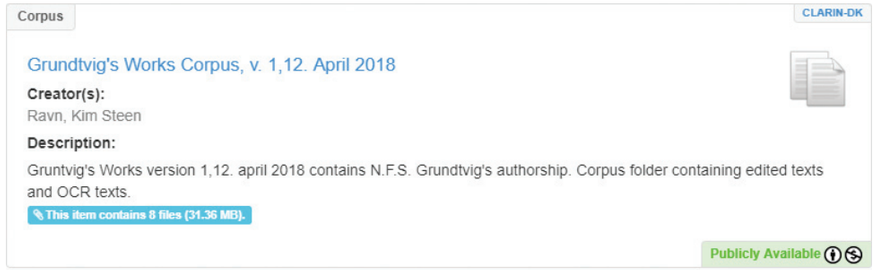
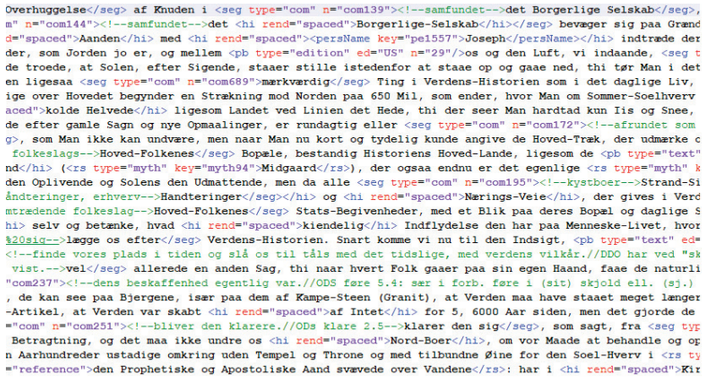


Figure 16: *The corpus in the CLARIN-DK repository*

The corpus, now deposited in CLARIN-DK's D-Space repository[12], consists of around 1,300 TEI encoded XML-files, of which approximately 450 are critical editions manually annotated with person names, place names, mythological names, Bible citations and comments. When new versions of the works are released, they will be uploaded as new versions of the corpus in the CLARIN-DK repository.



```
Overhuggelse</seg> af Knuden i <seg type="com" n="com139"><!--samfundet-->det Borgerlige Selskab</seg>,
m" n="com144"><!--samfundet-->det <hi rend="spaced">Borgerlige-Selskab</hi></seg> bevæger sig paa Grænd
d="spaced">Aanden</hi> med <hi rend="spaced"><persName key="pe1557">Joseph</persName></hi> indtræde der
der, som Jorden jo er, og mellem <pb type="edition" ed="US" n="29"/>os og den Luft, vi indaande, <seg t
de troede, at Solen, efter Sigende, staaer stille istedenfor at staae op og gaae ned, thi tør Man i det
en ligesaa <seg type="com" n="com689">mærkværdig</seg> Ting i Verdens-Historien som i det daglige Liv,
ige over Hovedet begynder en Strækning mod Norden paa 650 Mil, som ender, hvor Man om Sommer-Soelhverv
aced">kolde Helvede</hi> ligesom Landet ved Linien det Hede, thi der seer Man hardtad kun Iis og Snee,
de efter gamle Sagn og nye Opmaalinger, er rundagtig eller <seg type="com" n="com172"><!--afrundet som
g>, som Man ikke kan undvære, men naar Man nu kort og tydelig kunde angive de Hoved-Træk, der udmærke c
 folkeslags-->Hoved-Folkenes</seg> Bopæle, bestandig Historiens Hoved-Lande, ligesom de <pb type="text"
nd</hi> (<rs type="myth" key="myth94">Midgaard</rs>), der ogsaa endnu er det egenlige <rs type="myth" k
den Oplivende og Solens den Udmattende, men da alle <seg type="com" n="com195"><!--kystboer-->Strand-Si
åndteringer, erhverv-->Handteringer</seg></hi> og <hi rend="spaced">Nærings-Veie</hi>, der gives i Verd
mtrædende folkeslag-->Hoved-Folkenes</seg> Stats-Begivenheder, med et Blik paa deres Bopæl og daglige S
hi> selv og betænke, hvad <hi rend="spaced">kiendelig</hi> Indflydelse den har paa Menneske-Livet, hvor
%20sig-->lægge os efter</seg> Verdens-Historien. Snart komme vi nu til den Indsigt, <pb type="text" ed=
<!--finde vores plads i tiden og slå os til tåls med det tidslige, med verdens vilkår.//DDO har ved "sk
 vist.-->vel</seg> allerede en anden Sag, thi naar hvert Folk gaaer paa sin egen Haand, faae de naturli
"com237"><!--dens beskaffenhed egentlig var.//ODS føre 5.4: sær i forb. føre i (sit) skjold ell. (sj.)
, de kan see paa Bjergene, især paa dem af Kampe-Steen (Granit), at Verden maa have staaet meget længer
-Artikel, at Verden var skabt <hi rend="spaced">af Intet</hi> for 5, 6000 Aar siden, men det gjorde de
="com" n="com251"><!--bliver den klarere.//ODs klare 2.5-->klarer den sig</seg>, som sagt, fra <seg typ
 Betragtning, og det maa ikke undre os <hi rend="spaced">Nord-Boer</hi>, om vor Maade at behandle og op
n Aarhundreder ustadige omkring uden Tempel og Throne og med tilbundne Øine for den Soel-Hverv i <rs ty
="reference">den Prophetiske og Apostoliske Aand svævede over Vandene</rs>: har i <hi rend="spaced">Kir
```

Figure 17: *A look into* Haandbog i Verdens-Historien (Handbook in World History) *from 1833*

The language excerpt in Figure 17 shows the old orthography from before the Danish language revision in 1948, e.g.:

Original	… som Man i det attende Aarhundrede troede, at Solen, efter Sigende, staaer stille istedenfor at staae op …
Normalized Danish	som man troede i det 18. århundrede, at solen efter sigende står stille i stedet for at stå op
Literal English translation	… as thought in the 18th century, that the sun after what they said, is staying still instead of rising …

[12] http://hdl.handle.net/20.500.12115/31

Furthermore, the excerpt shows the manual mark-up of the corpus, done by philologists at the Grundtvig Centre. There are references to, for example, person names (Joseph), mythological places (Midgaard) and actual places (Europe) and comments on parts of the text (*Overhuggelse af Knuden* <com139>, literal English translation: the cut of the knot). The actual comment is not shown in the text.

The corpus is an excellent resource for researchers who wish to apply digital methods to investigate various aspects of Grundtvig and his epoch. For example, researchers might want to investigate Grundtvig as a historical person, address the 19th century's literary language or orthography, or dig into his work when studying the theoretical background of the Danish folk high school tradition. The corpus is also important for scholars applying Linked Data in order to investigate the 19th century, since the corpus contains the annotations of people, places and events.

Event | **Teach the Teachers – the Voyant Tools**

Written by **Lene Offersgaard** and **Dorte Haltrup Hansen**

Digital methods are only slowly gaining ground in the teaching of literary studies in Denmark. While many lecturers are interested in introducing digital methods to their students, they often lack the knowledge of existing tools. From previous workshops, CLARIN-DK learned that neither traditional NLP tools like lemmatizers, POS-taggers, and named entity recognizers, nor simple command line scripting, were suitable in such teaching scenarios. This is why CLARIN-DK started to explore other technologies, such as data visualization tools that could serve as a better and easier entry point to the use of digital methodologies for non-computational researchers and teachers.

We opted for Voyant Tools,[13] introduced to us by information specialists from HUMlab – a datalab at the Copenhagen University Library. Voyant Tools is an online environment that performs automatic text analysis with functionalities such as word frequency lists, frequency distribution plots, and KWIC displays (Figure 18). CLARIN-DK and HUMlab have organized several interactive workshops presenting the use of this environment to lecturers and researchers at the Faculty of Humanities at the University of Copenhagen. CLARIN-DK hosted a dedicated event at the Department of Nordic Studies and Linguistics on 21 November 2018, which was attended by 12 teachers and researchers.

Figure 18: *Voyant Tools*

[13] https://alf.hum.ku.dk/VoyantTools/

In order to tailor the events to the needs of the participants, CLARIN-DK asked some of them in advance which literary works were most relevant to be showcased and which research questions could be investigated and discussed during the events. They opted for novels written around the Modern Breakthrough period, an era in the Scandinavian literature which started at the end of the 19th century and in which Naturalism replaced Romanticism. The Archive of Danish Literature (http://adl.dk) provided a collection of 54 novels. The novels were preprocessed and uploaded to a local instance of Voyant Tools by the CLARIN-DK team and information specialists from HUMlab.

A research question addressed the use of terms before and after the Modern Breakthrough (1870–1890). If it was possible to visualize changes in the use of, for example, terms for emotions (like *love*) which are typical for the Romanticism period compared to the use of more concrete terms (like *work*) which should be more common in the Naturalism novels. Using the *Trends* tool in Voyant (Figure 19), it was found that the term for love is used relatively more often before 1875 than after 1888. Moreover, the term for *work* is not used before 1875 in the novels, while it was used after then. Therefore, the use of these terms indicates that there is a shift in the use of common themes around the Modern Breakthrough. However, by using this simplistic method, it is impossible to differentiate novels representing the Modern Breakthrough.

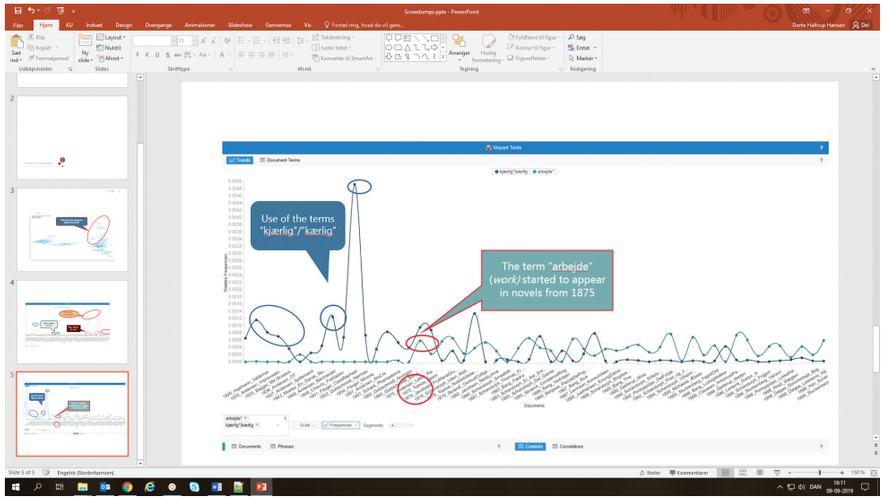


Figure 19: *The chronological distribution of the terms love vs. work for the period between 1826 and 1899 with regard to 54 novels*

We therefore investigated if other tools in Voyant could also confirm the differences between the two literary periods. In the *ScatterPlot tool* it is, among other things, possible to visualize the results of document similarity analysis. Figure 20 shows the document similarity using the TF-IDF frequency count for all novels in the corpus. In the figure, the novels by Herman Bang and a few novels by Sophus Schandorph are clearly separated from the other works. The novels from the late 19th century of these two writers are considered representatives of the Modern Breakthrough. It was now up to the researchers to interpret the similarities in the other groups of the scatter plot and from there to pose more research questions.

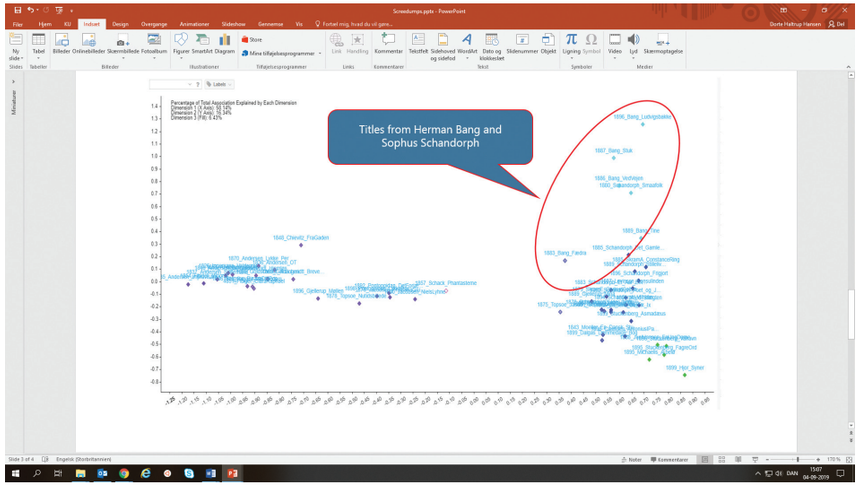


Figure 20: *Novel similarity based on TF-IDF counts*

In this and other workshops, the participants soon realized that studying texts through isolated words (word forms) was limiting, and there was a clear need for lemmatization. Moreover, the need for PoS-tagged texts became evident since some researchers were interested in investigating adjectives showing emotions, while others were interested in analysing events, requiring the automatic extraction of verbs. Despite this, Voyant Tools has proved to be very illustrative and useful to get a first quantitative overview of a collection of novels, and it allowed the comparison of two or more novels.

As a follow up to this event, the CLARIN-DK team will organize a workshop introducing corpus tools and corpus querying techniques in linguistically annotated texts for Literary Studies. The event will also showcase how automatic linguistic annotations are performed on texts from before and after the Danish orthographic reform of 1948, and discuss how it is possible to circumvent problems encountered when applying NLP tools developed for contemporary texts to older texts.

Interview | **Klaus Nielsen**

Klaus Nielsen is the chief editor at the Grundtvig Study Centre.

What is your scholarly background and your current academic position?

<

I obtained my PhD from the University of Copenhagen in 2012 and my thesis was a combination of traditional literary theory and book history, a philological field that focuses on a more mechanical-analytical study of the publication process of literary works. I focused on *Gittes monologer*, a famous collection of satirical poems by the Danish poet Per Højholt published in different versions between 1980 and 1984. I was able to observe crucial textual differences between their various published versions, which allowed me to arrive at a much richer interpretation of the poems that wouldn't be possible with the final, best-known 1984 version alone. This showed me how important it is to combine traditional qualitative literary analysis with analytical methods that also take into consideration non-textual information such as publication history.

I now work as chief editor at Grundtvig Study Centre,[14] where we are preparing a critical edition of the collected works of N.F.S. Grundtvig, a very prolific and multidisciplinary Danish author who published around 37,000 pages of text from 1804 to his death in 1872. We are making this corpus available in an online environment, with manual annotations that follow the scholarly standards of textual criticism. In a sense, my PhD was an important methodological stepping stone for my current work related to the Grundtvig's Works Corpus, which also involves a close study of the differences between the various published editions.

>

[14] http://www.grundtvigsvaerker.dk/

The Grundtvig's work corpus has been published through the CLARIN-DK repository. How did this collaboration start? How do you benefit from this collaboration?

<

We released the first version of our corpus through the CLARIN-DK repository in 2018 at the suggestion of Lene Offersgaard, with whom we were collaborating on a related project at the time. This was a great opportunity for us because we had been receiving feedback from some of our more devoted users who said they wanted the corpus in a downloadable format. We've also made an agreement with CLARIN-DK that as soon as we publish a new version of the corpus through our online environment, we'll also update the version deposited in the repository with the newest, more richly annotated one.

>

How is Grundtvig's corpus structured? What are some of the challenges you come across when annotating the corpus?

<

The corpus is extremely varied in terms of content, since Grundtvig was a polihistorian who wrote on a variety of different subjects. Perhaps most prominently, he wrote books on Danish history and Nordic mythology, carried out linguistic studies of Old Icelandic and Old English, translated from Latin, wrote political and philosophical texts, and composed around 1,500 hymns, many of which are still sung today in Denmark. For this reason, Grundtvig's views are representative of the intellectual and cultural zeitgeist of Denmark in the 19th century.

There's a downside to his varied repertoire, in that annotation is still manually intensive. We do use a database for place and person names that we feed into a named-entity recognizer, but even in this case, we often have to manually verify the results. For example, Grundtvig often refers to the philosopher Søren Kierkegaard, who was a contemporary of his, and our software is generally successful in identifying this particular named entity. However, Grundtvig often refers to him by his last name only, but since Søren Kierkegaard had a brother who was also a published author in the same period, we have to manually check the automatic recognition to make sure that the software made a link to the correct referent. In addition to this, we often come across obsolete words, in which case we manually add their possible historical meaning. This can only be done by closely reading and interpreting the surrounding text. Nevertheless, we will use the parts that have already been annotated as a baseline for a semi-automated processing of the remaining two-thirds of the corpus in the future.

One of the greatest challenges in terms of mark-up pertains to identifying Biblical references, especially in cases where Grundtvig doesn't use direct quotes taken from the Bible but his own modified variants, or where he makes indirect references to the more obscure motifs and quotes. Although we have theologians both internal and external who closely read the texts and manually identify such references, it would be invaluable if we could also make use of a language tool that would help automatize this process of identification. I don't think that such a tool exists yet, but it would be a very welcome addition to the CLARIN infrastructure in my opinion. Similarly, it would be great to have a tool that can automatically recognize proverbs and sayings, which abound in Grundtvig's works, given that his work is a major part of the Danish cultural heritage. Although I'm not an expert in digital technologies, it seems that developing such a tool wouldn't be too hard a task, as there already exist readymade digital collections of Danish proverbs that could be used as a baseline for training the tool.

>

Has the corpus been successfully used by an external research project?

<

Yes, Baunvig and Nielbo (2017)[15] have used our corpus in a case study to determine how digital methods can benefit the analysis of very large collections of written text, and to uncover new perspectives and interpretations. Grundtvig Studies is a popular subfield in literary history in Denmark, and many studies on Grundtvig have been published in the past fifty years. However, previous researchers weren't able to use digital methods and tools, which means that their claims were influenced by the limitations inherent to a purely manual approach to analysis. As I've said, Grundtvig produced around 37,000 pages in his lifetime, which is simply too much text for an individual researcher to read and then be able to recollect the finer details. For instance, there is an older study in which it is claimed that Grundtvig started suffering from a series of psychological problems in the 1830s, which was reflected in the texts he wrote in this decade. However, Baunvig and Nielbo (2017) were able to show, by using quantitative methods such as measuring the amount of information entropy in the corpus, that his psychological turmoil actually started earlier than was previously claimed, which is of course an important finding from a purely historical viewpoint. There has also been a follow-up study of our corpus conducted by Nielbo et al. (2018).[16]

>

[15] https://knielbo.github.io/files/valider_selvopgoer_kln.pdf

[16] https://doi.org/10.1093/llc/fqy054

What makes this corpus particularly valuable for the CLARIN infrastructure?

<

I think that our rather thorough manual approach to the corpus is an important contribution for a more accurate understanding of the historical developments of the Danish language, especially its orthography. What is important in this respect is that there were no orthographic rules in Grundtvig's time, only tendencies, which means that spelling was quite liberal in comparison to contemporary Danish. Consequently, we're often in doubt whether the way Grundtvig spelled a certain word is an instance of spelling variation that was attested at the time or if it is just a spelling mistake on his part. This is particularly problematic in cases where Grundtvig's idiosyncratic spelling can't be found in the historical dictionaries of 19th century Danish, since this intuitively makes you think that the spelling variant was a mistake. However, such dictionaries weren't compiled on the basis of the original edition but often used later published editions that had gone through the editing process, where spelling variation was normalized. This means that if a researcher wanted to study the vocabulary of 19th century Danish just on the basis of such dictionaries, he or she would miss the attested variations and consequently get a warped view of how people actually wrote at the time. By contrast, we spend a lot of time closely analysing and proofreading the materials, so we are able to present a resource that serves as a much more complex, as well as accurate, presentation of the linguistic situation at the time.

>

Could you give an example of such orthographic variation? How did you resolve it?

<

I actually came across a fairly interesting orthographic problem just recently when I was annotating Grundtvig's *History of the Northmen*, which is one of the few texts he wrote in English. In this text, Grundtvig used the word kempion in the sense of "champion" or "hero"; however, this spelling variant isn't listed in the *Oxford English Dictionary*, which only includes the variant campion with an a instead of an e. Because my colleagues and I weren't sure how to solve this issue, we consulted a Professor of Middle English, and he believed it to be a spelling mistake that should be corrected in the edited corpus, given that the Oxford *English Dictionary* is extremely comprehensive and thorough in its account of English etymology. However, when I searched for the variant kempion on Google, I found out that it was actually attested at the time, and it was for instance used by Sir Walter Scott in his 1822 novel *The Pirate*, which Grundtvig was alluding to.

>

Are there any other aspects of the CLARIN-DK infrastructure that are important for your work at the centre?

<

Yes, especially in relation to how proactively they reach out as part of their user-involvement initiative. Last year, CLARIN-DK organized a tutorial for the philologists at our centre where they demonstrated how Voyant Tools can simplify our annotation process. Using Voyant has turned out to be extremely helpful when we come across obsolete phrases the meaning of which we don't know and can't find in the historical dictionaries. By using Voyant's extended search capabilities and visualization tools, we are now able to easily chart the occurrences of this unknown phrase in the entire corpus, and then extract only those texts where this phrase seems to occur in a similar context, which then helps us determine its actual meaning.

I am also pleased to say that CLARIN-DK has already made the first version of our corpus available through their installation of the Voyant Tools. We plan on updating this test version with newer ones with regularity. In the long run, I believe the availability of the corpus through CLARIN-DK's Voyant Tools will significantly streamline user assistance.

>

Your professional website says that you're also interested in audio literature. Is this something that you're still actively researching?

<

No, my research on audio literature was mostly confined to my PhD project, because Per Højholt, who is the author of the poems that I was analysing, had read them aloud on Danish radio in the 1980s. By using an audio-analysis software called PRAAT I measured prosodic features such as the author's pitch and reading speed, and I was able to see how he deliberately changed his voice in accordance with the way the point-of-view character developed through the course of the poems' narrative. This was a rather small but important finding, since it hadn't been previously acknowledged in the relevant literature on Gittes Monologer how the author's spoken performance of his own work added new dimensions to the understanding of the poems themselves.

>

What kind of new research questions does audio literature offer in the context of Digital Humanities? Do you think that CLARIN could contribute to this field?

<

When I was writing my thesis, research on audio literature was still a very new field, but nowadays it is more readily agreed upon that audio recordings can serve as crucial material for textual analysis. Literary theorists are now conducting important research on the link between the reader of the audio text and the content of the text itself, and this opens up many interesting questions. Let's say, for instance, that we are dealing with a novel written in the first person, and that the narrator is a woman. Should the reader of the audio version then also be a woman, or conversely, what interpretative repercussions would arise if the reader were actually a man? That is, the person's voice crucially affects the way people perceive the text, much in the same way that the sort of typography of an old book can evoke various pre-conceptions in the reader about the book's content.

Given how audio literature opens up interesting questions relevant for the emerging Digital Humanities, I think that new digital tools for analysing recorded literary works would serve as very welcomes additions to the CLARIN infrastructure.

>

What are your hopes for CLARIN-DK in the future?

<

I think that one of the future challenges for Digital Humanities in Denmark is to find a common platform where our whole research community can have a more unified and interoperable access to as many carefully annotated resources as possible. I believe that CLARIN-DK is an excellent candidate in the country for this, because our experience with releasing the Grundtvig's Work Corpus has proven to us that their repository is a stable environment through which corpora can be released in a sustainable fashion and with well-presented metadata. On top of that, the repository also allows us to integrate our corpora with other services in the consortium. For this reason, it can only be a good thing if more Digital Humanities scholars in Denmark decide to deposit their resources in the CLARIN-DK repository.

>

SLOVENIA

Introduction

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

CLARIN.SI joined CLARIN ERIC in 2015.[17] The seat of CLARIN.SI is the Jožef Stefan Institute, the main Slovenian research organization for applied research in natural sciences and technology, where three units are involved in overseeing its operation: the Department of Knowledge Technologies, the Laboratory for Artificial Intelligence, and the Centre for Networking Infrastructure. CLARIN.SI is organized as a consortium, bringing together partners from all the main organizations that produce or use language resources in Slovenia, in particular the four Slovenian universities (University of Ljubljana, University of Maribor, University of Nova Gorica and University of Primorska), three research institutes (Scientific and Research Centre of the Slovenian Academy of Sciences and Arts, Jožef Stefan Institute, Institute of Contemporary History), three societies (Slovenian Language Technologies Society, Trojina Institute for Applied Slovene Studies, Domestic Research Society), and two HLT companies (Alpineon and Amebis). The national coordinator of CLARIN.SI is Tomaž Erjavec.

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

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

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

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

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

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

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

Tool | **CSMTiser**

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

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

CSMTiser, available on the CLARIN.SI GitHub site,[18] is a supervised machine learning tool that performs word normalization by using Character-level Statistical Machine Translation. The tool is a wrapper around the well-known Moses SMT package, which enables non-computer-scientists to train and run a text normalizer by editing the configuration file, running a script for training the normalizer, and then another one for applying it.

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

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

Figure 21: *Slovene text from three different periods. The column in bold shows the slice the text belongs to and, in brackets, its year of publication. Each example gives the original text in the first line and the modernized work tokens in the second line, to illustrate the kind of phenomena that must be handles in the modernization of words.*

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

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

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

References:

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

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

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

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

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

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

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

Resource | **Emoji Sentiment Ranking 1.0**

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

Emoji are Unicode graphic symbols, used as a shorthand to express concepts and ideas, and can play an important role in social media text analytics. In 2015, Petra Kralj Novak, Jasmina Smailović, Borut Sluban and Igor Mozetič from the Jožef Stefan Institute in Ljubljana, Slovenia released the first emoji sentiment lexicon, called Emoji Sentiment Ranking 1.0, and published it as a resource in the public language resource repository CLARIN.SI.[19] With 78,500 downloads to date, the lexicon is the most downloaded resource in the CLARIN.SI repository.

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

The process and analysis of the Emoji Sentiment Ranking is described in detail by Kralj Novak et al. (2015). The authors draw a sentiment map of the 751 emoji (see Figure 22), formalize sentiment and present a novel intuitive visualization of sentiment distribution in the form of a sentiment bar (Figure 23). Furthermore, they compare the sentiment of tweets with and without emoji, and find that tweets with emoji tend to be more positive. They also found differences between the more and less frequent emoji: the more frequently used emoji tend to be more positive. Another interesting aspect is the position of emoji in tweets: the more sentimental charge an emoji has, the more likely it is to appear at the end of tweets (see Figure 24). An exception is the soccer ball emoji, which is commonly used to replace a word but has a very positive sentiment associated with it.

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

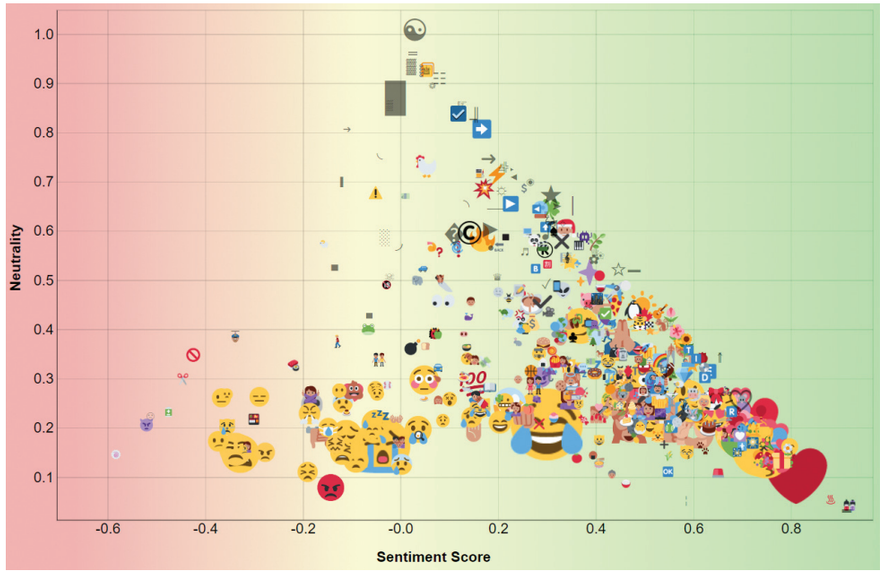


Figure 22: *Sentiment map of the 751 most frequently used emoji. The position of the emoji denotes its sentiment score and neutrality, while its size is proportional to the frequency of its usage. An interactive version is available here: http://kt.ijs.si/data/Emoji_sentiment_ranking/emojimap.html*

Image [twemoji]	Unicode codepoint	Occurrences [5...max]	Position [0...1]	Neg [0...1]	Neut [0...1]	Pos [0...1]	Sentiment score [-1...+1]	Sentiment bar (c.i. 95%)	Unicode name
😂	0x1f602	14622	0.805	0.247	0.285	0.468	0.221		FACE WITH TEARS OF JOY
❤	0x2764	8050	0.747	0.044	0.166	0.790	0.746		HEAVY BLACK HEART
❤	0x2665	7144	0.754	0.035	0.272	0.693	0.657		BLACK HEART SUIT
😍	0x1f60d	6359	0.765	0.052	0.219	0.729	0.678		SMILING FACE WITH HEART-SHAPED EYES
😭	0x1f62d	5526	0.803	0.436	0.220	0.343	-0.093		LOUDLY CRYING FACE
😘	0x1f618	3648	0.854	0.053	0.193	0.754	0.701		FACE THROWING A KISS
😊	0x1f60a	3186	0.813	0.060	0.237	0.704	0.644		SMILING FACE WITH SMILING EYES
👌	0x1f44c	2925	0.805	0.094	0.249	0.657	0.563		OK HAND SIGN
💕	0x1f495	2400	0.766	0.042	0.285	0.674	0.632		TWO HEARTS
👏	0x1f44f	2336	0.787	0.104	0.271	0.624	0.520		CLAPPING HANDS SIGN
😁	0x1f601	2189	0.796	0.127	0.296	0.577	0.449		GRINNING FACE WITH SMILING EYES
☺	0x263a	2062	0.799	0.062	0.218	0.720	0.657		WHITE SMILING FACE

Figure 23: *The sentiment distribution of each emoji is visualized in form of a sentiment bar. http://kt.ijs.si/data/Emoji_sentiment_ranking/index.html*

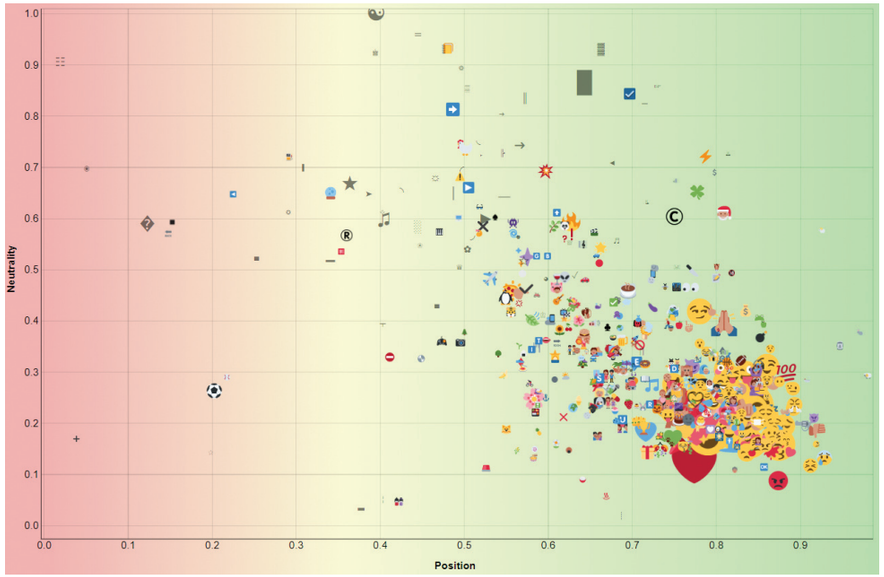


Figure 24: *Emoji position in tweets. The horizontal axis represents the length of a tweet. The vertical axis represents the neutrality of the emoji: top for very neutral and bottom for very emotional, either positive or negative. Emoji that act as word replacements, thus positioned in the middle of the tweets, tend to have a neutral sentiment. The emoji that act as sentiment conveyers are more likely positioned at the end of tweets.*

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

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

Reference:

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

Event | **JANES Express**

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

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

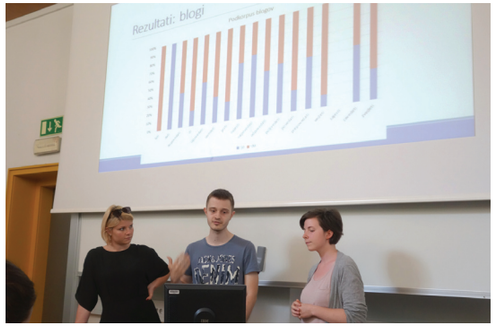
This was achieved in the scope of the Slovene basic research project JANES, which compiled a large and representative corpus that covered a large portion of publicly available user-generated text in Slovene, in particular tweets, blogs, forums posts, news comments and Wikipedia talk pages (Fišer, Ljubešić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

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

Particularly notable was the JANES Express seminar series for fellow researchers in corpus and computational linguistics which have been organized in Ljubljana (Slovenia), Zagreb (Croatia) and Belgrade (Serbia).[20] It was organized in collaboration with the Regional Linguistic Data Initiative. The seminar series presented the guidelines for manual annotation of training corpora of non-standard language varieties and the annotation platform WebAnno. As a result, three comparable gold standard corpora for tagging, lemmatization and normalization of non-standard Slovene, Croatian and Serbian have been developed and are available from the CLARIN.SI repository. Apart from testing and adapting the methodology originally developed for Slovene to two additional closely related languages, the JANES Express seminar series is also a best-practice example of knowledge transfer in the region.

More information about all these events as well as teaching materials are available on the project website.[21]

Summer camp on Slovene Netspeak



Summer school on Internet Linguistics

JANES Express seminar in Zagreb

JANES Express seminar in Belgrade

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

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

Interview | **Kaja Dobrovoljc**

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

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

<

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

>

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

<

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

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

>

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

<

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

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

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

>

You've also been part of the team that created the manually annotated ssj500k corpus.[22] Could you describe your role in its creation and annotation? Why is this corpus important for Slovenian linguistics?

<

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

>

Together with Joakim Nivre you have worked on annotating the Treebank of Spoken Slovenian[23] following the Universal Dependencies framework. What are the benefits of the Universal Dependencies framework and why is it important for Slovene to be part of the initiative? What are the challenges of creating a treebank of spoken language data? Why is it important for Slovene linguists and the society at large to have access to a treebank of the spoken language?

<

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

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

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

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

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

>

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

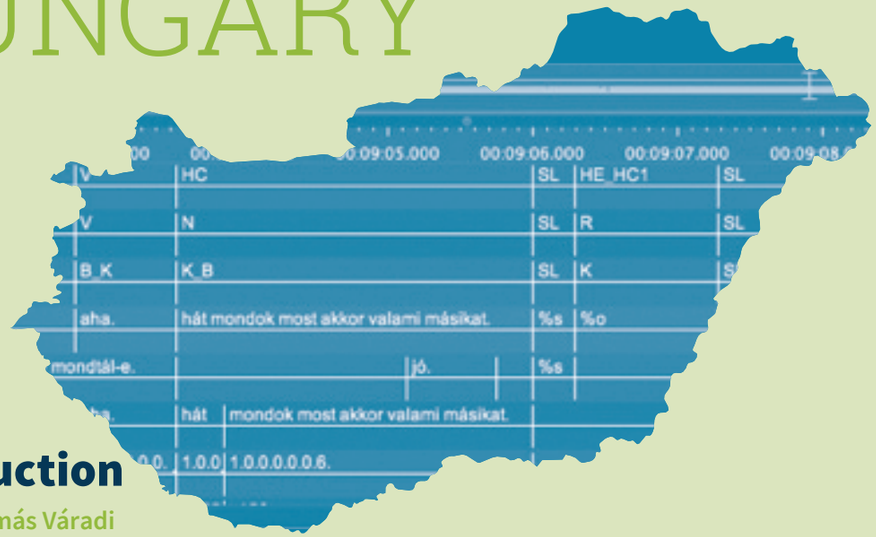
<

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

>

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

HUNGARY



Introduction

Written by **Tamás Váradi**

The national CLARIN consortium for Hungary, HUN-CLARIN, joined CLARIN-ERIC in 2016.[24] The Research Institute for Linguistics was one of the founding partners of CLARIN and took an active role in the preparatory phase of the history of CLARIN. The members of the consortium are the Research Institute for Linguistics, the MOKK Centre for Media Research and Education and the Speech Communication and Smart Interactions Laboratories of the Budapest University of Technology and Economics, the University of Szeged, the University of Debrecen, the Pázmány Péter Catholic University, the MorphoLogic LLC, the Institute for Computer Science and Control, and the Institute of Cognitive Neuroscience and Psychology. The national coordinator for HUN-CLARIN is Tamás Váradi.

As can be seen from the above list, the consortium covers a wide range of complementary expertise and research interests. It represents most of the leading research centres in Hungarian language and speech technology, which have closely cooperated in various national and international projects for more than a decade.

The resources developed by HUN-CLARIN members include corpora that are indispensable to research in the use of the Hungarian language, such as the Hungarian National Corpus, which has recently been upscaled to giga size, the Hungarian WebCorpus, which was the first of its kind in Hungarian, and the Szeged Treebank, the reference treebank for Hungarian. Bilingual resources include the Hunglish Corpus, a sentence-aligned Hungarian-English parallel corpus of about 120 million words in four million sentence pairs. A truly unique resource is HuComTech Corpus, a large scale multimodal corpus which offers a rich dataset on 47 annotation levels and was presented to the CLARIN community at the CLARIN 2018 Conference.

As regards tools, the Hun* set of tools developed by the MOKK Centre (such as HunAlign, HunTag, HunMorph, etc.) has also acquired recognition beyond Hungary for its versatility and free availability for languages other than Hungarian. A major recent achievement is the comprehensive processing chain e-magyar, which was developed through widespread collaboration within HUN-CLARIN members. This open and modular toolset was developed to suit the needs of Digital Humanities researchers and application developers alike, and is therefore available both as a web service and for download from GitHub repositories.

Severely limited by lack of funding for national activities, HUN-CLARIN, nevertheless, is making successful efforts to reach out to the Humanities and Social Science communities. It has established cooperation with the Centre for Digital Humanities at Eötvös Loránd University as well as the Centre for Social Sciences. Last year HUN-CLARIN embarked on a roadshow among Hungarian universities showcasing the central HUN-CLARIN tools and resources as well as local research projects. The three events so far at the universities of Szeged, Debrecen and Pécs have proved so popular that a second event is already being organized this autumn at Szeged University, at the institution's request.

In 2017 HUN-CLARIN hosted the CLARIN Annual Conference in Budapest. In the future, HUN-CLARIN plans to establish a K-Centre for Hungarian, continue with our outreach efforts and, subject to securing some national funding, set-up and operate a B-Centre as well.

The HUN-CLARIN team

[24] https://clarin.hu/en

Tool | **e-magyar: a Comprehensive Processing Chain for Hungarian**

Written by **Balázs Indig** and **Tamás Váradi**

The e-magyar toolchain was developed in 2016 as a major collaborative effort across the Hungarian NLP community.[25] The rationale for it was based on a clear vision of an open, modular, extendable and easy-to-use pipeline for Hungarian, which was suitable for non-specialists and developers alike. There existed pipelines created especially for Hungarian (e.g. the Hun* tools or Magyarlánc), and state-of-the art pipelines (e.g. StanfordNLP and UDPipe) also support Hungarian. However, they cannot fulfil the desired functions of modularity, extendibility and user-friendliness. For example, improving the existing methods and annotations on different levels of processing was extremely tedious, which prematurely cut short almost every attempt at natural improvement.

Therefore, the development of e-magyar started by collecting and integrating the best-practices and good features of the existing modules and pipelines while implementing the features that the community missed the most. The first version was integrated into the GATE framework.

The toolchain consists of the following tools (see Figure 25 for the general architecture):

- emToken, a rule-based tokenizer which adds Unicode handling and detokenization to its ancestor Huntoken;
- emMorph, a rule-based morphological analyser based on Helsinki Finite State Transducer, the flagship tool within e-magyar which integrates all previous efforts (including the commercial tool HUMOR) into a new, open-source tool for Hungarian;
- emPOS, a statistical PoS-tagger derived from HunPOS which is an improved version of the TnT tagger;
- emDEP, a dependency parser and emCONS a constituent parser taken directly from Magyarlánc;
- emNER, a named entity recognizer based on the HUNtag3 framework;
- emChunk, a NP recognizer based on the HUNtag3 framework.

[25] http://e-magyar.hu/en

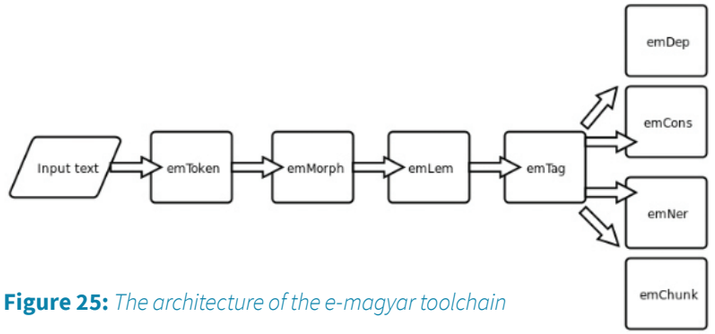


Figure 25: *The architecture of the e-magyar toolchain*

To further improve the efficiency and user-friendliness of e-magyar, the whole architecture was given a thorough overhaul in which the GATE framework was replaced with an inter-module communication framework that follows the toolbox philosophy. The new architecture makes e-magyar not only a truly modular, easy-to-use and extendable toolchain, but one that can very quickly be transformed into a webservice and a Python library as well. To illustrate the modularity and enhanced flexibility of the system, many new modules have already become part of the toolchain, providing alternative options to existing modules. For example, the well-known spellchecker and stemmer Hunspell presents an alternative to emMorph and the three UDPipe modules – tokenizer, PoS-tagger, dependency parser can be selected in preference to emToken, emPOS and emDEP. To see emMorph in action, see this demo, showing the analysis of the word *bokraim 'my shrubs'*.

The e-magyar toolchain was developed to suit non-technical users as well. They can use the drag-and-drop Text Parser webservice which accepts short texts and outputs their analysed version to the selected level of detail (see Figure 26). In addition, a more lightweight web option, a web service of emMorph, (showing the morphological analysis of individual words) was also set up to enable linguists to check the analyses of particular words during their annotation work.

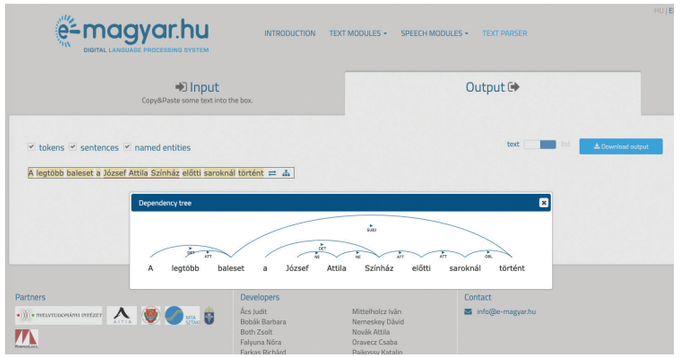


Figure 26: *Dependency parsing in the Text Parser webservice*

Resource | **Multimodal HuComTech Corpus**

Written by **László Hunyadi** and **Tamás Váradi**

The idea of building a multimodal corpus of Hungarian (containing annotations of text, prosody, gaze, gesture, etc.) was conceived 10 years ago. The aim was to improve human-machine communication applications (like chatbots) by empowering them with a comprehensive set of knowledge about human-human communicative behaviour. The underlying assumption was that there exist certain primitives of human behaviour. Such behavioural primitives form temporal patterns which can be assigned functional interpretations. For instance, a prosodic feature like falling intonation followed by a visual cue such as a downward gaze often signals that the speaker wishes to terminate his or her turn in conversation. Such primitives can further serve as a marker which, with a certain probability, points to a pattern with a given interpretation.

When building the HuComTech corpus[26] we first observed and annotated primitives of behaviour at multiple levels, which included recording intonation, morpho-syntactic annotation, video annotation, unimodal and multimodal pragmatic annotation, among others. Subsequently, we interpreted the complex raw annotation phenomena in terms of pragmatic and communicative function, and finally we identified actual patterns of behaviour based on the annotated raw and interpreted data. In total, about 50 hours of dialogues with 111 subjects were recorded in two (formal and informal) scenarios. HunCLARIN experts captured the multimodality of human-human communication by observing a wide range of both non-verbal and verbal behaviour. The primitives of non-verbal behaviour were either visual or audio in nature. The visual primitives included eye gaze (direction and blinking), eyebrows, head, hand and (upper) body movement, perceived emotions and a range of pragmatic and communicative categories (such as turn management, agreement, certainty, etc.). The non-verbal audio primitives included a range of prosodic features, perceived emotions and a range of pragmatic and communicative categories.

[26] http://tla.nytud.hu/

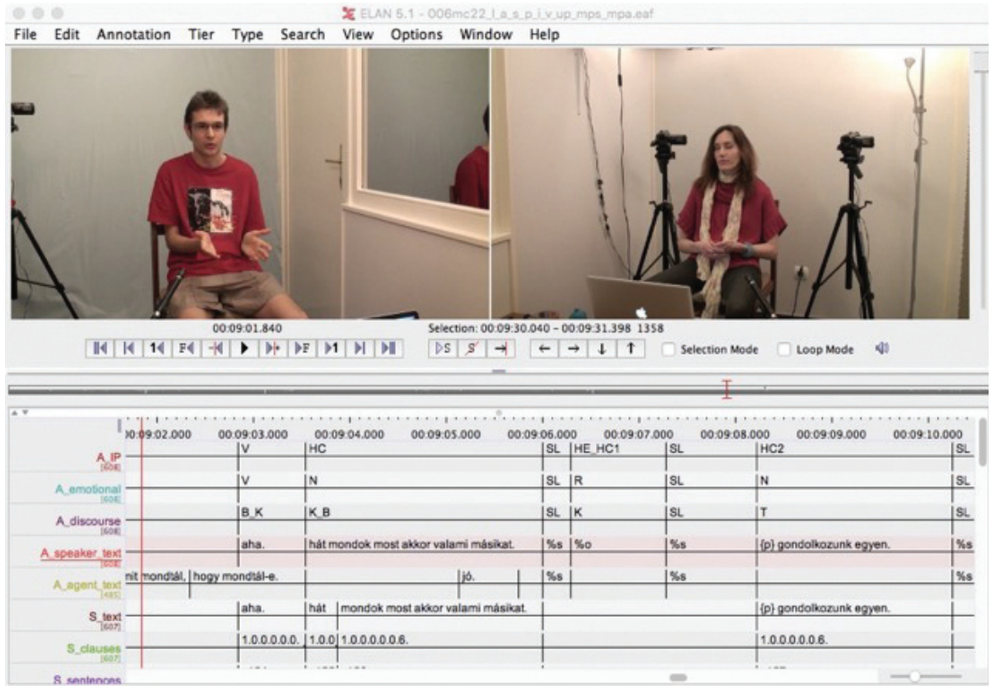


Figure 27: *Annotation of the HuComTech corpus with the ELAN tool*

The annotation of verbal primitives was aimed at offering the first of its kind of syntactic analysis of spoken language. This was partly done by using another Hungarian CLARIN tool, *magyarlanc*, while the German CLARIN *webmaus* tool was used for the alignment of words on the timeline. The specific features of the HuComTech Corpus include its unique conception of multimodality, which actually represents the synthesis of three approaches: the annotation of primitives and functions both based on visual observation alone, the same kind of annotation based only on audio observation, and genuine multimodality based on audio and visual clues together. This threefold distinction of primitives within multimodality allows for capturing behavioural patterns at three levels (vision, audio, and their joint complexity), facilitating the building of two-way communication systems. The corpus has also been successfully used in linguistic research; for instance, Hunyadi (2019) used HuComTech to study the multimodal expression of agreement and disagreement, while Szekrényes (2019) presented an approach to the post-processing of temporal patterns based on the multimodal data in the corpus.

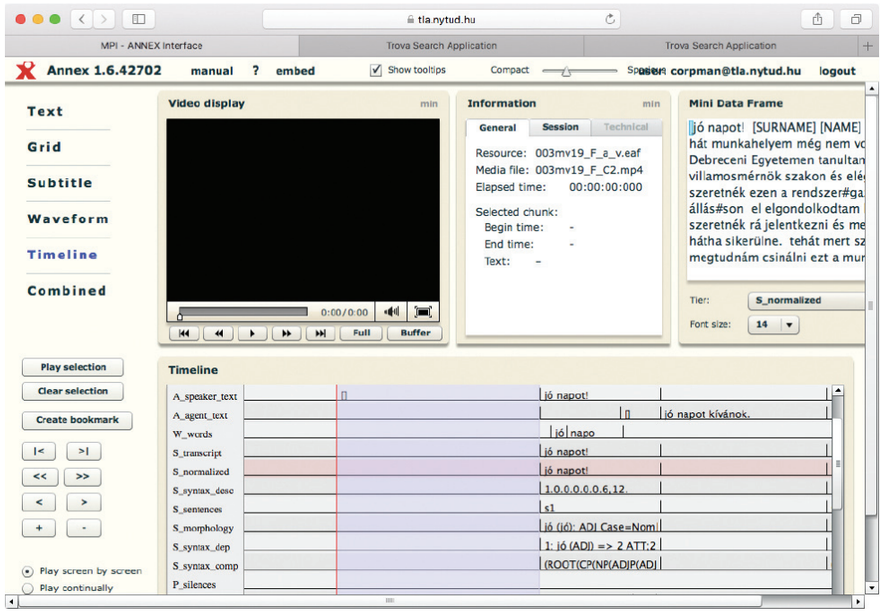


Figure 28: *Browsing the corpus at the HUN-CLARIN repository tla.nytud.hu with the Annex tool*

For more information about the HuComTech corpus see Hunyadi et al. (2018).

References:

Hunyadi, L. 2019. Agreeing/Disagreeing in a Dialogue: Multimodal Patterns of Its Expression. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2019.01373.

Hunyadi, L. et al. 2018. Human-human, human-machine communication: on the HuComTech multimodal corpus. In *Proceedings of the CLARIN Annual Conference* 2018, 56–65.

Szekrényes, I. 2019. Post-processing T-patterns Using External Tools From a Mixed Method Perspective. *Frontiers in Psychology.* https://doi.org/10.3389/fpsyg.2019.01680.

Event | **The HUN-CLARIN Roadshows**

Written by **Réka Dodé**

HUN-CLARIN, in line with the overall CLARIN mission, considers it a high priority to support humanities and social science research with suitable language technologies and language resources.

Unfortunately, Digital Humanities in Hungary has recently suffered a serious setback when the budding DH training at the master's level was suddenly stopped by the Hungarian Government. While this did not mean an end to Digital Humanities research at the Centre for Digital Humanities – Eötvös Loránd University, with which HUN-CLARIN have established good cooperation, it did certainly make the task of reaching out to DH researchers a lot more challenging.
To counter this difficulty, HUN-CLARIN devised the concept of a Roadshow series that is based on the idea of proactively (and literally) bringing language technology to where humanities research is actually done, namely, to Hungarian universities. The other key concept behind the Roadshows is that instead of a one-sided evangelization of language technology, the workshop should mobilize and showcase local initiatives.

The template for the workshops is therefore divided into two parts. The first part features an introduction to HUN-CLARIN and the value proposition of CLARIN in general. The introduction is then followed by an overview of the basic resources and tools HUN-CLARIN offers to Humanities researchers. The second part highlights local Humanities projects where there had already been an initial attempt to employ some language technology tool in research, or where an interesting dataset and/or a vaguely perceived need for some automated method for analysing the data had been identified. Both parts of the workshops have been designed for plenty of interaction between the visiting HUN-CLARIN team and the hosting researchers, but the afternoon sessions in particular are buzzing with excitement for both teams.

The visiting HUN-CLARIN team is fascinated by the genuinely interesting research questions and the data that the local colleagues have shown in their presentations. Researchers of the host institution, on the other hand, appreciate the advice they have received and the perspectives that it has opened with regard to their own work. Because the research profiles of the hosting institutes have been different, each event has had a different focus and offered intellectual excitement of a different kind, but with a consistently high level of intensity.
So far HUN-CLARIN has organized three such Roadshow events. The first took place at Szeged

University on 18 October 2018 and was held with the support of a grant by CLARIN ERIC.[27] The second workshop was staged at Debrecen University on 7 February 2019,[28] and the third event was organized in Pécs on 2 May 2019.

In Szeged the local contributions focussed on speech and language technology tools, in Debrecen the topics of the local presentations centred on questions of building and using the HuComTech multimedia corpus, while in Pécs the discussion revolved around computational linguistic approaches to discourse representation.

HUN-CLARIN plans to take the Roadshow to other universities in the future. We take encouragement by the fact the local organizers of the Szeged workshop have since indicated that in response to local demand they want to stage a follow-up workshop in November 2019.

The HUN-CLARIN Roadshows

[27] http://clarin.hu/en/content/seminar-speech-and-language-technology-tools-0

[28] https://clarin.hu/en/content/use-corpora-language-technology-tools-and-data-driven-methods-human-sciences

Interview | **Noémi Vadász**

Noémi Vadász is a PhD student and junior research fellow who works at the Research Institute for Linguistics. As a computational linguistic with a formal background in syntax and semantics, she collaborates with HUN-CLARIN in the e-magyar project.

Please describe your academic background and your current research position.

<

I am a junior research fellow at the Research Institute for Linguistics, Hungarian Academy of Sciences at the Research Group of Language Technologies. After my BA on Hungarian Literature and Linguistics I have finished two MA programmes: Theoretical Linguistics and Computational Linguistics. I then moved on to the Doctoral School for Linguistics at Pázmány Péter Catholic University and am currently working on my PhD thesis.

>

What is the topic of your PhD and why did you decide to focus on this problem? How are you approaching it and what do you hope to achieve with it once it is completed, both in terms of scientific results and to your research community? What are you currently busy with?

<

The topic of my PhD is coreference resolution, which is widely researched within the scope of computational linguistics. However, I assume that I could show something new because my approach differs slightly from the classical view of computational linguistics. The reason for that is that my way to computational linguistics has led through classic humanities and theoretical linguistics, therefore I investigate this topic rather as a theoretician but I keep in mind the applicability as well.

Currently I am building a coreference corpus which – beyond the usual analysis layers such as tokenization, part-of-speech tagging, morphological analysis and dependency parsing – will contain anaphoric and coreference relationships. In the example '*I called my mother. She was really tired*.' the personal pronoun 'she' refers back to its antecedent 'my mother' and this relationship is called anaphora. In contrast, coreference occurs when two expressions have the same referent and there are numerous forms of this relationship (e.g. repetition, name variants, synonymy, part-whole relationship, etc.). In the example '*I bought a bicycle. Tomorrow I will ride home my new bike*.' the base of the coreference relationship between 'bicycle' and 'bike' is synonymity.

Anaphora and coreference show similar behaviour across languages. However, in contrast with English, Hungarian is a pro-drop language, which means that some pronouns (namely the personal and possessive pronouns) can be dropped from the sentence following fairly subtle rules. In these cases, the person and number of the subject and the object can be calculated from the inflection of the finite verb, and the person and number of the possessor are calculable from the inflection of the possessed, therefore the use of zero pronouns can be handled in a simple rule-based manner. As a zero pronoun can also refer back to its antecedent, it needs to be indicated in the coreference corpus. I have created an application that inserts the dropped pronouns into the texts, therefore these pronouns can also play a role in anaphora resolution. The corpus could serve as a resource for further research on this topic, be it answering theoretical questions or a technical application for a certain purpose.

Building a corpus of gold standard quality is definitely complicated and time-consuming. But still, the process of corpus building allows one to study the object of anaphora and coreference very meticulously. The feedback of my annotators also gives lessons to be learned. Therefore, together with the corpus, I increase my own knowledge about the phenomena. At the end of the pilot phase I am going to be in possession of the know-how that allows further enlargement of the resource.

>

How did you get involved with HUN-CLARIN and what is your experience with it?

<

My department has multiple connections with HUN-CLARIN. Firstly, the Old Hungarian Corpus (http://oldhungariancorpus.nytud.hu/) was produced in my institute. Initially, I was involved in this project as an annotator, I manually corrected the output of the optical character recognition on Old Hungarian texts. Later, to speed up the work, I developed a small script for helping manual normalization (standardization of old or non-standard texts). It turned out that manual work could be considerably cut down with the help of this pre-normalization tool.

Secondly, I am involved in the e-magyar project, a text processing pipeline for Hungarian, which is also connected to HUN-CLARIN. Last year I developed two small but useful modules for e-magyar, both of which are responsible for conversion between certain formats. One of them converts from the e-magyar tagset to an international standard part-of-speech tagset of Universal Dependencies (UD). The converter is needed for intermodular communication inside the pipeline, but could also serve as a useful output formalism due to the prevailing nature of UD. The other converter is applied between the internal format of e-magyar and the CoNLL-U format, a widely used international standard. The conversion between these two formats allows further work, annotation or visualization of the output with other tools related to the ConLL-U format. Both of the converters were needed for my own purposes in my corpus building project, but soon it turned out that the covered formalisms could be useful for other users as well, and therefore the converters have now been integrated in the e-magyar framework.

>

In addition to contributing to the development of e-magyar, you also have extensive experience in using it in practice. Could you briefly describe this dual role of yours and the advantages it brings?

<

I have a double relationship with e-magyar as I am an everyday user of it and a member of the developer team as well. This duality brings benefits: on the one hand, my needs are fulfilled thanks to the work of my colleagues, and on the other hand, my everyday experience with e-magyar serves as a useful feedback, which is important for maintenance and further development of e-magyar.

I use e-magyar principally in my corpus building project. Initially, the selected texts are analysed with the tokenizer, morphological analyser and part-of-speech tagger modules of e-magyar. Then, the output of e-magyar must be corrected manually, because the quality of the other annotation layers can be strongly influenced by the initial step. Next, the texts with the corrected annotation layers are further analysed by a sentence parser module of e-magyar, which produces

the dependency trees of the sentences. This layer needs manual correction as well. At this point of the workflow, the texts with the corrected annotations are accessible for further, higher-level analysis, such as anaphora resolution.

I am working on three applications in connection with my PhD. The first one is responsible for inserting zero pronouns, the second resolves anaphora and the third resolves coreference. Indeed, the output of these applications also needs manual correction, but finally, besides a high-quality gold standard corpus, I obtain valuable observations of the quality of my applications. I hope that these three applications can also be added to the e-magyar chain as modules in the future.

>

You have recently also been involved in the development of Normo, a tool for the normalization of historical Hungarian. How is historical Hungarian different from contemporary Hungarian and why is such a tool needed? How does it work and who it is intended for? Has it been used on any text collection that is important for Hungarian humanities researchers?

<

Normalizing old texts is an important step in the workflow, because of the heterogeneity of the old orthographic system applied in historical texts. Normalization makes the text readable for humans and also for computers. There are multiple approaches to normalization – our project aimed to preserve the structures of the old language variety making them investigable for historical linguists, and therefore the task of normalization here mainly means the standardization of the spelling (thus covering the differences between the Middle and Modern Hungarian alphabet).

Since manual normalization is time-consuming and requires highly skilled and delicate work, application of automatic methods can help a lot. According to our measurements and the feedback from our annotators, Normo, our pre-normalization tool, eases and shortens the manual normalization work.

Normo consists of two main modules. the first one is a memory-based module with a relatively small dictionary of the most frequent words in the New Testament and their modern equivalents. Based on this dictionary, the most frequent words can simply be replaced with their modern forms. The second one is a rule-based module which works with manually defined rewrite rules. These rules come from two sources: some of them were defined on the basis of known changes in the history of Hungarian, others were defined through corpus-based observations. While the character-level rules are applied inside a word (e.g. replacement rules for handling characters that are not used in Modern Hungarian), the so-called token-level rules operate across word boundaries for splitting or joining words according to the rules of the modern orthography. Normo has been used in the project of building the Old Hungarian Corpus and has been applied to our five Middle Hungarian Bible translations.

>

What are your plans and dreams for the future?

<

My biggest future plan is to work further on my coreference corpus and to make it available for others. With this it will be all set for seeking answers to other exciting questions. I also have to write up my dissertation. Apart from the work on my PhD I have recently been working on some other topics a lot. For instance, I became interested in morphological tagsets. I assume that I could exploit my theoretical–computational hybrid attitude in this field as well. Lastly, I have some favourite topics which I have already been working on (e.g. authorship attribution), I would like to work further on these topics later.

>

How can research infrastructures such as HUN-CLARIN best serve early-stage researchers and how can a new generation of researchers best contribute to the research infrastructure?

<

Recently, for example, I have attended a CLARIN workshop on NLP tools for historical data, which was a great opportunity for me. On the one hand, the event gave me a chance to get to know other researchers of a specific field. On the other, as I think it is essential for beginners to gain self-confidence among their colleagues, which comes gradually through presenting your research often, not to mention the fluent use of English. Additionally, CLAIRN conferences and workshops serve as a good platform to share new ideas with colleagues who have more experience and get useful feedback. The world of conferences, workshops and networking is of course only one aspect of the CLARIN infrastructure's benefits. However, according to my recent experiences, it is one really worth mentioning.

>

BULGARIA

Introduction

Written by **Petya Osenova** and **Kiril Simov**

Bulgaria has been a founding member of CLARIN ERIC since 2012. In 2014, following the strategic plan of the Bulgarian Government and Ministry of Education and Science, the CLARIN and DARIAH Infrastructures merged into a single infrastructure called CLaDA-BG (CLARIN and DARIAH in Bulgaria) and obtained funding in 2018.[29]

In Europe such models have already proved to be successful in the Netherlands, Austria and Greece. The CLaDA-BG consortium is very heterogeneous; its members come from universities, other academic institutions, museums, libraries, non-government organizations and companies. It includes a group of language and semantic technology oriented partners, on the one hand, and expert and content oriented ones, on the other.

The first group includes: the Institute of Information and Communication Technologies at the Bulgarian Academy of Sciences (Coordinator for CLaDA-BG and CLARIN-BG), Institute of Mathematics and Informatics at the Bulgarian Academy of Sciences, Ontotext AD (Sirma AI), Sofia University “St. Kliment Ohridski” (Coordinator for DARIAH-BG), New Bulgarian University, Konstantin Preslavsky University – Shumen, and Bulgariana – an NGO promoting CH in Bulgaria. The second group includes: the South-West University “Neofit Rilski” – Blagoevgrad, Sirma Media, the Cyrillo-Methodian Research Centre at the Bulgarian Academy of Sciences, Institute of Balkan Studies and Centre of Thracology “Alexander Fol” at the Bulgarian Academy of Sciences, Institute of Ethnology and Folklore Studies with Ethnographic Museum at the Bulgarian Academy of Sciences, Burgas Free University, “Ivan Vazov” Public Library – Plovdiv, and Sofia History Museum.

The mission of the infrastructure is to build a scientific ecosystem for supporting research in Social Studies and the Digital Humanities. The main goal is to construct a Bulgaria-Centric Knowledge Graph (BGKG) – repository where all types of linguistic and encyclopaedic knowledge are stored and linked. Thus, they will be used for extracting content with respect to particular tasks.

In their first year of operation the consortium worked on: structuring of the various resources, extending and building contemporary and old corpora, and modelling cultural objects, contextualizing the knowledge through connecting events, artefacts, and descriptions.

Some of the main resources to mention are: the syntactic corpus BulTreeBank (215,000 tokens), the BTB-Wordnet that is integrated with Wikipedia (22,000 synsets), the Valency Lexicon (6,000 verb frames), the Inflextional Lexicon (over one million wordforms) (the Institute of Information and Communication Technologies), the large Bulgarian corpus with statistics on collocations with a span of one to six tokens (eight million webpages have been processed) (Ontotext AD), the Corpus of Child Speech (33 hours of records and 355 pages of transcripts) (Shoumen University), the Ethnographical Museum exhibition on 3D representation, the epigraphic collection of ancient inscriptions in Greek – TELAMON[30] (Sofia University), bilingual corpora (New Bulgarian University), and so on.

Among the most important tools for Bulgarian are: the NLP pipeline and the online concordance webclark (IICT-BAS). Several other tools are also in development: an old-to-new spelling transformation tool, a conceptual and keyword search tool over a huge corpus of contemporary Bulgarian, and a semantic annotator of Bulgarian. CLaDA-BG’s plans include the creation of CLARIN B and K centre, and applying for assessment during the second year of the project.

[29] http://clada-bg.eu/

[30] https://telamon.uni-sofia.bg

Tool | **BTB-Pipe: a Language Pipeline for Bulgarian**

Written by **Petya Osenova** and **Kiril Simov**

The BTB-Pipe language pipeline for Bulgarian has been developed incrementally over the last twenty years, starting with the Bulgarian-German BulTreeBank project for the creation of a Bulgarian treebank. The BTB-Pipe comprises the following modules:

- Tokenizer and sentence splitter
- Morphosyntactic tagger
- Lemmatizer
- Dependency parser

Bulgarian is an analytical language with rich word inflection, predominantly in the verbal area. The rich morphology inevitably leads to a lot of morphological ambiguity. Consequently, morphosyntactic tagging is more complex in Bulgarian than in languages like English. BTB-Pipe is a hybrid system combining a rule-based module and a statistical module (Simov and Osenova 2001) and uses the BulTreeBank Morphosyntactic Tagset (Simov, Osenova, and Slavcheva, 2004).

The lemmatizer in BTB-pipe comprises a set of transformation rules that have been developed based on the 1998 inflectional lexicon (Popov, Simov, and Vidinska 1998). Since the rules in the lexicon are implemented through the CLaRK system, they can also be used on unknown words in order to produce some guesses with regard to their word lemmas.

This is an illustrative example of a lemmatization rule:

if pos-tag = Vpitf-o1s then
{ remove -ox; concatenate -a }

When the lemmatizer applies this rule to the verb form *четох* (roughly /četoh/), where the inflection *–ox* encodes the features 1st person singular and the past indefinite tense (“I read”), it produces the lemma *чета* (/četa/).

The parser uses MaltParser and Mate Dependency Parser for training dependency trees. The input is the result from the tagger and the lemmatizer, and the output a dependency tree or trees for the sentences in the text, using either an internal set of dependency relations developed for the CoNLL 2006 Shared Task or the Universal Dependencies.

The current version of the BTB-pipe can be used in three different modes: as a standalone application, as a command line, and as a web service. The output of the pipe can be in the WebLicht standard developed within CLARIN-D (Hinrichs et al. 2010) or in the NAF format (Fokkens et al. 2014). Currently, ClaDA-BG is redesigning and reimplementing some of the modules using spaCy with the goal of improving the performance of the pipeline.

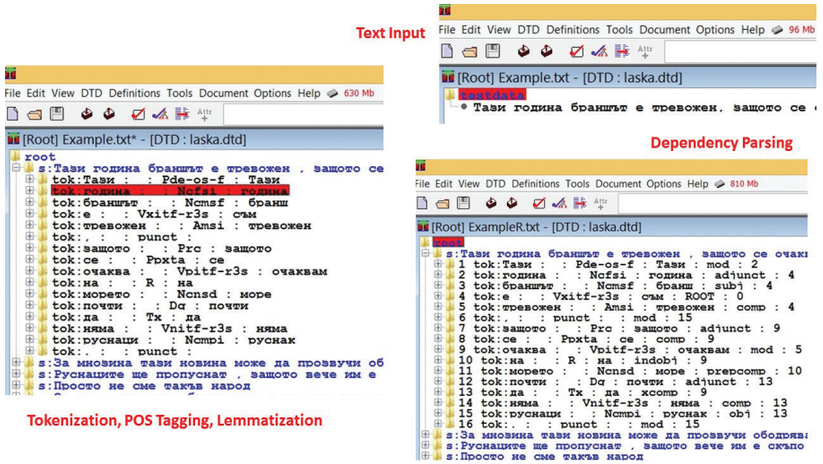


Figure 29a: *Linguistic annotation in BTB-Pipe*

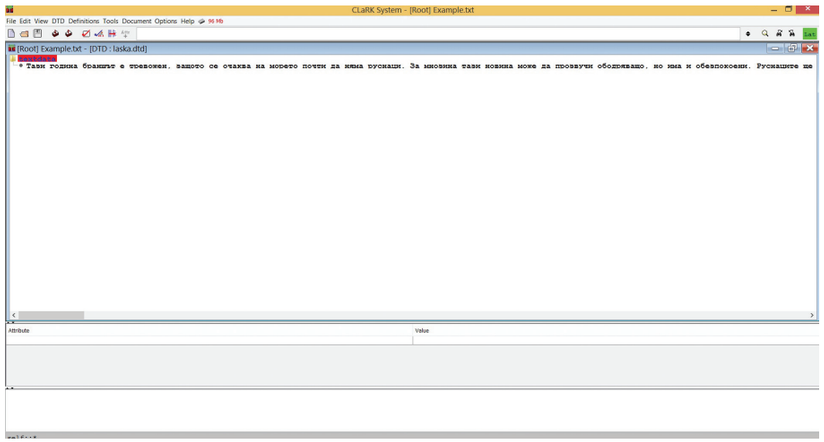


Figure 29b: *BTB-Pipe annotation in the CLaRK XML Editor*

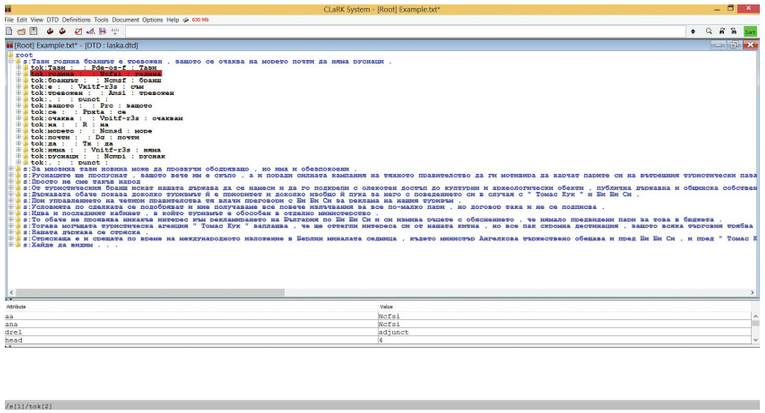


Figure 29c: *Tokenization, lemmatization and morphosyntactic tagging with BTB-Pipe*

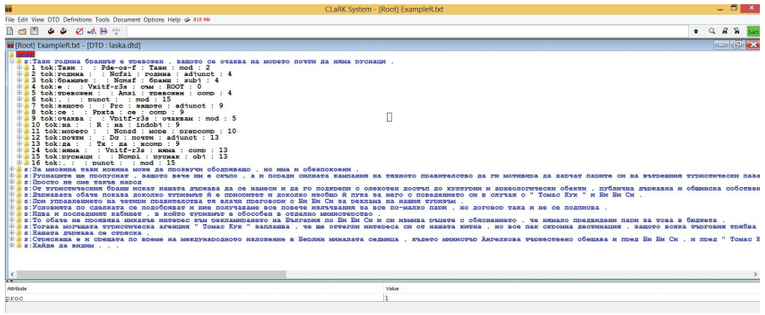


Figure 29d: *Syntactic relations annotated with BTB-Pipe*

References:

Fokkens, A., Soroa, A., Beloki, Z., Rigau, G., van Hage, W.R., and Vossen, P. NAF: the *NLP Annotation Format*. Technical Report NWR-2014-3. Version 1.1. NewsReader project: Building structured event indexes of large volumes of financial and economic data for decision making – ICT 316404.

Hinrichs, E., Hinrichs, M., and Zastrow, T. 2010. *WebLicht: Web-Based LRT Services for German. In Proceedings of the ACL 2010 System Demonstrations*, 25–29, Uppsala, Sweden.

Popov, D., Simov, K., and Vidinska, S. 1998. *A Dictionary of Writing, Pronunciation and Punctuation of Bulgarian Language*. Atlantis LK, Sofia, Bulgaria.

Simov, K., Osenova, P., and Slavcheva, M. 2004. BTB:TR03: *BulTreeBank morphosyntactic tagset BTB-TS version 2.0*. Technical Report.

Simov, K. and Osenova, P. 2001. *A Hybrid System for MorphoSyntactic Disambiguation in Bulgarian. In the Proceedings of the RANLP 2001 Conference*, Tzigov Chark, Bulgaria, 5–7 September 2001, 288–290.

Resource | **Bulgarian Child Language Corpus**

Written by **Petya Osenova** and **Kiril Simov**

The first systematic study of child speech in Bulgaria is attributed to the early 20th century philosopher Prof. Ivan Georgov, but interest in child language truly took off in the last decade of the 20th century and the beginning of the 21st, with Yuliana Stoyanova and Velka Popova who worked on longitudinal data. The contemporary systematic study of child speech, based on solid empirical material, is also associated with the creation of the first Bulgarian corpus in the CHILDES framework by the research team of the Laboratory of Applied Linguistics (LABLING) at Shumen University. The corpus is based on an array of longitudinal data from Popova's personal archive.

The CHILDES framework is reputed for its openness and rationality, which are leading factors in the processes of cooperation and globalization in the Humanities. This is a guarantee of both the broad social validity of the research results based on corpora, and their integration into initiatives for exchanging linguistic data and technologies aimed at overcoming the current fragmentation of the research field. Moreover, CHILDES and the sister initiative TalkBank are already integrated into CLARIN as one of the Knowledge Centres. The Bulgarian child language corpus enables cross-lingual research and contributes to a modern, convenient standard for the study of linguistic ontogeny, which, thanks to its universal parameters, enables rapid, accurate and reliable comparison with a large number of languages and the development of solid typologies and modern theories.

The corpus comprises two types of speech resources: CORPUS A (spontaneous speech material of four children at their early age – from one to three years old) and CORPUS B (comprising stories based on a series of pictures with 90 children at pre-school age (from three to six years old). For the sake of integrity and processing, the speech resources are presented in two formats – in Cyrillic as well as Latin. Figure 30 illustrates the encoding of two children.

@Begin	@Begin
@Participants: ALE Alexandra Target_ Child, VEL Velka Mother	@Participants: ALE Alexandra Target_ Child, VEL Velka Mother
@Birth of ALE: 29-JAN-1989	@Birth of ALE: 29-JAN-1989
@Date: 4-JUL-1990	@Date: 4-JUL-1990
@Filename: al10506	@Filename: al10506
@Age of ALE: 1;05.06	@Age of ALE: 1;05.06
@Situation: at home	@Situation: at home
*VEL: Njama li da spish ti?	*VEL: Няма ли да спиш ти?
*ALE: Dzak.	*ALE: Дзак.
*VEL: Dzak li?	*VEL: Дзак ли?
*VEL: Ja da nankash!	*VEL: Я да нанкаш!
*VEL: Kakvo si na mama ti?	*VEL: Какво си на мама ти?
*VEL: Mamino kokiche.	*VEL: Мамино кокиче.
*VEL: Kakvo da ti donese mama – mlechice ili kompot?	*VEL: Какво да ти донесе мама – млечице или компот?
*ALE: Popot [:kompot].	*ALE: Попот [:компот].
……	….....
@End	@End

Figure 30: *Excerpt from the Bulgarian CHILDES corpus*

Future development of the corpus includes annotation with part-of-speech and morphological information, and integration with the WebCLaRK online service, a Bulgarian portal for language services on the web.[31] Video data also exists for the same material, the processing of which is in progress and will be included in one ClassTalk session in the TalkBank database. The data comprises recorded classroom interactions in a number of kindergarten groups. Video transcription of the video data will follow the same basic principles as audio transcription. These Bulgarian corpora could be used not only for research of classroom interactions between the teacher and children, but also as sample material for training students of pedagogy.

[31] http://webclark.org/

Event | **CLaDA-BG Dissemination Activities**

Written by **Petya Osenova** and **Kiril Simov**

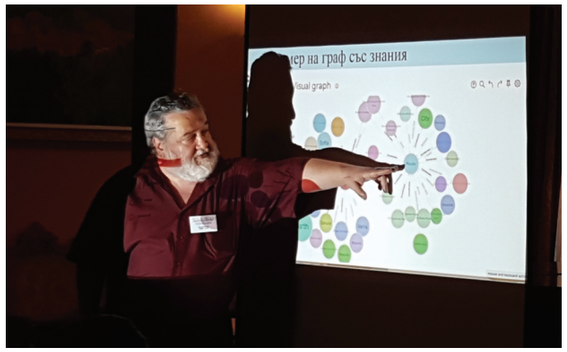
With partners belonging to both CLARIN and DARIAH, the CLaDA-BG consortium is very heterogeneous. For that reason, it regularly organizes seminars and dissemination activities that are aimed at presenting the infrastructure to researchers with backgrounds in the Humanities, such as history, ethnography, library science, and museology. These are 'hosted events', which means that CLaDA-BG experts visit consortium institutions and their teams. During CLaDA-BG's first year of operation we mainly disseminated the goals of the infrastructure to the interested audience and partners. At these meetings we received valuable feedback and also learnt a lot about the needs of the potential users, which are summarized below.

The first awareness raising event was called "Open Science Infrastructures for Big Cultural Data: International Advanced Masterclass". It was organized by UCL Qatar in collaboration with DARIAH-EU and the National Library Ivan Vazov - Plovdiv. The event was held in Plovdiv, Bulgaria, from 13 to 15 December 2018. Members of CLaDA-BG presented the available resources and tools. Dimitar Iliev from Sofia University presented Telamon, which is a corpus of Greek inscriptions found in Bulgaria, while Dimitar Minev, who is the Director of the Plovdiv National Library, talked about how museology datasets can be turned into Linked Open Data resources. A feedback session followed where interesting comments were provided by colleagues from the British Library, especially on the integration of the processing language and image information. The event was concluded by a panel discussion on *the Next Steps in Bulgarian Open Humanities*. It was chaired by CLaDA-BG coordinator Kiril Simov. The panellists were Roumiana Preshlenova (the Institute of Balkan Studies and Centre of Thracology, BAS, CLaDA-BG) and Georgios Papaioannou (UCL Qatar).

On 15 February 2019, a one-day seminar was organized at the Institute of Balkan Studies with a Center of Thracology. At the event, Kiril Simov presented the mission, constitution and organization of the infrastructure, while Petya Osenova presented the resources and tools it offers, after which researchers from the Institute presented their work. In the second part of the seminar, the participants discussed with the lecturers how to make their data and dictionaries machine readable and searchable, and how to OCR and process old books or newspapers at the Institute. The lecturers explained the basic principles of constructing structured data and processing it with the NLP pipe for Bulgarian. In addition, a decision was made to use the data provided by the Institute for the creation of a normalization model to modernize old texts, making them processable by the existing NLP modules.

On 30 and 31 May 2019, Kiril Simov delivered a dissemination lecture at the 12th National Conference “Education and Research in the Information Society”. The lecture was titled “Integrated Language and Knowledge Resources for CLaDA-BG”. The participants were representatives of Bulgarian libraries, universities and educational institutions. The libraries were especially interested in the aspects of improving and speeding up the digitization of their data. In response to their interest, CLaDA-BG started working on the creation of an appropriate normalization model for older texts.

On 23 August 2019, CLaDA-BG experts attended an informal seminar at the Cyrillo-Methodian Research Centre, BAS which has a rich collection of medieval manuscripts in Old Bulgarian, Russian, German and other languages. The main problem of researchers who study Cyril and Methodius is the proper handling of old lexica and their compilation into searchable online dictionaries, a prerequisite for which is OCR and editing, which calls for the reuse of the existing services available in CLARIN centres in other European countries.



Kiril Simov presenting CLaDA-BG to humanities researchers

Interview | **Aneta Nedyalkova**

Aneta Nedyalkova is an MA student of Bulgarian philology. Under the auspices of CLaDA-BG, she is working on an associative dictionary of verbal expressions.

You are a really early-stage-researcher since you are just finishing your MA. How did you get interested in psycholinguistics?

<

My interest was inspired by a course on interdisciplinary approaches in linguistics at Shumen University, especially the project-oriented psycholinguistics practicum which addressed questions that have always excited me. Under the guidance of Prof. Velka Popova I conducted an experiment on word associations with ten people from the Osmar village, Northeastern Bulgaria, where I live. The course assignment then grew into a master’s thesis project and I found myself in the role of a junior but enthusiastic researcher.

>

How did you get involved with CLaDA-BG?

<

I was in an advanced phase of my associative investigation on my master’s thesis when I attended a presentation on CLaDA-BG by the local coordinator Prof. Dimitar Popov and immediately realized that my research interests would be a great fit for CLaDA-BG, as it would give me an opportunity to collaborate with experienced linguists and researchers with similar interests and so learn from them.

>

What are you working on at the moment?

<

I am finishing my master's thesis, titled "Specific features of the contemporary Bulgarian native speakers' dictionary. A psycholinguistic research", which presents a pilot survey on the mental lexicon of non-expert native speakers based on three psycholinguistic procedures: a free associative experiment, spontaneous elicitation of definitions and example sentences for a given word. My results motivated me to extend experimental work and create a dictionary of verbal associations of 100 people from the Osmar village which will be my main contribution to CLaDA-BG.

>

What makes such a dictionary important from a psycholinguistic perspective? Why did you choose this region?

<

One of the key challenges of psycholinguistics is the study of the mental lexicon. A free associative experiment is one of the most popular approaches for this, as it allows psycholinguists to uncover very broad semantic patterns that exist in human consciousness, that is to say, cognitive links between words that are not based on lexical-logical relations, such as synonymy and antonymy, but on more ephemeral associative links. In an associative experiment, you ask participants to write down certain word combinations that they associate with a target notion.

One of the main tasks of CLaDA-BG's Shumen team is the creation of several contemporary associative dictionaries which will serve as a basis for investigations of language awareness in Bulgarian society and for researching the sociolinguistic aspects of the Bulgarian mental lexicon. In the Bulgarian lexicographic tradition, there already exist two associative dictionaries. The first is the *Bulgarian Standards of Verbal Associations from 1984 and the second is The Slavic Associative Dictionary: Russian, Belarusian, Bulgarian, and Ukrainian* from 2004. While the 1984 dictionary is outdated, The Slavic Associative Dictionary, albeit more recent as well as multilingual, has a major methodological flaw in that it includes data from only one social group, students between the ages of 18 and 25. Consequently, the application of the dictionary in research is quite limited.

CLaDA-BG aims to create new associative dictionaries that will be broader in scope with regards to sociolinguistic variables, and will account for differences in territorial origin, gender, age, education, and profession. Consequently, they will be useful resources for a wide range of users across the Humanities and social scientists, such as linguists, psycholinguists, sociolinguists, ethnolinguists, cognitologists, psychologists, teachers, and political scientists. My task to create an associate dictionary of verbal expressions on the basis of the inhabitants of the town of Osmar is just one of CLaDA-BG's associated dictionaries. I have chosen this town for two reasons. First, Osmar is an urban-type settlement, in-between a typical town and a typical village. This is reflected in the inhabitants' specific lifestyle, clothing, and attitude towards technological progress, that is also reflected in the collective features of their mental lexica. The second reason is personal – I live and work as a secretary in Osmar's local library.

>

Could you describe in more detail the compilation of the dictionary of verbal associations and related preparation of questionnaires and experiments? What are the inspiring parts of this work and what are the challenges?

<

The compilation of an associative dictionary involves several stages. First, you need to design the associative experiment that will be the basis for the dictionary. This involves selecting the participants, determining the word-stimulus pairs, setting up the research design (written or spoken) and developing the research materials. Then the experiment is run with every participant separately, which is followed by processing and summarization of the results. Direct contact with the participants is the most inspiring part for me. It is a challenge for me to prepare and motivate them to participate in the experiment. The actual experiment is always interesting, and sometimes very funny or even emotional. The final part of summarizing the data in a systematic way in the dictionary is the hardest and most exhausting, but the curiosity to see the results keeps me inspired and enthusiastic even in this last stage.

>

How did the services and knowledge expertise in CLaDA-BG support you in your work?

<

In general, research is often lonely and challenging for a beginner, so being part of the CLaDA-BG team helps a lot. After the initial training of the young researchers by Prof. Popova, we were offered guidance by the local project coordinator Prof. Popov. In addition, the Student Linguistics Club was established – a small community for like-minded students in which we discuss our research in the infrastructure.

Several different associative dictionaries are currently being developed, some of which have resulted from the joint work of students under the guidance of CLaDA-BG. One of the PhD students is the coordinator and synchronizer of the collected data, which is organized and submitted in separate batches. They are then reviewed by the scientific supervisor Prof. Popova. This way, the data goes through two levels of verification, which provides control and guarantees the objectivity of the results.

The first year of work on the associative dictionaries has shown me the importance of the guidance that CLaDA-BG has offered me related to the compilation of the associative dictionary. Their other language services, such as models and standards for data processing, are also important and useful for us young researchers. The further expansion of these services and the related software environment, which are part of the consortium's future agenda, could be considered as an optimal perspective for the accomplishment of higher quality of the research work.

>

What tools and/or resources of CLaDA do you find most useful for your current and future work and why?

<

For my research, the most helpful CLaDA-BG resources are the corpora of spontaneous speech and the lexicons which provide me with material for the verification of my hypotheses and theoretical models. In addition, I also use the language processing modules, such as the part-of-speech tagger and the sense annotator, because they make my data structured and searchable. Finally, I extensively use the WebClark concordancer for detecting additional contexts that provide explanations for various associations.

>

What do you envisage focussing on after the completion of your MA?

<

After my MA I intend to continue my research in the field of psycholinguistics. I plan to participate in the work on the expansion of associative data and speech corpora by CLaDA-BG, which will be beneficial for theoretical as well as applied activities of a wide range of specialists. In conclusion, I would say that I am very happy to be a member of the CLaDA-BG team, since with my experience and data I am able to help other researchers but also have the opportunity to further develop my competencies.

>

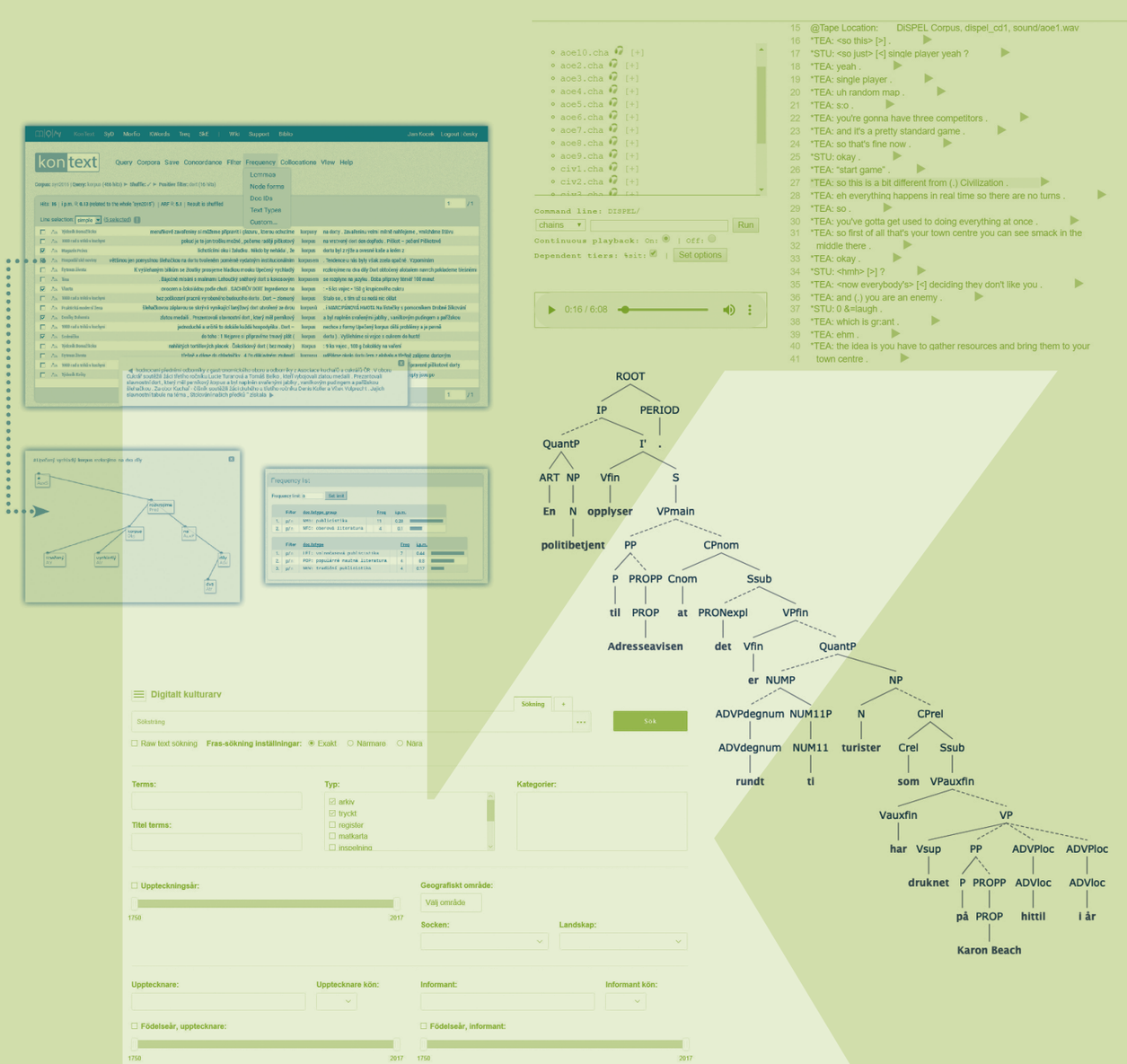
K-centres featured in this volume:

CLARIN Knowledge Centre
for Treebanking

CLARIN Knowledge Centre
for the Languages of Sweden

The TalkBank
Knowledge Centre

The Czech CLARIN Knowledge Centre
for Corpus Linguistics



PART 2

K-CENTRES

CLARIN Knowledge Centre for Treebanking

Introduction

Written by **Koenraad De Smedt** and **Jan Hajič**

On June 25, 2015, CLARIN recognized a virtual Knowledge Centre for Treebanking operated by a consortium consisting of the following:

1. the CLARINO Bergen Centre at the University of Bergen, Norway;
2. LINDAT/CLARIN at the Charles University in Prague, Czech Republic.

The aim of the CLARIN Knowledge Centre for Treebanking is to provide support for researchers interested in the following:

1. building, depositing, and/or disseminating their treebanks;
2. exploring existing treebanks available at the consortium.

Knowledge is transferred through training events, written documentation, personal advice, hands-on assistance and resource hosting. The members of the consortium offer serviced open platforms for constructing, managing and exploring treebanks. Access to some treebanks is open, while access to others is restricted to registered users after signing in. Users can be authenticated through single sign-on at members of the CLARIN Service Provider Federation or of eduGAIN.

The CLARINO Bergen Centre operates INESS[32] (Infrastructure for the Exploration of Syntax and Semantics), an integrated treebanking environment with the following online services:

- accessing, searching and visualizing treebank data in various formats (dependency, constituency, LFG, HPSG);
- building LFG treebanks by parsing and discriminant disambiguation;
- editing dependency treebanks.

Most services can be accessed using a web browser, but uploading treebanks and grammars requires manual support. Currently INESS has more than 200 treebanks available in more than 70 languages.

[32] http://clarino.uib.no/iness/page?page-id=Getting_started

Written knowledge-sharing on the INESS site includes the following:

- a welcome page;
- a page for getting started;
- an overview of the extensive documentation (including walkthrough and documentation of grammar, query language, web interface, annotation and formats);
- an FAQ (list of frequently asked questions);
- a user forum;
- a list of publications;
- links to related information (including a video and slides from a demo);
- project background.

INESS provided interactive knowledge-sharing at the following events:

- Tutorial at the CLARA Thematic Course on Consolidating and Harmonizing Treebank Annotation, Prague, 2010
- INESS Training Workshop, Solstrand, 2013;
- INESS Training Workshop for NAOB, Solstrand 2014;
- INESS Training Workshop at the ParGram meeting, Warsaw, 2015;
- INESS Training Workshop, Solstrand, 2016;
- tutorial on Multiword Expressions in Treebanks at the 2nd PARSEME Training School, La Rochelle, 2016 (with written notes);
- workshop at MONS, Solstrand, 2017.

LINDAT (Prague) offers interactive services to:

- deposit treebanks in a repository;
- visualize treebank data using Treex;
- search and visualize the treebanks using PML-TQ;
- search treebanks using Kontext.

Written knowledge-sharing at LINDAT includes the following:

- a step-by-step guide for depositing resources;
- an FAQ (list of frequently asked questions);
- a user forum.

Until now, the user forums at the Knowledge Centre have been little used, but the organized events mentioned above have been well attended. Knowledge transfer through personal contact with experts at the Knowledge Centre has proved important for projects aiming at curation of their resources.

Interview | **Helge Dyvik**

Helge Dyvik is Professor Emeritus at the Department of Linguistics at the University of Bergen in Norway. Professor Dyvik is one of the main developers of INESS, a CLARIN K-Centre that is operated by the CLARINO Bergen centre and which provides an integrated treebanking environment for accessing, searching and visualizing syntactically parsed data in various formats.

Please describe your academic background.

<

I have been a Professor of General Linguistics at the University of Bergen since 1983. I studied at the University of Bergen and at the University of Durham, working with Old Norse and Old English phonology and syntax as a graduate student, and also with foundational issues in generative syntactic theory, which became the topic of my PhD thesis. I also did some work in runology, interpreting a number of recently uncovered Medieval runic inscriptions in Bergen. When Lexical Functional Grammar emerged around 1980, I started working within that and some related frameworks. LFG was later used as the annotation framework for the Norwegian and some other treebanks in INESS. I was also involved in some early work in experimental machine translation in Norway. From the late 1990s, I worked on developing an automatic method called Semantic Mirrors, which derives thesaurus-like lexical information from translation corpora. Around the turn of the millennium, Victoria Rosén and I started to develop the first version of the Norwegian Computational Grammar (NorGram), which is a project that we're still working on and now also involves other researchers, some also from the CLARINO network.

>

What is your role in INESS? Could you describe the main goals of this project?

<

I am one of the developers of INESS, which stands for Infrastructure for the Exploration of Syntax and Semantics. This project, which began in 2010, had two main goals. The first one was to establish an infrastructure for various treebanks across languages. There are now around 400 treebanks, large and small, in INESS, covering about 70 languages. The second was to develop the first Norwegian treebank based on 'deep' parsing, which is what most of my work is related to. I was mainly responsible for the further development of the grammar NorGram, in continuous interaction with the annotators (or more properly, the disambiguators) working with the disambiguation of the parse forests of sentences. The treebank, which is called NorGramBank,[33] currently covers around 80 million words (and with that size is obviously for the most part stochastically disambiguated). This is quite a large number for a syntactically parsed corpus, and it is still growing, as we have fashioned it to be a dynamic resource. In 2015, INESS was – in cooperation with the Czech infrastructure LINDAT, which also specializes in the development of treebanks – recognized as a Knowledge Centre in the CLARIN Knowledge Sharing Infrastructure.

>

What are the main goals of INESS as a CLARIN Knowledge Centre?

<

<

We have started actively working on making the search facilities of INESS more user-friendly, which is one of our main goals as a CLARIN Knowledge Centre. Paul Meurer, who was awarded the 2017 Steven Krauwer Award for CLARIN achievements,34 has developed a querying system called INESS Search. This query language is very powerful and can handle various syntactic frameworks, such as Lexical Functional Grammar, Dependency Grammar and Head-Driven Phrase Structure Grammar.

While the query language itself is in many ways simpler than other existing syntactic query languages, the treebank annotations are so complex that the query task may seem complicated to a novice.

To make it more accessible, I have been working on user-oriented, example-based documentation, and with Paul Meurer on developing query templates.

[33] http://clarino.uib.no/iness/page?page-id=iness-descr

[34] https://www.clarin.eu/news/paul-meurer-awarded-2017-steven-krauwer-award-clarin-achievements

The example-based documentation, which is currently only in Norwegian, is based on the structure of the Norwegian Reference Grammar and examples found there, and shows in a step-by-step fashion how to search for the exemplified constructions. The query templates are ready-made queries with parameters to be supplied by the user, and they are integrated into the search environment itself. You can see examples of such templates, originating in cooperation with lexicographers, if you choose "Select query templates" under the Sketch tab on the INESS webpage (Figure 31). This gives you various query formulas for a wide range of both simple and complex syntactic constructions, which is a useful showcase for grammarians and philologists who are not used to working with more complex query languages. The idea is to be able to aid users by supplying new query templates on demand.

>

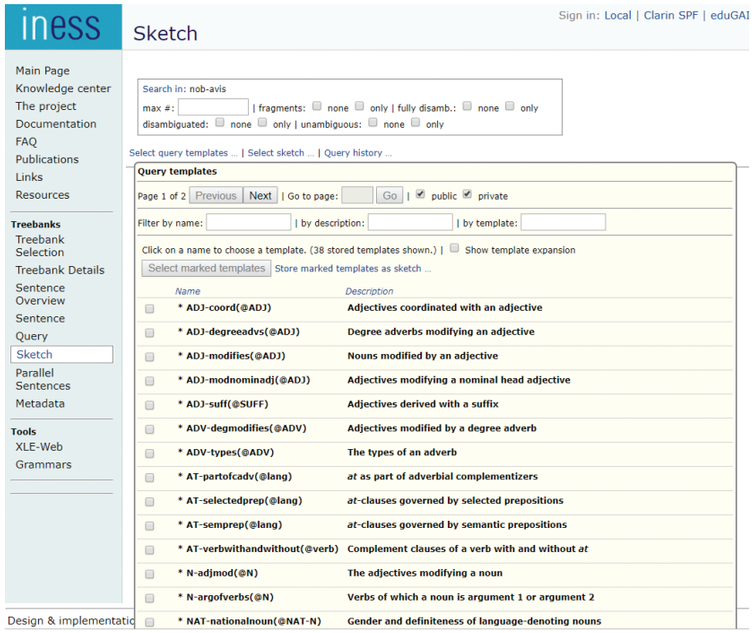


Figure 31: *Query templates in INESS Search*

Could you give a simple example of how INESS Search works?

<

Let's say that we want to find relative clauses that function as modifiers in nominal phrases, like the bolded clause in the English sentence *The **man who is working in the field** is my father*. To find such constructions in the treebank, we only need to enter the following search query in INESS search:

NP > CPrel

This instructs the concordancer to look for all syntactic constructions for which the following holds:

1. There exists a tree structure node which is an NP category and there exists a node which is a clause structure (of type CP, a "complementizer phrase") headed by a relative-clause subordinator, such as *som* in Norwegian.
2. The CP node must be embedded within the NP node, which is specified by the > operator.

Technically, this query is an abbreviation of the quantified expression #x_:NP > #y_:CPrel, where the operator # stands for an existential quantifier that binds a variable for which a specific categorial property is defined (in this case, NP for variable x and CPrel for variable y). However, Paul Meurer has simplified the query language so that it is not necessary to refer to the tree nodes with quantified variables if the variables aren't used more than once in the query.

The results of such a search query are shown in Figure 32:

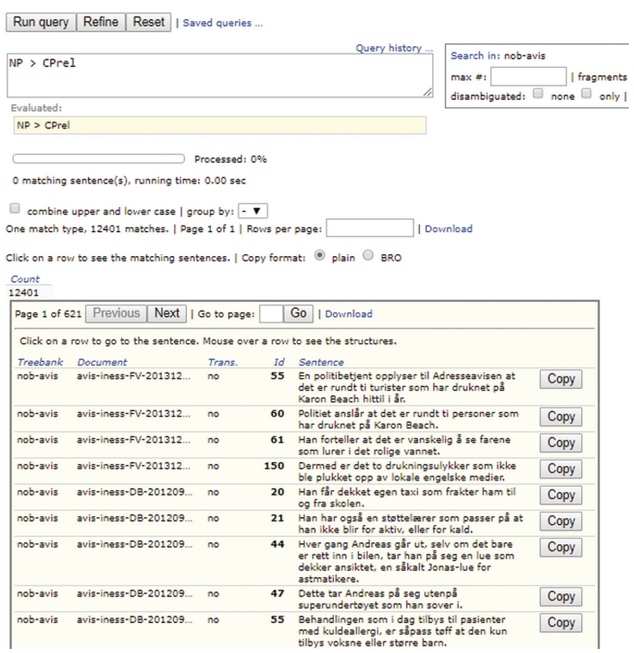


Figure 32: *Using INESS search for identifying relative clauses embedded in NPs*

The first result is the sentence *En politibetjent opplyser til Adresseavisen at det er rundt ti turister som har druknet på Karon Beach hittil i år,* which roughly corresponds to English “A police officer informs the Adressa newspaper that there have been around ten tourists who have drowned at Karon Beach so far this year”. So, the clause *som har druknet på Karon Beach hittil i år* (“who have drowned at Karon Beach so far this year”) is the relative clause embedded in the NP turister (“tourists”), which is what the query was looking for. Clicking on the example leads you to a tree representation of its C(onstituent)-structure and a representation of the corresponding F(eature)-structure, which lists the grammatical features of the nodes in the tree and shows their grammatical functions.

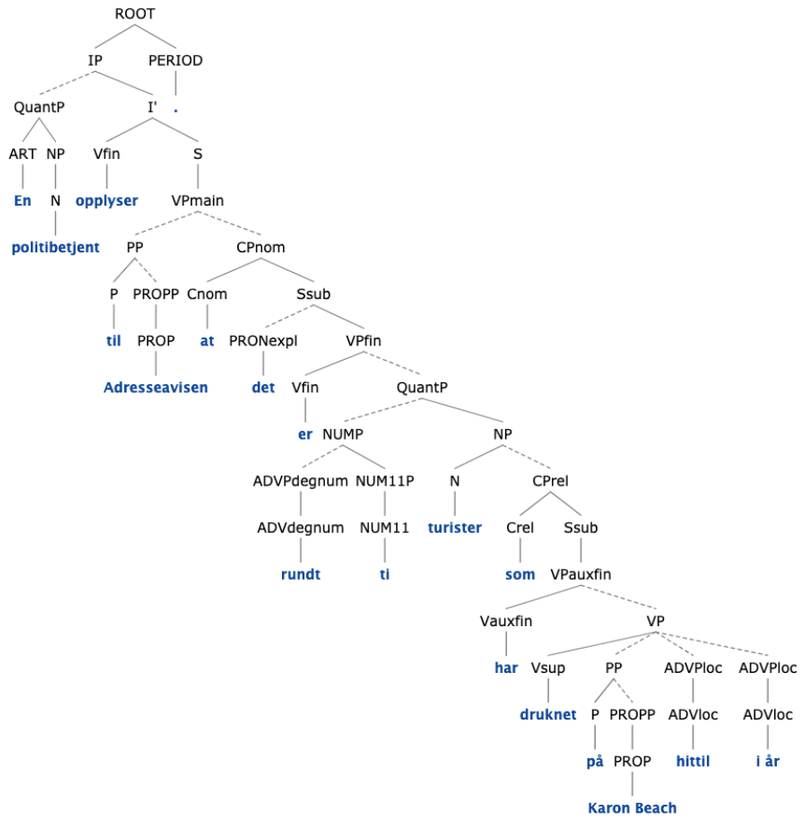


Figure 33: *The LFG C-structure representation of the* sentence En politibetjent opplyser til Adresseavisen at det er rundt ti turister som har druknet på Karon Beach hittil i år *(“A police officer informs the Adressa newspaper that there have been around ten tourists who have drowned at Karon Beach so far this year”)*

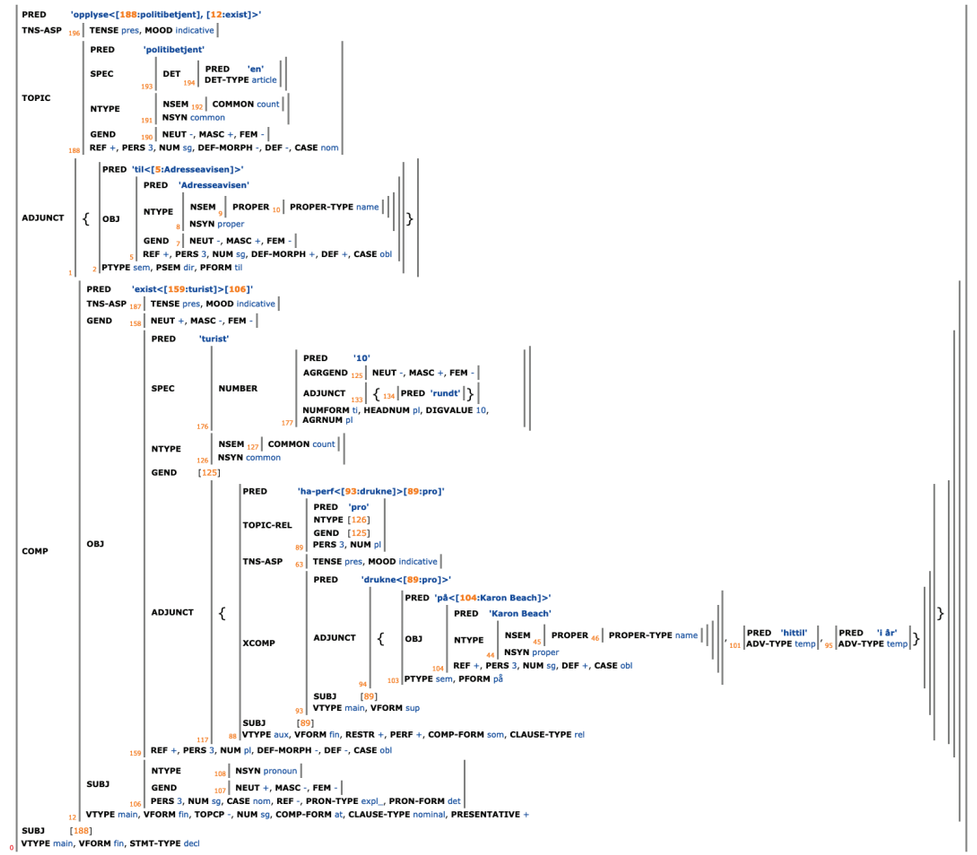


Figure 34: *The corresponding F(eature)-structure*

Additionally, it is possible to formulate search queries that take into account both the C-structure and the corresponding F-structure information, as in the following example:

NP > #x_ >> #f_ >CLAUSE-TYPE ‘rel’

In this case, the operator >> denotes the mapping from a C-structure node #x_ to an F-structure #f_ that contains the grammatical information CLAUSE-TYPE ‘rel’. This means that the concordancer is now looking for all tree nodes that are embedded within NP and whose F-structure contains the value “rel(ative)” for clause type. This search query now allows us to find relative clauses in NPs both with and without overt subordinators (e.g. *påstanden du nevnte*, “the claim you mentioned”, where the subordinator som is omitted). Since the latter types of relative clauses lack the CP layer in the LFG representation on account of the omitted subordinator, it is more complicated to search for them by only referring to their C-structure, as in the case of the simpler query NP > CPrel.

Such syntactic constructions would be much more difficult – if not impossible – to extract from a corpus that isn't syntactically parsed, since you wouldn't be able to specify any kind of syntactic relations in the query language. As a Knowledge Centre, we also provide help with formulating new search queries, so if a researcher is interested in any kind of syntactic or to some extent semantic phenomenon but doesn't know how to extract it from the treebanks, he or she need only contact us. Additionally, if researchers are interested in a more detailed explanation of the kinds of formal relations that underlie INESS, we have prepared a short walkthrough in English that explains the basic idea behind the query language.[35] A fuller documentation of the query language is also provided.

>

Have the Treebanks of INESS been used in any successful project?

<

Helene Uri, who, aside from being a linguist, is a famous novelist and children's writer, wrote a book called *Hvem sa hva: Kvinner, menn og språk (Who Said What: Women, Men and Language*). She discussed the different ways men and women use language, as well as the different ways in which men and women are written about in various types of discourse. Part of her research was done on the Norwegian Treebank, which, in addition to the syntactic dependencies, provides semantic representations such as predicate-argument structure. Specifically, she used the treebank to find out which verbs are mostly associated with female agents and which verbs with male agents. Helene's book was very successful and she won the Brage Prize 2018 for it.

The NorGramBank treebank has also proven itself important in relation to the rather unique language situation in Norway, where there are two written standards. Bokmål, which is the majority standard, goes back to the beginning of the 20th century and is adapted from Danish orthography and based on educated urban speech. It is therefore the more traditional written standard in that it reflects the fact that Norway was in union with Denmark for 400 years until 1814 and Danish was our only written language at that time (actually not much more distant from spoken Norwegian than standard languages in some other countries are from some of their dialects). The other standard, Nynorsk (originally called landsmål) was constructed towards the end of the 19th century by the poet and linguist Ivar Aasen, who based the standard on the more archaic dialects that were spoken in the rural areas of Norway and were thus not influenced by Danish. What's important for the current language situation is that, from around 1920, the Norwegian parliament introduced policies that tried to merge the two standards. This was an extremely controversial decision that was met with resistance by proponents of both standards and was ultimately abandoned. Both varieties have been recognized as official standards of written Norwegian ever since 1885. However, a result of the failed merging attempt is that there is considerable freedom of choice, particularly with regard to inflectional forms, in both of the official standards. Still, the actual choices made by authors of published texts do not in general reflect the full scope of the official possibilities that still remain – there is an emerging de facto standard, especially within Bokmål. Charting this development in the language therefore becomes an important task.

The NorGramBank treebank is especially useful for observing this rather complex language situation in Norway, as it consists of a wide variety of textual materials like newspaper articles, popular research and parliamentary debates in both standards. It is for this very reason among the resources used by the Norwegian Language Council, which is responsible for language standardization. In addition, the lexicographic project NAOB in Oslo is using the treebank in the further development of a new comprehensive web-based dictionary of Bokmål which was published last year. The Oslo lexicographers now try to help finance further development of the treebank, since they understand the importance of having an up-to-date resource that can provide relevant examples chosen from the literature based on the actual syntactic use of the dictionary lemmas. There are also other lexicographic projects using NorGramBank.

>

You have conducted a few linguistic analyses of your own by using the Norwegian treebank. Could you discuss some noteworthy examples?

<

I am running a blog in which I discuss grammatical phenomena based on the NorGramBank treebank. I focus on some of the well-known syntactic constructions, since there are many misconceptions about their usage in popular discourse. For instance, in Norwegian, one of the stylistic pieces of advice that you hear time and time again is to avoid the passive, the reason being that it supposedly makes a sentence less informative by omitting the agent. However, such advice often isn't accompanied by any contextual justification, so it boils down to a prescriptive rule that doesn't hold water if you look at how passives are used in actual texts. In the treebank, I've noticed that passives are especially prominent in popular science. In articles from the *forskning.no* website, the treebank showed that passives were used in almost 25% of the sentences, on average. Looking at their function in relation to the surrounding context, we usually

[35] http://clarino.uib.no/iness/page?page-id=INESS_Search_Walkthrough

find that their use *in lieu* of the active voice is well motivated. For example, there were many passive sentences like *Disse funnene har ikke vært beskrevet tidligere* ("These findings have not been described previously"), in which the omitted agent of the verbal action is referentially non-specific, which is something that you would expect given that popular science abounds in generalizations. This means that using active variants, like *Researchers have not previously described these findings*, would not make these sentences any more informative. If anything, they would only disrupt the information flow from the perspective of the surrounding discourse. So, when people give out stylistic advice like "avoid the passive", what they generally overlook is the function of the construction – that is, the passive voice is a device that makes it possible for the writer to adapt the information structure of the sentence to what is prominent in the context; its use is communicatively oriented and in fact very useful in most cases. Besides, Scandinavian languages have an especially rich variety of passive types, something which makes this stylistic advice particularly harmful.

In another blog post, I looked at how sentence complexity (defined as number and degree of embedding of subordinate clauses) varies between different text types. Perhaps the most interesting finding is that the transcriptions of the Norwegian parliamentary debates contain the most complex sentences by far, even more so than the written genres. I was also able to observe an interesting difference between the two Norwegian written standards in the domain of literature. I found that children's books written in Nynorsk are the least complex, but children's books in Bokmål contain, on average, more complex sentences than novels in Nynorsk. (However, the limited size of the Nynorsk children's books corpus is a caveat here.)

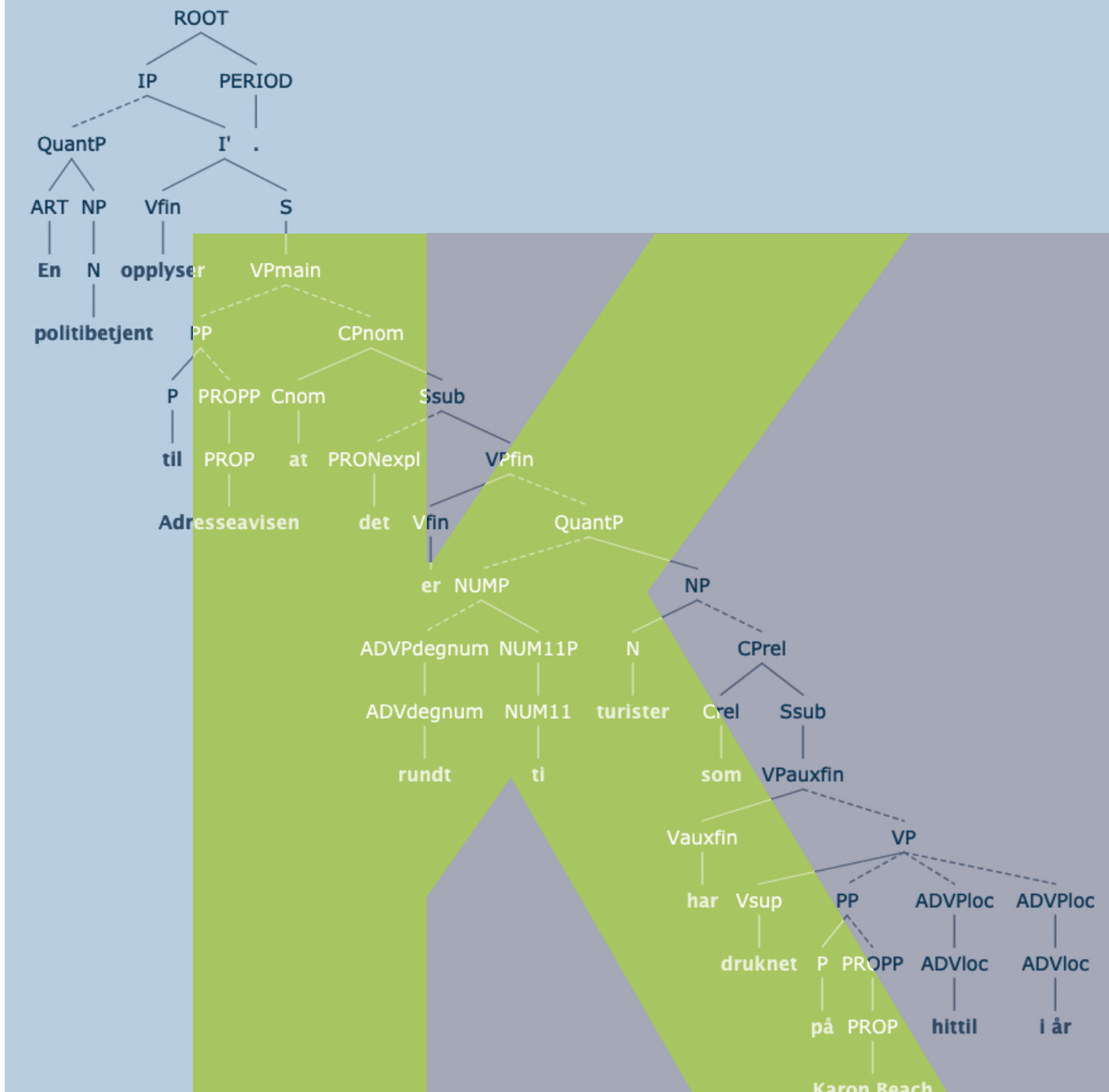
>

What are the future plans with regard to INESS?

<

We plan to expand the corpus of literary Bokmål texts significantly, as part of our cooperation with the NAOB dictionary project, and to continue making the search facilities more accessible, expanding the documentation with ready-made search examples. We also plan on expanding the Norwegian treebank with more texts in Nynorsk.

>



CLARIN Knowledge Centre for the Languages of Sweden

Introduction

Written by **Rickard Domeij**

The SWELANG Knowledge Centre is an information service offering advice on the use of digital language resources and tools for Swedish and other languages spoken in Sweden, as well as other parts of the intangible cultural heritage of Sweden.
The centre is based at the Language Council of Sweden (Stockholm) and is run in cooperation with the other sections of the Institute of Language and Folklore (ISOF) in Uppsala and Gothenburg. The institute is sanctioned by the Swedish government to collect, preserve, process and disseminate scientific knowledge and material concerning the Swedish language, the national minority languages, the Swedish sign language and Swedish dialects, as well as other parts of the intangible cultural heritage of Sweden.

Development of Digital Tools and Services

The SWELANG knowledge centre cooperates closely with SWE-CLARIN and the National Language Bank of Sweden (Nationella språkbanken). The knowledge centre focuses on developing methods for collecting two types of data:

- Official texts and terminology for research in official communication and social conditions. The material is multilingual with parallel texts in Swedish and translations into easy-to-read, plain language of the five national minority languages (Finnish, Sami, Romani, Yiddish and Meänkieli), as well as other minority languages used in official communication.
- Folk narratives, as well as other text and speech material from the dialect and folklore archives. The material consists of inventories, dialect word databases, letters, recordings, transcriptions, etc. It is important both in terms of content and linguistic quality, as it includes a large number of geographical, social, and stylistic varieties.

In addition, the centre is developing methods to manage and make widely available contextualized digital archive material through a map-based research interface called Digitalt kulturarv (Digital Cultural Heritage). The interface is connected to a database of 16,000 complete records. Apart from text material, consisting of transcribed records that were scanned using OCR or HTR, the database also contains metadata, such as year of recording, categories and location, as well as information about the person recording and informants (i.e. name, year of birth, gender). The interface shows not only a list with search results, but also visualizes statistics from the metadata. For example, a map illustrates the geographical distribution of the records. A limited public version of Digitalt kulturarv called Sägenkartan (Map of Legends) can be accessed on the web (in Swedish only). A log-in version with richer content for researchers is on its way.

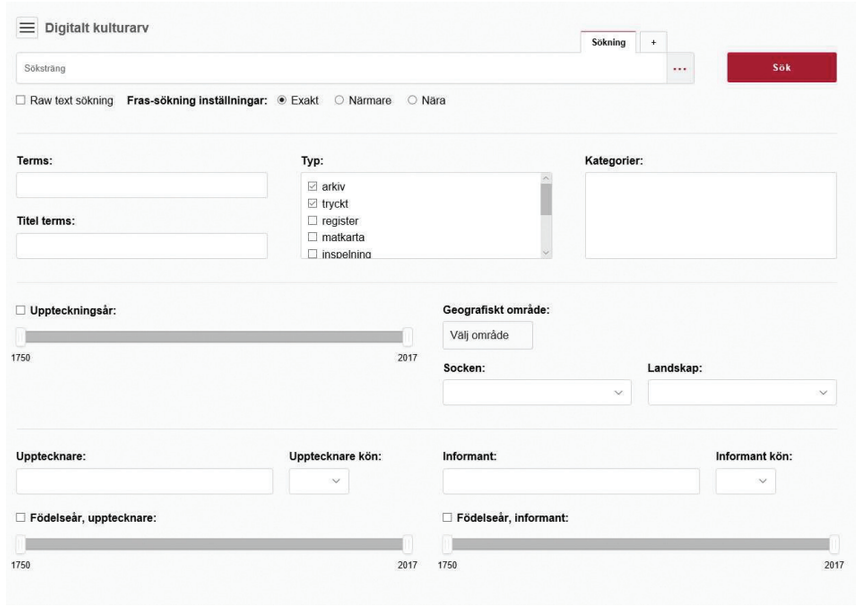


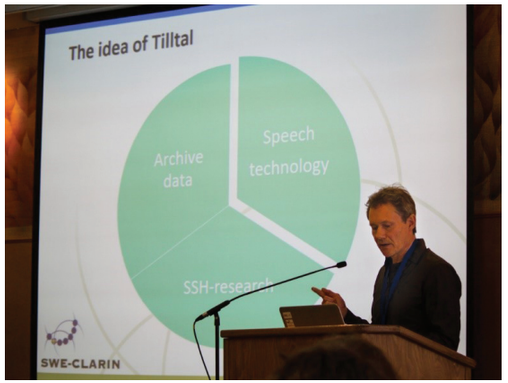
Figure 35: *The search interface of the Digital Cultural Heritage*

The knowledge centre is also developing an infrastructure for dictionaries, the aim of which is to store and make available official terminology and dialect words in collaboration with Språkbanken (a CLARIN B centre). Resources offered by the SWELANG Knowledge Centre are already available through the SWE-CLARIN catalogue. This mostly includes bilingual dictionaries that pair Swedish with other languages spoken in the country, such as the Swedish-Bosnian dictionary and the Swedish-Azerbaijani dictionary.

Interdisciplinary collaboration within the TillTal project

In the TillTal project we examine how speech and language technology methods can make the historical speech recordings more accessible for research in cooperation with data holders, researchers and speech and language technologists. For instance, there are immense amounts of recorded interviews which currently have to be played in real time in order to be analysed. These materials conceal a wealth of information of great interest for the Humanities and Social Sciences.

With digital tools we see possibilities to explore the recordings in new ways. For example, we are exploring methods to visualize and browse large amounts of audio data together with the CLARIN Knowledge Centre of Speech Analysis at KTH (Malisz et al. 2017). This is done by projecting sound segments on a two-dimensional plane with a technique used to find similarities in images, so that representations of similar sounds are clustered together. We hope that this will make it possible to find interesting features in audio files without actually listening to them one by one, for example to identify applause and singing from speech, or even find similar vowel pronunciations. Our archives also include a wide range of information in written form, including descriptions of recording situations and manual transcripts, which we use to provide further pathways into the speech materials (Domeij et al. 2019).



Rickard Domeij presenting the TillTal project

Associated project collaborations

The K-Centre is part of the following national and international language infrastructure collaborations:

- CLARIN — the European research infrastructure for language resources and technology
- ELRC — European Language Resource Coordination
- eTranslation TermBank — collection and provision of terminological resources for machine translation within the EU
- META-NET, a Network of Excellence consisting of 60 research centres from 34 countries, is dedicated to building the technological foundations of a multilingual European information society
- SWE-CLARIN
- TillTal project

References:

Borin, L., Forsberg, M., Edlund, J., and Domeij, R. 2018. Språkbanken 2018: Research Resources for Text, Speech, and Society. Poster DHN I: Mäkelä, Eetu, Tolonen, Mikko and Tuominen, Jouni (eds.) *Digital Humanities in the Nordic Countries 3rd Conference,* 504–506. http://ceur-ws.org/Vol-2084/poster7.pdf.

Berg, J., Domeij, R., Edlund, J., Eriksson, G., House, D., Malisz, Z., Nylund Skog, S., and Öqvist, J. 2016. Tilltal – making cultural heritage accessible for speech research. Paper presented at the CLARIN Annual Conference 26–28 October 2016, Aix-en-Provence, France.

Berg, J, Domeij, R, Edlund, J., Eriksson, G., House, D., Malisz, Z., Nylund Skog, S., AND Öqvist, J. 2017. Involving users and collaborating between disciplines in making cultural heritage accessible for research. Paper presented at the CLARIN Annual Conference 18–20 September 2017, Budapest, Hungary.

Dagsson, T. and Skott, F. 2018. Digital Cultural Heritage — a Digital Folklore Archive [Blog post]. https://sweclarin.se/eng/digital-cultural-heritage-%E2%80%94-digital-folklore-archive.

Domeij, R. and Eriksson, G. 2018. Språkbanken Sam. A CLARIN knowledge center for the languages of Sweden. Poster presented at SLTC 2018 , 20–22 November at Stockholm University.

Domeij, R., Eriksson, G., Lindström, E., Magnusson Petzell, E., Nylund Skog, S., Skott, F., and Öqvist, J. 2019. Text as an entry point to speech – a journey into the most inaccessible areas of the archives. Book of abstracts 4th Conference of the Association Digital Humanities in the Nordic Countries Copenhagen, March 6–8 2019.

Nylund Skog, S. 2018. From personal letters to scientific knowledge: The creation of archived records in a tradition archive. In *Visions and Traditions: Knowledge Production and Tradition Archives*. Helsinki: Academia Scientiarum Fennica, FFC 315.

Malisz, Z., Öqvist, J, Fallgren, P., Edlund, J., and House, D. 2017. Visualizing vocalic variability in space and time – automatic exploration of “found data”. Paper presented at the 47th Poznań Linguistic Meeting, 18–20 September 2017, Adam Mickiewicz University, Poznań, Polen.

Interview | **Susanne Nylund Skog**

Susanne Nylund Skog is an ethnologist and folklore researcher who collaborates with the SWELANG K-Centre in the TillTal project.

Please describe your academic background

<

I am a researcher at the Institute for Language and Folklore in Uppsala, Sweden, where I work at the Department of Dialectology and Folklore Research, and an Associate Professor of ethnology at Uppsala University and of Nordic folklore studies at Åbo Akademi University, Finland.

I defended my doctoral dissertation in ethnology and childbirth stories at the University of Stockholm in 2002, and have since then done extensive ethnographic research on Jewish life in Sweden and on stories by birdwatchers. With performance and narrativity in focus, I have explored issues such as anti-Semitism, whiteness, intertextuality, emotions and materiality. I am currently doing research on archive collections within the project TillTal aimed at making spoken cultural heritage accessible for research, which is funded by Riksbankens Jubileumsfond, the Swedish Foundation for Humanities and Social Sciences.

>

How did you get involved with the K-Centre for the Languages of Sweden? What is the main goal of the Tilltal project?

<

I first came in contact with Rickard Domeij and Gunnar Eriksson from the K-Centre at the SWE-CLARIN exploratory workshop for researching audio materials from a cross-disciplinary perspective. The workshop ended in a joint research grant proposal for the multidisciplinary project Tilltal by the Institute for Language and Folklore, KTH Royal Institute of Technology and the Swedish National Archives.

The overall goal of the project is to make Sweden's archive of recorded speech more accessible for Humanities and Social Science research, which is also one of the main goals of the K-Centre. I am involved in the project as a qualitative researcher who studies the recordings, and I collaborate with language technologists like Gunnar Eriksson who help me with technological solutions for my research questions.

>

How do speech recordings differ from other materials used in Digital Humanities research? What does the TillTal project do to promote the use of speech recordings in multidisciplinary approaches?

<

Speech recordings represent a seriously underutilized resource of the Swedish memory institutions, at least for Humanities and Social Sciences purposes, where researchers often only work with secondary materials, such as transcriptions of the spoken materials, instead of investigating the recordings themselves. One problem is that the number of speech recordings is very large. The archives of the Institute for Language and Folklore alone contain around 25,000 hours of recorded speech. Paradoxically, this contributes to the fact that such materials are not often used by Humanities researchers, as speech is extremely challenging and time-consuming to work with and can be quite unmanageable without appropriate tools.

To help overcome this problem, the TillTal project has established three different case studies and one user study.[36] In the case studies, research agendas from three different Humanities and Social Sciences fields are being pursued with the help of speech technologies. These are case 1: from personal experience narratives to cultural heritage, which focuses on speech recordings in ethnology, case 2: linguistic variation in time and space, which involves collaboration between speech and language technologists and sociolinguists, and case 3: interaction patterns over time and type of conversation, which extends previous work within interaction analysis. In the user study, we are applying an activity-theoretical approach with the aim of involving researchers, such as me, and investigating how we use – and would like to be able to use – these archival speech resources.

>

[36] http://www.sprakochfolkminnen.se/download/18.46a737b116a496e255833f9/1556021072709/Domeij pres.pdf

Could you describe your research in collaboration with the K-Centre? Have there been any prominent results from this inter-disciplinary collaboration?

<

I am directly involved in case 1: from personal experience narratives to cultural heritage, where I mostly work with a collection of Swedish folklore that was created by Karl Gösta Gilstring, a clergyman and high school teacher who lived in Sweden between 1915 and 1986. Gilstring worked on his collection for more than fifty years, and the result is regarded as the largest folklore collection assembled by a single Nordic researcher in modern times. It consists of more than 8,000 original letters, as well as 250 hours of recordings (mainly interviews conducted by Gilstring himself), from which Gilstring made 70,000 folklore records, divided into approximately one hundred parish collections and organized by subject matter, which aside from folk tales also includes descriptions of rural daily life and traditions.

In our case study, I am interested in establishing the motivations and scientific premises that Gilstring used to create his collection of folktales and to investigate the reasons as to why it has become an integral part of the cultural heritage of Sweden. In the TillTal project, I explore the differences between the unedited audio interviews and his edited written versions that later appeared in the collection. A prominent finding in this respect is that when Gilstring wrote down the folktales he had collected from letters and by conducting oral interviews, he sometimes omitted parts of the story that he felt were his informant's modern interpretations and not part of a "traditional" incarnation of the folktale. This goes to show that cultural heritage is socially constructed, in that Gilstring's rather conservative attitude, which involved a rejection of modern ideals, directly influenced the content of what we nowadays perceive as our "traditional" folklore in Sweden.

The collection is also valuable because of the geographic distribution of the materials. Gilstring's approximately 700 informants not only came from all over Sweden, but also from the Åland Islands and Finland, while around 60 of them were Swedes who had emigrated to America. This is important with respect to the map-based interface Digitalt kulturarv, which the SWELANG K-Centre is developing, since the interface allows me to trace the geographic origins of the letters that were sent to Gilstring by his informants. For instance, I have been able to observe – on the basis of the geocoded information specifying the location of an informant at the time he or she sent the letter – that after emigrating Swedish Americans typically did not stay at a particular place in North America for a long time, but rather moved all over the country, and sometimes even came back to Sweden for a time. Additionally, it was possible for me to observe that the emigrants often presented Sweden in a romanticized manner in their letters to Gilstring, painting the country in broader strokes in comparison to the descriptions in the letters by their compatriots who never left Sweden. This highlights the fact that the ways in which people perceive and remember a particular place (Sweden in the case of the Swedish Americans) are always socially and culturally constructed, and shaped by the individual who reports them.

>

What are the main obstacles of working with audio data? How does the K-Centre help you overcome them?

<

Just recently, I was conducting research on an audio recording that was made with one of Gilstring's informants – a Swedish American called Carl Nelson, who came to America in 1896 when he was 18 years old. What's interesting about the interview is that in certain parts Nelson repeats the same folk stories that he had already described to Gilstring in their previous written correspondences. Additionally, Nelson often jumps from one story to another and then later on returns to comment on a story he's already told. Aside from Nelson's rather messy narration with frequent digressions, the recording is 10 hours long in total, so it took me weeks to go through it. This shows that it is time consuming to analyse audio recordings, so it is incredibly important for me that TillTal gives me the opportunity to collaborate with language technologists like SWELANG's Gunnar Eriksson, who provides me with guidance on the use of automatic speech-analysis methods with which I am able to go back and forth between the different segments of a long audio interview in a time-efficient manner and to interlink them with other related materials in different formats and secondary sources. Indeed, one of the plans of SWELANG is to make available to the research community an environment in which various kinds of materials (e.g., audio recordings, written letters) can be combined so that, for example, dynamic links can be made from a recorded interview to a letter where the same subject or narrative is mentioned twice.

>

As a qualitative researcher, do you think there's any room for improvement in the way data is presented and made available by large-scale research infrastructures?

<

I often feel as though the various domain-specific resources (e.g., historical corpora) available through the repositories are mostly intended for large-scale projects that deal with quantitative "big-data" questions, but it isn't obvious to me how they are suitable for qualitative research. The problem is that many resources contain metadata describing only surface-level features, such as size and linguistic annotation, but lack metadata that are specific to the needs of my field, such as detailed descriptions of the collection process itself, information on who the contributors were in the case of folklore resources, where they came from, when they lived, and so on.

Nowadays, it is easier to get grant money if you propose a humanities project that will – aside from solving research questions that are intrinsic to the field – also involve digitization and collaboration with researchers working in computational fields. While I of course agree that it's extremely valuable to make the data that you're working on accessible in online environments through such collaborations, it often feels as though only the quantity of the data is seen as a measure of success, rather than the presentation of the content of the materials themselves.

I therefore think that it's important for such collaborative projects to re-focus, at least in part, on improving access to and the presentation of the resources that are already available, which is precisely what we are doing in the TillTal project by creating a user-friendly environment for the speech analysis of audio data where the presentation and accessibility of the recordings is tailored to the needs of researchers outside computational fields, like myself.

>

What are the future goals of the TillTal project and the SWELANG K-Centre?

<

One of the future aims of the TillTal project – and by extension the K-Centre – is to increase the amount of available content and bring together related materials (recordings, reports of recordings etc.) through digital methods, which will be done in collaboration with the National Language Bank and SWE-CLARIN. We also plan to release a search system tailored specifically to working with recorded interviews. The system will be accompanied by a tool that will enable us to explore other related non-audio materials while listening to the recording. With this tool, we'll also be able to add additional information about a recording on the fly, such as laughter, or mark sections with fast or otherwise intensive dialogue.

We also plan to develop crowdsourcing tools for transcription and improvement of archive materials, and further work on the mapping interface Digitalt kulturarv, with which researchers will be able to follow audio recordings through time and place, and thereby efficiently study all the documents that were created along the way. In the long run, the plan is to integrate these different technologies in a rich digital tool box, which will offer new possibilities to work with the archival materials of the Institute for Language and Folklore.

>

The National Language Bank and SWE-CLARIN is funded by the Swedish Research Council (2017-00626).

The TalkBank Knowledge Centre

Introduction

Written by **Brian MacWhinney**

TalkBank, which was recognized as a CLARIN Knowledge Centre in 2016, is the world's largest open access integrated repository for spoken language data.[37] It provides language corpora and other audio resources to support researchers in Psychology, Linguistics, Education, Computer Science, and Speech Pathology. The National Institutes of Health and the National Science Foundation have provided support for the construction of five components of TalkBank:

- AphasiaBank for the study of language in aphasia in six languages;
- CHILDES for the study of child language development in 42 languages from infancy to age six;
- FluencyBank for the study of language fluency and disfluency in stuttering, aphasia, second language learning, and normal processing;
- HomeBank for the study through automatic speech recognition of untranscribed daylong recordings in the home and elsewhere; and
- PhonBank for the analysis of children's phonological development in 18 languages.

The five components, which involve multiple corpora collected and encoded according to the same principles contributed by individual researchers from all over the world, form very large collections that are being used extensively to study the cognitive, neurological, developmental, and social bases of language processing and structure. In addition to our support for these five areas, TalkBank also promotes the growth of corpora in nine other related areas:

- ASDBank for the study of language in autism spectrum disorder;
- BilingBank for the study of bilingualism and multilingualism;
- CABank for the study of conversation using the methods of Conversation Analysis;
- ClassBank for the study of language in the classroom;
- DementiaBank for the study of language in dementia;
- RHDBank for the study of language in right hemisphere damage;
- SamtaleBank for the study of conversations in Danish;
- SLABank for the study of second language learning; and
- TBIBank for the study of language in traumatic brain injury.

[37] http://talkbank.org/

TalkBank Principles

The TalkBank system is grounded on six basic principles: maximally open data-sharing, use of the CHAT transcription format, CHAT-consistent software, interoperability, responsivity to research group needs, and adoption of international standards.

Maximally open data-sharing

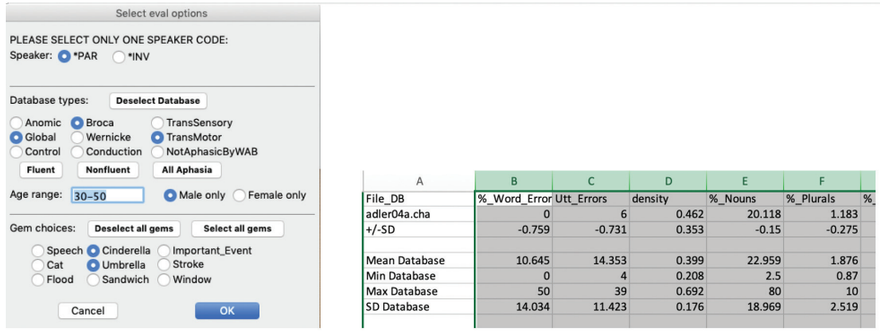
In the physical sciences, the process of data-sharing is taken as a given. However, data-sharing has not yet been adopted as the norm in the Social Sciences and Humanities. This failure to share research results – much of it supported by public funds – represents a big loss to science. Researchers often cite privacy concerns as reasons for not sharing data on spoken interaction. In response to this, TalkBank provides a variety of options in which data can be made available to other researchers, while still preserving participant anonymity, such as password protection and pseudonymization of the participants' first and last names.

CHAT Transcription format

As individual researchers sample from a great diversity of language contexts, they tend to develop idiosyncratic, incompatible methods for language transcription and analysis. In order to provide maximum harmonization across these formats, TalkBank has created an inclusive transcription standard, called CHAT, that recognizes all the features required by different disciplinary analyses. Furthermore, CHAT allows researchers to link transcripts directly to the audio or video, which significantly speeds up transcription and improves accuracy.

CHAT-consistent software

The basic program for analysis of TalkBank data is called CLAN. For language analysis, CLAN automatically computes clinical measures, such as the mean length of the utterance (MLU), the Type-Token Ratio (TRR), Brown's morphemes (for children), and several other values, without errors. Figure 36 illustrates the use of a dialog in CLAN's EVAL program for comparing a transcript from a single participant with those from matched participants in the larger AphasiaBank database.



	A	B	C	D	E	F	
	File_DB	%_Word_Error	Utt_Errors	density	%_Nouns	%_Plurals	%_
	adler04a.cha	0	6	0.462	20.118	1.183	
	+/-SD	-0.759	-0.731	0.353	-0.15	-0.275	
	Mean Database	10.645	14.353	0.399	22.959	1.876	
	Min Database	0	4	0.208	2.5	0.87	
	Max Database	50	39	0.692	80	10	
	SD Database	14.034	11.423	0.176	18.969	2.519	

Figure 36: *A dialog (on the left) in CLAN's EVAL program for comparing a transcript from a single participant (Adler04a) with those from matched participants in the larger AphasiaBank database. On the right is a small segment of the Excel output of the analysis with means and standard deviations.*

Much of the morphosyntactic analysis in CLAN depends on the use of automatic part-of-speech taggers and grammatical dependency taggers that we have constructed for Cantonese, Chinese, Danish, Dutch, English, French, German, Hebrew, Japanese, Italian, and Spanish. The TalkBankDB database search engine permits rapid searches of the database, CQL queries, graphic displays, and downloading of data in CSV format for further statistical analysis. A user-friendly guide for using CLAN that does not presuppose technical knowledge was written by Nan Bernstein Ratner (University of Maryland) and Shelley B. Brundage (George Washington University).

Interoperability

The PhonBank component of TalkBank has developed a separate program called Phon, which provides extensive support for the analysis of phonological data. Crucially, the entire code and functionality of the popular PRAAT software for phonetic transcription are now included inside Phon. Compatibility with other common formats, including Anvil, CONNL, DataVyu, ELAN, EXMARaLDA, LENA, Praat, SRT, SALT, and Transcriber is achieved through translation programs inside CLAN. Recently, Christophe Parisse from INSERM/CNRS and Ortolang, also the repository of the French CLARIN observer, has built a powerful new editor called TRJS, which is used for the transcription, editing and visualization of data and corpora of spoken language, and works directly with the CHAT, ELAN, and TEI formats.

Responsivity to research community needs

TalkBank seeks to be maximally responsive to the needs of individual researchers and their research communities, as well as instructors and clinicians. Our most basic principle is that we attempt to implement all features that are suggested by users in terms of software features, data coverage, documentation, and user support.

Each corpus page includes a link to a facility called the TalkBank browser that allows users to play back linked multimedia corpora directly in their web browser (Figure 37). Users can choose to have continuous playback or playback of specific sections or utterances. For AphasiaBank, FluencyBank, RHDBank, and TBIBank, there are web pages with example videos and instructional commentary designed for use in teaching about language disabilities.

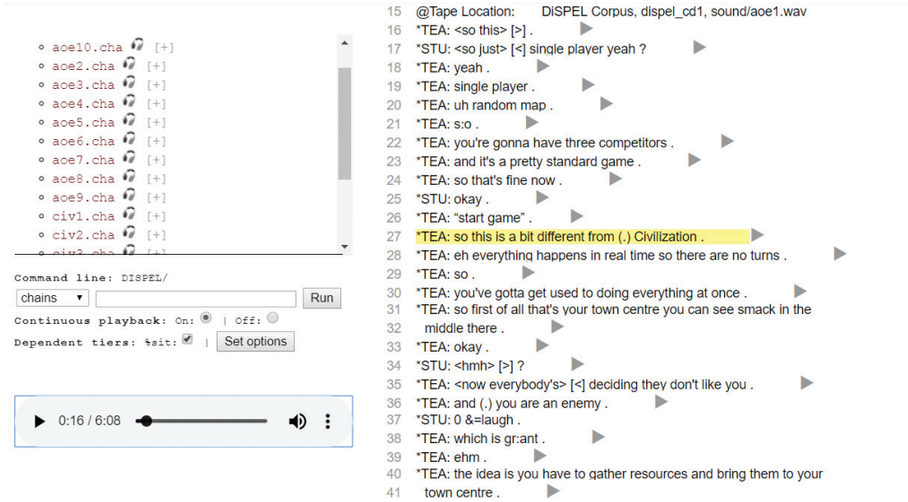


Figure 37: *An audio recording and its assorted transcription in the Dispel Corpus. Using the CLAN editor, the transcriptions have been aligned with the recording, as shown by the yellow highlight. Features inherent to spoken language are also transcribed. For instance, the symbol (.) in the highlighted text stands for a verbal pause made by the speaker.*

TalkBank provides several avenues for user support. In addition to detailed manuals, configuration as a CLARIN Knowledge Centre, and GoogleGroups lists for user support, we have created screencast tutorials that explain how to use the database and the tools. These are hosted both on our own servers and through YouTube. We also conduct presentations and workshops each year at international conferences, such as IASCL, ASHA, LSA, the Academy of Aphasia, LREC, and CLARIN.

International Standards

The sixth basic TalkBank principle is our commitment to international standards for database and language technology. Toward this end, TalkBank has joined the CLARIN federation and is now one of the two members of the CLARIN ERIC infrastructure outside Europe. In 2017, TalkBank received the approval of the Core Trust Seal, which emphasizes the adoption of international standards in data access, protection of confidentiality, organizational infrastructure, data integrity, data storage, data curation, and data preservation. To achieve this, TalkBank maintains incremental GIT repositories for all of its datasets, where researchers interested in replicating earlier analyses can obtain copies of segments of the database from any particular date. In addition, 74 520 language resources in the Virtual Language Observatory (about 10% of all resources listed) derive from TalkBank corpora. Moreover, these resources are all available through open access in a single, consistent, fully documented, and validated format.

Interview | **Nan Bernstein Ratner**

Nan Bernstein Ratner is a Professor at the Department of Hearing and Speech Sciences, University of Maryland, College Park, as well as a Fellow and Honors recipient of the American Speech, Language and Hearing Association. Professor Bernstein Ratner is, along with Brian MacWhinney, one of the PIs of FluencyBank, a shared database for the study of the development of fluency in typical and disordered populations. FluencyBank is part of TalkBank, a CLARIN K-Centre.

Please describe your academic background.

>

I began as a Child Study major at Tufts University, which offered a large number of language classes. After graduation, I originally planned to seek a PhD in Linguistics, my advisor joked that linguists had a hard time finding jobs. She recommended something “applied” that involved language, so I started a federally subsidized MA in speech-language pathology (SLP) from Temple University in Philadelphia. I soon felt that SLPs weren’t making good use of basic language acquisition research. For instance, we were just beginning to explore the ramifications of Roger Brown’s work for clinical practice. Consequently, I decided to do a PhD in Applied Psycholinguistics at Boston University. While at Temple, I wrote an argumentative term paper on why stuttering might be a language disorder with a physiological origin, which turned into a thesis that got published and well-received. But my PhD advisors Paula Menyuk and Jean Berko Gleason, the inventor of the famous Wug Test, still wanted me to pursue first language acquisition and I’ve been a split personality ever since, straddling child language development/disorder and fluency/fluency disorders. Now that I work as Professor at the Department of Hearing and Speech Sciences at the University of Maryland, I am able to combine these interests. As time goes by, they seem less and less separable – fluency and language share interesting interactions.

>

What was the motivation for the FluencyBank project?

>

It is a well-kept secret that even researchers, let alone clinicians, have a lot of trouble accurately transcribing disfluency behaviours like stuttering. What you hear and where you hear it happen can be very variable. Furthermore, fluency researchers were generally very siloed, so there was little collaborative research combining data from different projects. Most of the studies in stuttering also involved too few participants, and there weren't enough longitudinal studies. In response to this, we started the FluencyBank project[38] under the TalkBank initiative because we wanted to make our data available as part of a large-scale interoperable multi-media archive which gives access to utilities specialized for processing audio materials.

There was also a lack of a structured approach to analysing stuttering and related disfluency profiles. Researchers didn't agree on how to code these behaviours, nor were they able to combine their data because everyone made up their own codes for annotation. In this sense, FluencyBank, like the entire TalkBank initiative, was created as an open site where annotation follows a uniform standard to enable multiple data sets to be combined for greater power. Although past work that wasn't consented directly for use in FluencyBank is being kept password-protected and researchers must explain what they want to do with the data to obtain access, we aim to make all the ongoing data contributions open access, which is also in line with TalkBank as a CLARIN K-Centre. All of our teaching materials are open-access now; they are being used across the globe to teach SLP students about the behavioural, affective and cognitive features of stuttering in adults and now children.

>

Could you describe a tool offered by TalkBank that's especially important for your research?

>

The most important tool that TalkBank offers is the transcription program CLAN and its media linkage capacity. Its key advantage is that it offers an easy way to chop up the audio or video signal into very small segments and link them to lines of transcription. Researchers using this program can more reliably process what they have transcribed while listening to the relevant segment.

We think this has real implications for improving the reliability of fluency transcription. For years, I have taught a class of graduate clinicians how to code for stuttering and I would ask my students to transcribe a sample that is available through FluencyBank. Even though the segment is very short, only about 250 words long, my students strongly disagreed on how many stutters or typical disfluencies it contained. Since this sort of disagreement is common among researchers and experienced clinicians as well, we now have a study in progress in which we're trying to compare the accuracy of the CLAN transcriptions with the traditional practice where clinicians simply play the audio and write down their observations. We're doing this to raise awareness as well as to help clinicians do a better job in analysing and understanding their data.

>

How does stuttering differ from other types of disfluency? How can TalkBank help?

>

Generally, articulation and language disorders are there from the very beginning and can be noticed as soon as a child starts speaking. Stuttering, however, is unique in that it seemingly appears out of nowhere in otherwise clinically typical children between the ages of two and four years. This has spurred wide speculation in the literature as to the exact nature of this disorder. For a long time, environmental factors, such as traumatic events, were claimed to precipitate stuttering. For instance, Freud claimed that parents are to blame for stuttering and neo-Freudians promoted the view that children who stutter are suffering from some kind of psychological neurosis, despite the fact there were no data to suggest this was true. Unfortunately, this belief persists in minds of parents world-wide and is difficult to eradicate.

We now know that stuttering has neurophysiological origin and genetic predisposition. Contemporary neurological studies using brain imaging techniques suggest that there's more limited brain connectivity between the regions associated with language planning and motor execution in stutterers compared to typically fluent speakers. The underlying cause of stuttering, however, remains a mystery, so it's valuable to compare it to other forms of disfluency in terms of typology, distributions, and response to linguistic variables, such as the complexity of the intended targets.

TalkBank is an especially good environment for such comparative studies, because the FluencyBank data are interoperable with other similar collections, such as CHILDES and Phon. CLAN offers a wonderful utility called KidEval, which performs a plethora of useful statistical analyses in English and some other languages, such as clause density, counts of important morphemes that are acquired over early childhood and often missing in disordered children's speech, or mean utterance length in morphemes/words, in addition to lexical diversity measures. It then exports the analysis to an Excel spreadsheet and even compares findings to hundreds of children of the same age and sex in the CHILDES Archive. This is important for our work in fluency

[38] https://fluency.talkbank.org/

development and disorder because we now know that linguistic complexity, defined in multiple ways, can impact the fluency of a child's speech. For example, in prior research we have found that it is more likely that someone will stutter on a word like boys than on boy, even though both are phonologically equally complex.

>

What makes the application of language technologies for the analysis of speech challenging for data collection and research, and how do you overcome these challenges in FluencyBank?

>

We would love to be able to automatically differentiate stuttering from the other disfluencies, which is even more challenging in the case of children in comparison to adults, because many children don't show the active struggle in speaking and secondary behaviours that make stuttering in adults so much more obvious. There also aren't any robust pre-existing models of kids' rate and fluency development, and how typically developing children's fluency might be distinguished from that of kids with language impairment (although we have some studies suggesting that kids with language impairment are less fluent than typical kids), kids who are grappling with trying to learn to talk in more than one language, as well as kids who stutter.

It is both tedious and frustrating to document the distributional patterns of fluency in speech samples. Through my career I have repeatedly seen SLPs who make mistakes even just counting the number of words in a read paragraph. However, we have greatly streamlined this process with FluCalc, which is a computational measure in CLAN that gives a detailed breakdown of disfluency behaviours, both over intended words and syllables. Crucially, FluCalc does this by comparing the disfluency behaviours against a weighted score, which on the one hand distinguishes disfluencies that are considered more atypical (i.e., clinically relevant) from those that are considered typical (i.e., disfluency that can be found in otherwise non-disfluent speakers, who may repeat words or phrases when anxious or tired), as well as ranks the atypical disfluencies according to their pathological severity on the basis of a criterion-referenced cut-off point.

For instance, a type of atypical disfluency is the prolongation of a word-initial consonant, such as when a person articulates a word like really as /r-r-r-r-r-eally/, repeating the /r/ sound. FluCal would mark this as more severe than repeating the entire word (really really big), which speakers do all the time in everyday communication when they want to emphasise something. By contrast, blocks are a terrifying form of stuttering where a speaker opens his or her mouth but nothing comes out. A typical speaker would only experience a behaviour like this in a nightmare; thus, they are given higher weight because they would rarely appear in a typical speaker's speech. FluCalc implements a weighted score that examines what types of disfluencies you see in a person's speech, and how many repetitions, or how long a prolongation is, as measures of severity. In the research community there is now an agreement that a child can be considered as stuttering if they receive a weighted score higher than 4% on a speech sample, and FluCalc can calculate this percentage automatically, which is especially important for teachers, clinicians, and doctors.

>

Could you describe some of the recent results achieved in the project?

>

Recently, we teamed up with Purdue University, where Anne Smith and Christine Weber had previously prepared a large-scale longitudinal study in which they followed a large sample of kids who stutter and compared them with their typically fluent peers. Since TalkBank utilities gave us the ability to map multiple language measures easily from the Purdue participants' language samples, we were able to use growth modelling to show that children's expressive language skill was a statistically relevant predictor of recovery from stuttering during early childhood – that is, the better a child's general language skills were, the more it was likely that they would outgrow stuttering on their own over a three-year window of observation (Leech et al. 2019[39]).

It is estimated that 80% of children who start to stutter stop on their own, for reasons we still don't understand well. Our major clinical and research problem is separating those children from those who won't recover and should get therapy early to ensure that the child can learn to speak more easily and not develop handicapping speaking fears. In light of this fact, we are currently working with the Purdue team to determine if other linguistic factors permit us to distinguish between children likely to recover and those who are likely to be persistent. Because the Purdue data are longitudinal, we can do a cross-sectional analysis that will detangle the persistent stutterers, especially given CLAN's ability to link fluency on the speaking tier with grammatical analysis of a dependent tier.

>

[39] https://doi.org/10.1044/2019_JSLHR-S-18-0318

Could you describe the educational component of FluencyBank?

>

Yes, from the very beginning we thought that we would achieve better awareness of the project if we included a teaching component. All the other Banks in TalkBank have teaching resources. We first went to stutterers' support group meetings and asked the attendees if they wanted to participate in a recorded interview that would be transcribed, annotated and put on the FluencyBank page for educational purposes. All of the participants have consented that the interviews – both the videos and the corresponding transcriptions – are made available as open access under *Voices of Adults and Voices of Children Who Stutter and Clutter* categories in the teaching component in FluencyBank. We have standardized these interviews so that the participants are always asked to talk about the impact that stuttering has had on their lives, their experiences with treatment, and to point out those aspects of their disorder that they want clinicians to understand better.

The teaching component has become widely used in education, and I keep getting thanks from professors of stuttering courses about it. The reason for its popularity partly has to do with the fact that more than half of the training programs world-wide lack a resident stuttering "expert", so they mostly have to resort to descriptions in textbooks, which are of course much less illustrative when it comes to explaining the phenomenon or how best to work with clients/patients. We have also designed a set of exercises aimed at university teachers, and we've received positive feedback from various instructors who use the Voices interviews as homework for their graduate students. Additionally, the latest editions of the two most widely used textbooks on stuttering, which are Barry Guitar's Stuttering: *An Integrated Approach to Its Nature and Treatment* and Walter H. Manning's *Clinical Decision Making in Fluency Disorders,* now explicitly mention FluencyBank as a both clinical and research resource.

>

What are your future goals with the project?

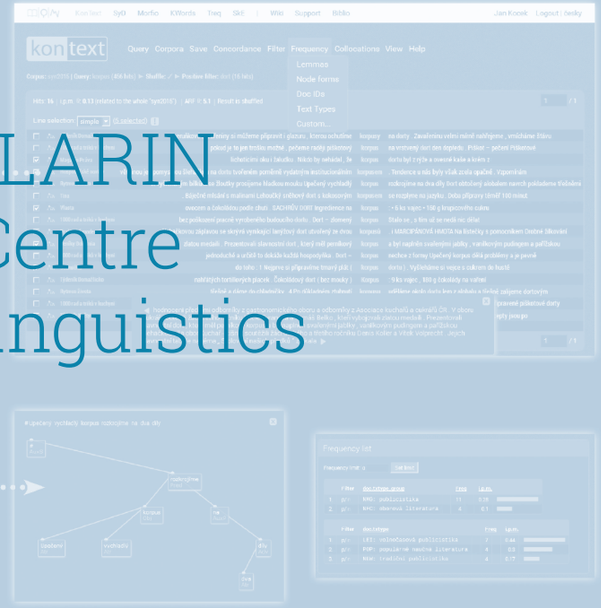
>

We want to get more data. We are already trying to recover and preserve precious data from the "baby boomer" generation of professors who are now retiring. We also want to change the culture of the field to be more like that of child language – that data do more good when shared than when kept close to the vest of their collector. In the case of non-stuttered disfluency, we aim to show that disfluency profiles may inform subtle levels of language impairment or need for remediation that would go undetected by crude language testing, which is known to be non-specific and non-sensitive in identifying older kids with language learning needs. We also seek to show that the elevated disfluency seen in some bilingual children isn't stuttering; it's the natural profile of a child learning to talk in two languages.

For both FluencyBank and CHILDES, we also want to make the research initiative appealing, useful and easy to use for practicing clinicians. Right now, language assessment takes a lot of time and energy – we want to speed it up, make it more informative, and guide more effective therapy goal selection, follow-up and documentation of outcomes. Less time diagnosing the problem and more time available to work towards helping children speak more like their typical peers.

>

The Czech CLARIN Knowledge Centre for Corpus Linguistics



Introduction

Written by **Michal Kren**

Czech CLARIN Knowledge Centre for Corpus Linguistics was recognized by CLARIN on December 4, 2018.[40]

The centre is based at the Institute of the Czech National Corpus, Faculty of Arts, Charles University, Prague. Czech National Corpus (CNC) is a long-term academic project with the main aim to continuously map the Czech language by building, annotating and providing access to a variety of large general-purpose corpora. CNC also develops specialized web-based applications for user-friendly access to the corpora and offers wide-ranging user support which includes a user forum with Q&A, bug reporting, detailed documentation and a knowledge base.

CNC is recognized by the Ministry of Education, Youth and Sports of the Czech Republic as a research infrastructure and included on the Roadmap of Large Research Infrastructures of the Czech Republic for 2016-2022. CNC is an associated member of the CLARIN-CZ consortium with established long-term collaboration with LINDAT/CLARIN. CNC is also a CLARIN FCS endpoint and it supports single sign-on (Shibboleth) as one of the options for accessing the CNC resources. In addition to this service-oriented line of work, CNC is also a research centre that promotes an empirical approach to language and runs a PhD programme in Corpus Linguistics.

[40] http://www.korpus.cz/

The CNC activities can be divided into the three main areas:

- **Data collection.** Focusing on quantity, quality, and variety, the CNC corpora feature careful text selection, reliable annotation and rich metadata. The following areas are currently covered:
  - contemporary written Czech: SYN-series corpora (total size 4.2 billion running words) which also include representative 100-million corpora released every five years;
  - contemporary spoken Czech: corpora consisting solely of spontaneous informal conversation of the ORAL and ORTOFON series (total size 5.3 million running words);
  - InterCorp multilingual parallel corpus: manually aligned and proofread fiction texts supplemented by collections of automatically processed texts from various domains (total size 1.5 billion running words in all 40+ languages);
  - specialized corpora include historical Czech (DIAKORP), Czech dialects (DIALEKT), and many more.

- **Annotation involves data curation,** metadata annotation, morphological tagging and syntactic parsing. For all these procedures, CNC uses open-source software, third-party tools, as well as specialized tools developed in-house. The third-party tools include the Czech morphological lexicon MorfFlex CZ, MorphoDiTa tagger, and Onion deduplication tool, to name just a few. The tools developed in-house include mainly the Phras module for identification of idioms, Mluvka for management of distributed spoken data collection, and the parallel text alignment editor InterText.

- **User application development.** We recognize the key importance of presenting corpora in an intuitive way that makes them accessible to researchers from various fields of social sciences and humanities. The following web applications are currently offered:
  - KonText – a general-purpose corpus query interface and concordancer with an advanced subcorpus manager, parallel corpus support and support for word-to-sound alignment;
  - SyD – corpus-based analysis of language variants, both synchronic and diachronic;
  - Morfio – identification of derivational models in Czech including estimation of their morphological productivity;
  - KWords – corpus-based keyword analysis;
  - Treq – translation equivalent search interface based on the InterCorp parallel corpus.

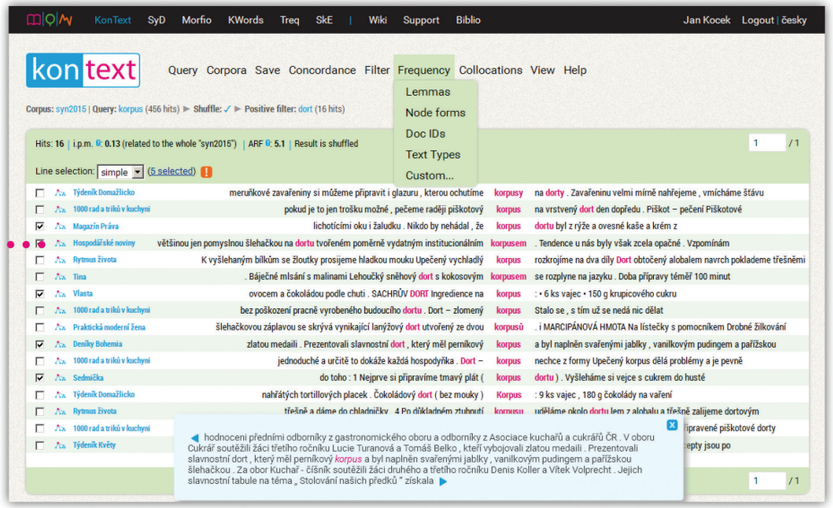


Figure 38: *KonText user application being actively developed by the CNC*

Currently, CNC has 7,500+ registered active users who perform (on average) 3,000+ corpus queries per day. The repository of CNC-based research outputs has yielded more than 150 theses (bachelor, master or doctoral) defended a year.

The CNC user support and related services are available also through the CLARIN K-Centre. This includes:

- K-Centre Helpdesk and CNC User Forum, virtual platforms for active user support and feedback. The CNC user forum features also a discussion forum (with Q&A) that can handle requests for new application features as well as bug reports.
- Documentation and knowledge base for CNC applications, data and services. It provides interdisciplinary guidance and promotes empirical methods in language research. It also features an online tutorial aimed at both beginners and advanced users.
- Repository of CNC-based resources and research output that can also serve as a bibliography for looking up information concerning the corpus research on Czech.
- Repository of corpus-based exercises (Czech only) for L1 and L2 language teaching.
- Consulting, education, and training: in addition to the general user workshop held on a regular basis twice a year, workshops on various topics are held upon request.
- Corpus hosting: the service includes final technical processing of user-compiled corpora, quality checks, and public access with related services.
- Customized data packages: data sets prepared on demand and extracted from the CNC corpora while observing the legal limitations that may not allow for redistribution of the texts per se.

Our expertise includes not only data formats, text curation, annotation, metadata encoding and corpus querying, but also empirical research on the Czech language, corpus linguistics methodology and statistical methods. The centre can also provide external pointers to other institutions regarding any aspect of Czech language including language resources and natural language processing.

CNC User Workshop

Interview | **Ondřej Tichý**

Ondřej Tichý is a corpus linguist who is deputy chair of the Department of English Linguistics at the Faculty of Arts at Charles University. Dr Tichý collaborates with and is a regular user of the Czech National Corpus.

Please describe your academic background and current position. What inspired you to take a digital humanist approach to linguistics?

>

I earned my PhD in English Linguistics at the Faculty of Arts, Charles University in 2014. I have been teaching and conducting research at the same faculty since 2008, specializing in historical and corpus linguistics, quantitative and computational linguistics, digitization and digital humanities. Between 2014 and 2018, I served as a vice-dean for information resources and since 2018 I have been the deputy head of the Department of English Linguistics. Parallel to my academic carrier, I have been working in IT since late 1990s and it has been primarily due to my background in IT and my academic interests in diachronic linguistics that I took the digital approach.

Another motivation for my involvement in the digital approach was to make important resources, that I used for my own research, available to the wider public as well, resulting in the digitization of an Anglo-Saxon dictionary for my MA thesis and then conducting automatic analysis of Old English morphology for my PhD. Finally, the projects based on the Helsinki corpora that were compiled when corpus linguistics started to emerge as one of the major linguistic diachronic methodologies in the 1990s have been very inspiring to me from the very beginning.

>

What is your involvement with the CNC K-Centre?

>

I am both a dedicated user of their infrastructure and a collaborating researcher. I have been invited by the centre to give talks on diachronic corpus topics (for instance, on lexical obsolescence in Late Modern English or on the quantification of orthographical variation in Early Modern English, which are two of my current research interests), I have consulted on some of these projects with a number of colleagues at the centre and I hope our fruitful collaboration to continue in the future as well. But mainly, I use the centre's infrastructure, tools and expertise to host and analyse corpora I need for my own projects. Many of these corpora are not in the public domain (either by the decision of their compilers or due to the licensing restrictions of their source material) and are only hosted for licensed users for research and teaching, but in cooperation with the centre we have also started publicly hosting data from the Early English Books Online (EEBO) project, and are about to host the Old Bailey Corpus, which is based on a selection of the Proceedings of the Old Bailey, the published version of the trials at London's Central Criminal Court.

>

Which data collections in CNC do you use in your own research? Could you present and discuss some of your research that has resulted from your use of the CNC corpora?

>

I mostly use English diachronic corpora that the centre specifically processed and hosts for our department and students, but I have also used the DiaKorp, InterCorp and the SYN corpora for a contrastive angle.

One example of the research I do using the centre's infrastructure is my recent work on spelling variation in Early Modern English based on the Parsed Corpus of Early English Correspondence. I introduced a novel methodology for the quantification of spelling regularity, which allowed a more objective assessment of its progression in time and which also makes use of the metadata provided by the CEEC such as gender, letter authenticity or relationship/kinship between the author and the recipient. I have explored interactions of such variables from the diachronic perspective using quantified levels of spelling regularity.

The measure introduced for this purpose is based on weighted information (Shannon) entropy, as a measure of predictability of a spelling of individual functionally defined types, and its calculation is partly based on the morphological tagging of the parsed version of the corpus.

I have also tackled the problem of underrepresentation in certain periods by establishing a size-based sampling for scalar variables like time. For instance, I was able to show that letters written by women showed a greater degree of entropy – so a greater degree of variability – in spelling regularity than letters written by men through the whole period (roughly from 1410 to 1680). However, this difference turned out to be a function of another sociolinguistic variable that I was accounting for besides the author's gender; namely, the relationship between the author and the recipient. Female authors corresponded significantly more with other members in the family than male authors who mostly corresponded with acquaintances outside the family. In a familial context, there might be less pressure to conform with spelling standards, hence the greater degree of variation.

Another example is an older study on measuring the typological change in English that was based on the parsed versions of the Helsinki corpora. In this paper [41], my colleague Jan Čermák and I proposed a quantitative, but also holistic, methodology for establishing the level of morphological syntheticity within a language – that is, how much a language relies on morphological markings to convey syntactic information. The methodology is based on a series of corpus-based probes into the morphological behaviour of selected high-frequency nouns, adjectives and verbs from Old English to Present-Day English in corpora hosted by the CNC. We thereby managed to establish several levels of syntheticity that correspond to the well-known typological re-shaping that happened in the history of English, which shifted from a heavily synthetic language in its early days to an analytic one in the present day. For instance, Old English was highly synthetic, its nouns ending in seven different inflections corresponding to the complex case system, whereas Present-Day English nouns only use the -s affix to mark plurality, and our proposed methodology was able to capture this quite precisely.
It should be also noted that CNC often consults and helps out indirectly, not with their corpora or tools, but with their scientific and technical expertise. For example, in my research into the obsolescence of multi-word expressions in the history of English, it was only thanks to a colleague at CNC and the centre's computing resources that I was able to pre-process most of the Google Ngram dataset (about 2 terabytes of data).

>

[41] https://doi.org/10.3726/978-3-0351-0640-4%2F17

Which challenges does one face when doing diachronic linguistics with corpora? Do CNC corpora employ any features that are specifically tailored to diachronic analysis? Is there any additional feature that you would like to see implemented in the future?

>

The specific challenges of diachronic corpus linguistics are numerous. Those that often trouble me are the scarcity of data coupled with their representativeness, the quality of the data and, in the case of English, the formal variation that can be found on almost all levels of linguistic description. Such variation is more often than not problematic for tools that are geared for the analysis of Present-Day English. The CNC tools (rather than corpora), while not specifically tailored towards diachronic analysis (except perhaps for SyD), do however yield to it quite well. I am very happy with KonText[42] and how our colleagues at the CNC are both able and willing to tweak it to make things work for specialized users, especially the treatment of metadata and the ways these can be analysed and searched seem better to me than in, for instance, the CQP web or SketchEngine.

Another advantage we are just going to make use of is the possibility to analyse metadata at utterance level, which means that we will associate metadata with parts of texts rather than with entire texts only. As an example: a user can start by limiting the query (search for a particular form/function) by the gender of the speaker or a specific timeframe, then view the frequency analysis based on the properties of the text containing the direct speech (e.g. by the type of offence in trial proceedings) and finally create a table interrelating two attributes (like social class of the speaker and the orthography of the keyword). This makes corpora like the Old Bailey Corpus much more approachable to less experienced users, since they do not have to overcome the steep learning curve of CQL or similar query languages and can also see some of their results in a neat tabular format without the need to export the results and run a statistical tool on them. It should also be noted that while many similar features may be available in similar tools, KonText is open source and free to use.

>

What are the main benefits of the KonText search interface? Do you use any of the other CNC tools, such as SyD, Morfio, in your work?

>

In my research, I mainly use KonText and recently the brand-new Corpus Calculator. I use SyD for teaching – as a tool roughly comparable to Google Ngram Viewer – since it provides a very user-friendly way to compare lexemes across the CNC corpora both synchronically and diachronically.

[42] https://kontext.korpus.cz/

As I noted in my previous answer, I like KonText because it allows me to quickly search and analyse metadata. I like to focus on social aspects of language changes. I also like the CQL since it is easy to teach and learn. Furthermore, it is very well documented in the CNC Wiki and is very similar to query languages used in other concordancers. From a teacher's perspective, CQL and other search options in KonText make it easy to start with and yet are very powerful at the same time.

>

What kind of feedback have you provided on the CNC corpora and its user interface? What is your experience with the CNC User Forum? Why is it important for the CNC K-Centre to offer such user support?

>

Since CNC often accommodates me by hosting all kinds of corpora that tend to be different than the Czech corpora they are predominantly focused on, I often request changes or new features – mostly by e-mail to specific colleagues but also through GitHub. While the CNC may not always immediately implement all my outlandish ideas it has in general been very forthcoming about my requests. Here[43] is one example, where I requested that headers be added to the .csv and .xlsx files exported from KonText, and the CNC team quickly implemented the change.

>

How do you use the CNC corpora in your teaching? Have your students obtained any interesting results from the CNC corpora?

>

I use KonText in most of my classes focused on the History of English to showcase specific changes, and I also teach how to use the interface in my English Diachronic Corpora course. Almost all of the students at our department learn to use KonText and InterCorp, and the majority of theses in our linguistic programmes are corpus-based, so most of the final theses (several dozen a year) are based on CNC-hosted corpora.

A lot of the theses are based on the contrastive approach focusing on features of Present-Day English and Czech, but there have been a number of diachronic theses and papers as well. One of my PhD students is now working with the CNC-hosted EEBO data to research lexical losses in Early Modern English, another of our PhD students is developing a parallel corpus of Old English and Latin translations that will again be hosted by CNC that has already extended its support in this. Finally, one of my PhD students prepared lessons in English available on the CNC wiki for using the diachronic EEBO corpus, which show how KonText can be used to account for spelling variation, looking at diachronically competing word forms, analysing morphology, among other uses. We hope that some of our students will develop a similar online course for the Old Bailey Corpus.

>

[43] https://github.com/czcorpus/kontext/issues/2038

Notes:

COLOPHON

Coordinated by
Darja Fišer and **Jakob Lenardič**

Edited by
Darja Fišer and **Jakob Lenardič**

Proofread by
Paul Steed

Designed by
Tanja Radež

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

Publication number
CLARIN-CE-2019-1537
November 2019

ISBN
9789082990911

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

www.clarin.eu

CLARIN
Common Language Resources and Technology Infrastructure

CLARIN 2019