CLARIN

Common Language Resources and Technology Infrastructure

Tour de CLARIN

VOLUME **THREE**

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

Common Language Resources and Technology Infrastructure

Tour de CLARIN

VOLUME **THREE**

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. In 2019, we expanded the initiative to also feature the work of CLARIN Knowledge Centres, which offer knowledge and expertise in specific areas provide to researchers, educators and developers alike. Initially conceived as a series of blog posts published on the CLARIN website, Tour de CLARIN soon proved to be one of our flagship outreach initiatives, which has been released in the form of two printed volumes.

This third volume of Tour de CLARIN is organized into two parts. In Part 1, we present the six countries which have been featured since January 2020: Norway, the United Kingdom, Germany, France, South Africa, and Iceland. Each national consortium is presented with five chapters: an introduction to the consortium, their members and their work; a description of one of their key resources; the 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 work.

In Part 2, we present the work of the six Knowledge Centres that have been visited since the publication of the second volume in November 2019: the Impact-CKC K-Centre, the Knowledge Centre for Polish Language Technology, the Phonogrammarchiv Knowledge Centre, the Knowledge Centre for Atypical Communication Expertise, the LUND University Humanities Lab Knowledge Centre, and the Spanish Knowledge Centre. Each Knowledge Centre is presented with two chapters: a presentation of what the K-centre offers to researchers, and an interview with a renowned researcher who has collaborated with the K-centre.

The creation of this volume has been affected by the COVID-19 pandemic, which put a lot of additional burden on our entire community both at work and at home, but especially colleagues who were involved in teaching, and those with organizational and managerial responsibilities. This is why we feel all the more grateful that we were able to complete the project on time and present such a rich collection of inspiring stories from the CLARIN community.

This volume would not have been possible without the contributions and dedication of the national user involvement representatives and national coordinators: Koenraad de Smedt and Kristin Hagen from Norway, Martin Wynne from the United Kingdom, Nathalie Walker, Melanie Grumt Suárez, Thorsten Trippel, and Erhard Hinrichs from Germany, Nicolas Larousse, Christophe Parisse, and Francesca Frontini from France, Eiríkur Rögnvaldsson from Iceland, and Liané van den Bergh, Juan Steyn, and Langa Khumalo from South Africa.

We are equally grateful for the contributions by the Knowledge Centre coordinators: Isabel Martínez-Sempere from the IMPACT-CKC K-Centre, Jan Wieczorek from the Knowledge Centre for Polish Language Technology, Kerstin Klenke from the Phonogrammarchiv, Henk van den Heuvel from the Knowledge Centre for Atypical Communication Expertise, Johan Frid from the LUND University Humanities Lab Knowledge Centre, and Mikel Iruskieta from the Spanish CLARIN Knowledge Centre.

We would also like to thank all the researchers who have kindly agreed to be interviewed for their time and invaluable insights: Yvonne van Baal, Michaela Mahlberg, Eva Gredel, Yana Strakatova, Amalia Todirascu, Menno van Zaanen, Jóhannes Gísli Jónsson, Mikel Iruskieta, Dominika Hadro, Beate Eder-Jordan, Katarzyna Klessa, Anita Lorenc, Gerd Carling, and Jose Pérez-Navarro.

Darja Fišer and **Jakob Lenardič**

Ljubljana, Slovenia
November 2020

Consortia featured in this volume:

Norway

The United Kingdom

Germany

Iceland

France

South Africa

PART 1
CONSORTIA

NORWAY

Introduction

Written by **Koenraad de Smedt** and **Kristin Hagen**

Norway joined CLARIN ERIC in 2015 after having been an observer since 2013. The Norwegian national infrastructure for language resources and technology is called CLARINO.[1] Its five-year construction phase has been funded through the eponymous national project since 2012, and the infrastructure is currently in full operation. From 2020 the infrastructure will enter a three-year upgrade phase, which is also be funded nationally.

The Norwegian national coordinator is Prof. Koenraad De Smedt at the University of Bergen. Other institutions that currently comprise the Norwegian CLARIN consortium are the following:

- University of Oslo (UiO)
- Arctic University of Norway (UiT)
- Norwegian School of Economics (NHH)
- Norwegian University of Science and Technology (NTNU)
- Uni Research Computing (UNI)
- National Library of Norway (NB)

In the upgrade phase, the Norwegian Centre for Research Data (NSD) will also join the consortium.

[1] http://clarin.w.uib.no/

CLARINO offers its infrastructure services through four centres.

The CLARINO Bergen Centre at the University of Bergen (UiB), in cooperation with the Norwegian School of Economics (NHH), has been certified as a CLARIN B-centre. The centre is operated by the Department of Linguistic, Literary and Aesthetic studies (LLE) and the University Library (UBB). It offers a repository based on CLARIN DSpace with LINDAT extensions, where researchers can download and upload digital language datasets. The centre also offers the following online tools: the INESS treebanking platform with treebanks for more than 70 languages, the Corpuscle corpus exploration system providing access to corpora currently covering 18 languages, and the COMEDI metadata editor which is in use worldwide. These systems run on a dedicated high-performance computer. The centre will also operate the Terminology Portal after its migration from Oslo.

Språkbanken, the Language Technology Resource Collection for Norwegian at the National Library of Norway (NB), has been certified as a CLARIN C-centre. Its mission is to provide language resources, primarily for Norwegian, that are not only useful to academic researchers but also suitable for research and development in applied language technology. NB also provides an online n-gram counter and exploration tools for its digitized library holdings. NB maintains the CLARINO National Metadata Registry (CNMR), a national catalogue of language resources based on metadata harvested from all CLARINO nodes.

The Text Laboratory at the University of Oslo (UiO) has been certified as a CLARIN C-centre. Through the centre, a wide range of corpora are available via the corpus exploration system Glossa (33 corpora for several different languages, including several African languages). Glossa supports CLARIN Federated Content Search (FCS). The corpus system Glossa itself is also downloadable, together with the Oslo-Bergen Tagger for morphosyntactic analysis and annotation of written Norwegian. A variety of databases and word lists based on frequency and text genre are also offered. Other units at UiO have developed an Interactive Dynamic Presentation (IDP) system for digital editions, and the CLARINO Language Analysis Portal that provides workflows for language analysis.

TROLLing, the Tromsø Repository of Language and Linguistics at The Arctic University of Norway (UiT), has been certified as a CLARIN C-centre. Based on Dataverse, the repository specializes in packages of replication-enabling datasets bundled with statistical code, presentations and other documentation, including scientific articles. UiT also hosts Giellatekno, which has datasets and tools on Sami and other Arctic minority languages.

CLARINO has two certified Knowledge Centres:

1. The CLARINO Bergen Centre has, together with the CLARIN/LINDAT Centre in Prague, established a K-centre for treebanking, which was featured in Tour de CLARIN Volume II.
2. The Norwegian Centre for Research Data (NSD) operates a K-centre on data management (including legal issues).

A recent Norwegian Parliamentary Paper on the Humanities describes CLARINO as the common infrastructure for language databases in Norway, having an impact primarily on the Language Sciences, but also enabling substantial research potential in other SSH disciplines, as well as in industrial research and development, for instance through multilingual technologies.

From 2020, the three-year CLARINO+ project will upgrade the technical infrastructure, promote uptake by researchers and further sustainability. As a result, the upgraded CLARINO will be a more up-to-date, dynamic and user-centred infrastructure providing better services for more language data stemming from past, present and future research. It will be better adapted to new state-of-the-art CLARIN core services and standards, and will better serve an extended target audience based on a sustainable business model. An updated portal will give easier access to the distributed services located in the four Norwegian centres and to central CLARIN services.

Figure 1: The Text Laboratory group (Anders Nøklestad, Joel Priestley, Janne Bondi Johannessen and Kristin Hagen) at LREC2016 in Portorož, Slovenia.

Tool | **Glossa: a User-Friendly Corpus Search System**

Written by **Janne Bondi Johannessen, Kristin Hagen, Anders Nøklestad** and **Joel Priestley**

The Glossa search system is an advanced, user-friendly interface for searching in written text and transcripts of audio or video.[2] Glossa was developed at the Text Laboratory at the University of Oslo, and reimplemented in 2014–2018 in the CLARINO project. Glossa is used to query a series of speech, text and parallel corpora available at CLARINO Text Laboratory Centre. The Glossa tool itself is also downloadable from the CLARINO Text Laboratory Centre site and from GitHub. Glossa allows researchers to search for a word, a part of a word, individual sounds, or several words and to search for grammatical information, as well as to filter the results on the basis of various sociolinguistic metadata, such as the gender and age of the informant. The search results are presented as concordances, as shown in the figures below, but can also be rendered as PDFs, plotted on maps (via geographical coordinates and Google Maps), or shown as frequency lists. The data can be automatically translated, for most corpora via the Google Translate API, and also downloaded in a spreadsheet format. For speech corpora, the integrated media player provides access to the corresponding media sequence of each query result.

Glossa supports login with eduGAIN, CLARIN and Feide (the latter for Norwegian users). The Glossa installation is based on widely available technologies such as Java, MySQL and IMS Corpus Workbench. Search and result processing functions are accessed using familiar browser buttons and menus, requiring no special technological background. For those familiar with the IMS Corpus Workbench, advanced CQP queries can be performed directly using a specific search box. Glossa also supports CLARIN Federated Content Search (FCS).

The figures below show Glossa being used with the speech corpus CANS – Corpus of American Nordic Speech and the NORM Corpus, a learner corpus of more than 5,000 pupil essays.

[2] https://www.hf.uio.no/iln/english/about/organization/text-laboratory/services/old-glossa/index.html

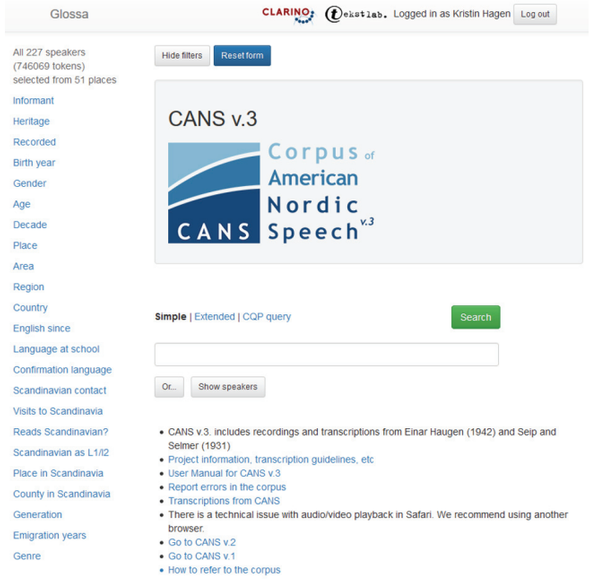


Figure 2: The main search page of CANS. To the left are all the searchable metadata categories with the number of selected speakers and tokens at the top. The *Show speakers* button will show the speakers that are currently selected. Below the corpus logo, there is a search box with three different search alternatives: *Simple,* which allows a search for a simple word or phrase, *Extended* (see Figure 3), and *CQP query.*

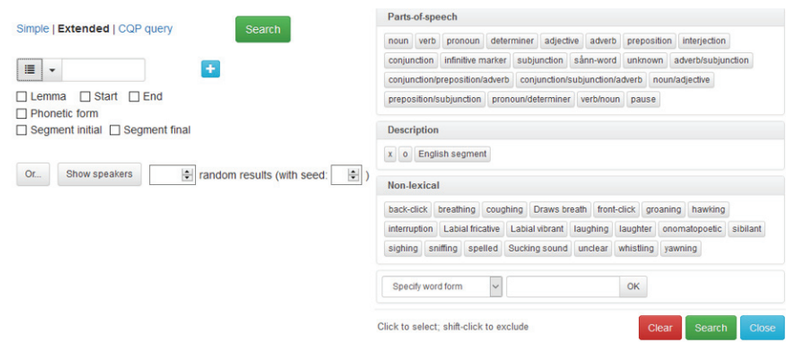


Figure 3: The *Extended* search interface. In addition to the various search term modifiers available via the checkboxes, clicking the menu icon reveals a list of additional attributes, such as part of speech. The plus icon is used to extend the search term with additional text boxes, whereas the *Or* button yields further text boxes for or-searches (disjunction). Note that you can tick the *Phonetic form* box if you want to search directly in the phonetic transcriptions of CANS (CANS has two aligned transcriptions of each sequence, one orthographic and one phonetic).

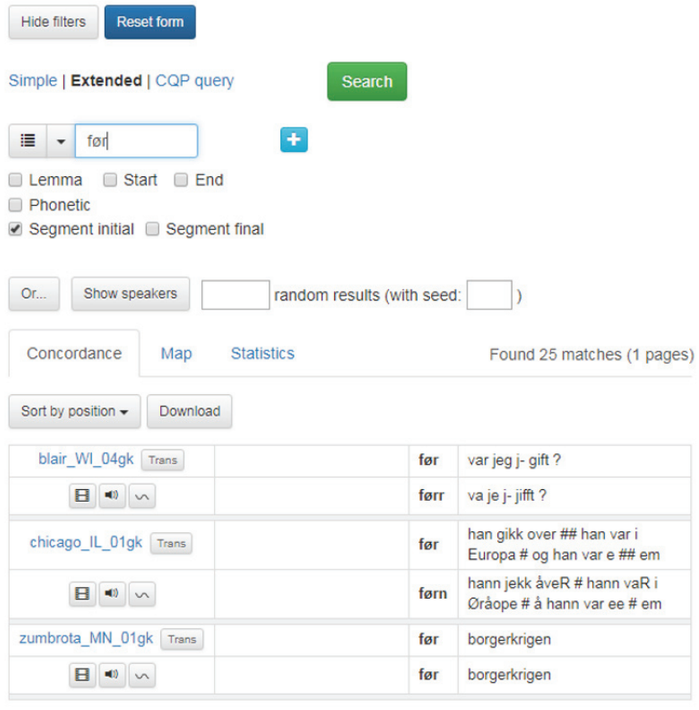


Figure 4: A search for *før* (“before”) in segment-initial position returns 25 matches. Both the orthographic and the phonetic transcription of the sequence are shown, together with options to see the metadata pertaining to the speaker, to translate, to play video or audio, or to see a spectrogram of the sequence.

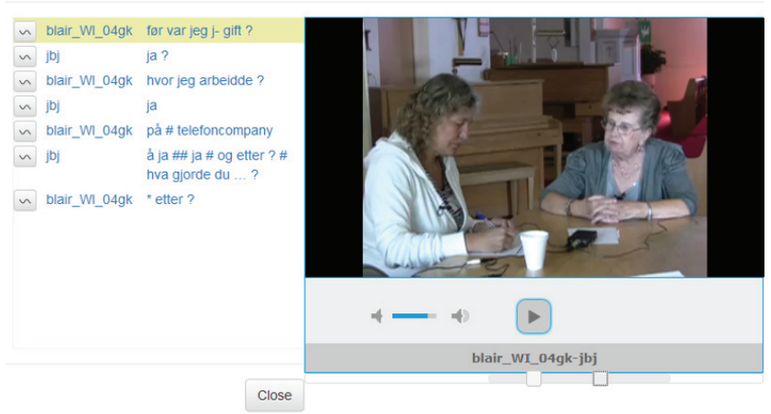


Figure 5: Video playback of the search result. More context can be accessed by moving the square buttons on the slider bar.

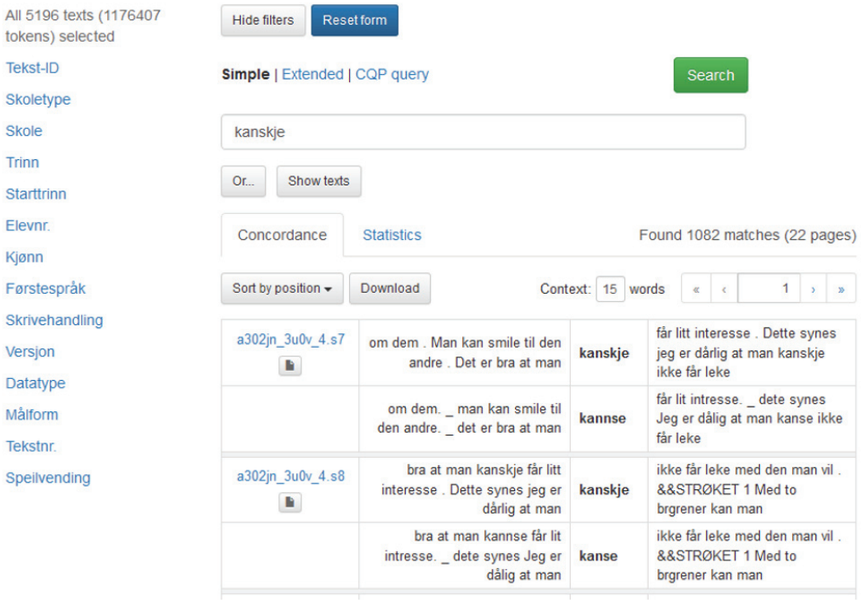


Figure 6: A search result for *kanskje* (“maybe”) in the pupils’ text corpus NORM. The search interface is the same as in the speech corpus CANS, but the metadata menu on the left is different. If you click on the document symbol to the left of the search result, you will see a PDF of the pupil’s text as shown in Figure 7. Note that this corpus has two aligned versions of the text, the original and corrected ones. You can search in both versions, although the orthographic search is the default.

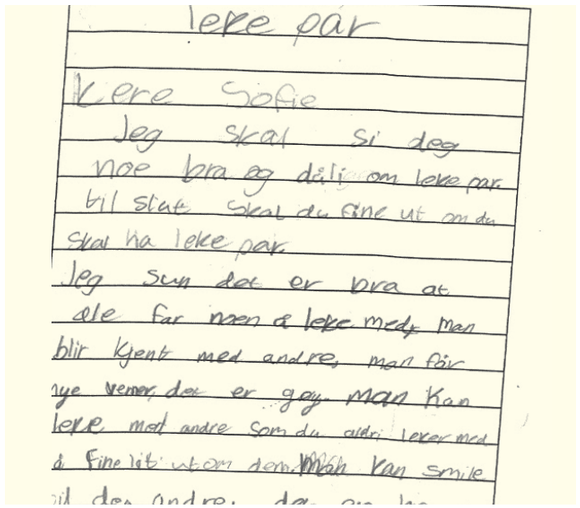


Figure 7: An essay in the NORM corpus.

Glossa is an essential search system for researchers working with language variation and change, since it allows them to compare language use across different age groups, different periods (there are nearly a hundred years between the oldest and the newest recordings in some of the speech corpora available through Glossa) and different locations. The dialect corpora and the heritage language corpora, such as the Corpus of American Nordic Speech v.3 and the Nordic Dialect Corpus v. 4.0, are often used in this kind of research and queried with Glossa. The written language corpora, such as the NORINT Corpus, which contains the use of Norwegian as a second language, and the Lexicographic Corpus for Norwegian Bokmål, have been successfully used for practical work on dictionaries and language planning and standardization, where it is vital that different kinds of texts can be distinguished easily on the basis of the sociolinguistic metadata filter options.

References:

Nøklestad, A., Hagen, K., Johannessen, J.B., Kosek, M., and Priestley, J. 2017. A modernized version of the Glossa corpus search system. In *Proceedings of the 21st Nordic Conference on Computational Linguistics (NoDaLiDa)*, 251–254.

Kosek, M., Nøklestad, A., Priestley, J., Hagen, K., and Johannessen, J.B. 2015. Visualisation in speech corpora: maps and waves in the Glossa system. In G. Grigonytė, S. Clematide, A. Utka and M. Volk (eds.): *Visualisation in speech corpora: maps and waves in the Glossa system, Proceedings of the Workshop on Innovative Corpus Query and Visualization Tools at NODALIDA 2015*, 23–31.

Resource | **The Nordic Dialect Corpus**

Written by **Janne Bondi Johannessen, Kristin Hagen, Anders Nøklestad** and **Joel Priestley**

The Nordic Dialect Corpus (NDC) is a speech corpus available at the CLARINO Text Laboratory Centre.[3]

NDC is a corpus of Danish, Faroese, Icelandic, Norwegian and Swedish (including Övdalian) spoken languages. The corpus consists of spontaneous speech data from dialects of the North Germanic languages across all Nordic countries. The recordings and transcriptions in the corpus are part-of-speech tagged and come from various sources. The Danish, Icelandic and Norwegian recordings were collected with the financial support of individual national research councils; the Faroese recordings were added by a cross-Scandinavian research project, while the Swedish recordings were mainly donated to the corpus from SweDia 2000, a previous dialect collection project.

Country	Informants	Places	Tokens
Denmark	81	15	220,360
Faroe	20	5	64,803
Iceland	48	8	94,338
Norway	438	111	1,997,920
Sweden (incl. Övdalian)	150	44	376,868

Figure 8: The recordings and transcriptions in NDC.

NDC contains over 2.75 million words from conversations and interviews done from 1998 to 2015, see Figure 8. Older Norwegian recordings can be found in the corpus LIA Norwegian – Corpus of Old Dialect Recordings, also available through the CLARINO Text Laboratory Centre. The informants, whose recorded speech constitutes the NDC corpus, were instructed to avoid interpersonal topics due to information privacy laws. This is why they mostly focused on topics like holidays, school, life in the old days, and factual, non-private mentions of individuals, while sensitive topics like politics are not addressed.

[3] http://www.tekstlab.uio.no/nota/scandiasyn/

Since the recordings in NDC are classified as personal data, the corpus is accessible only via Glossa, a search tool developed at the Text Laboratory and reimplemented in the CLARINO project. The corpus has three login possibilities: eduGAIN, CLARIN and Feide (for Norwegian users).

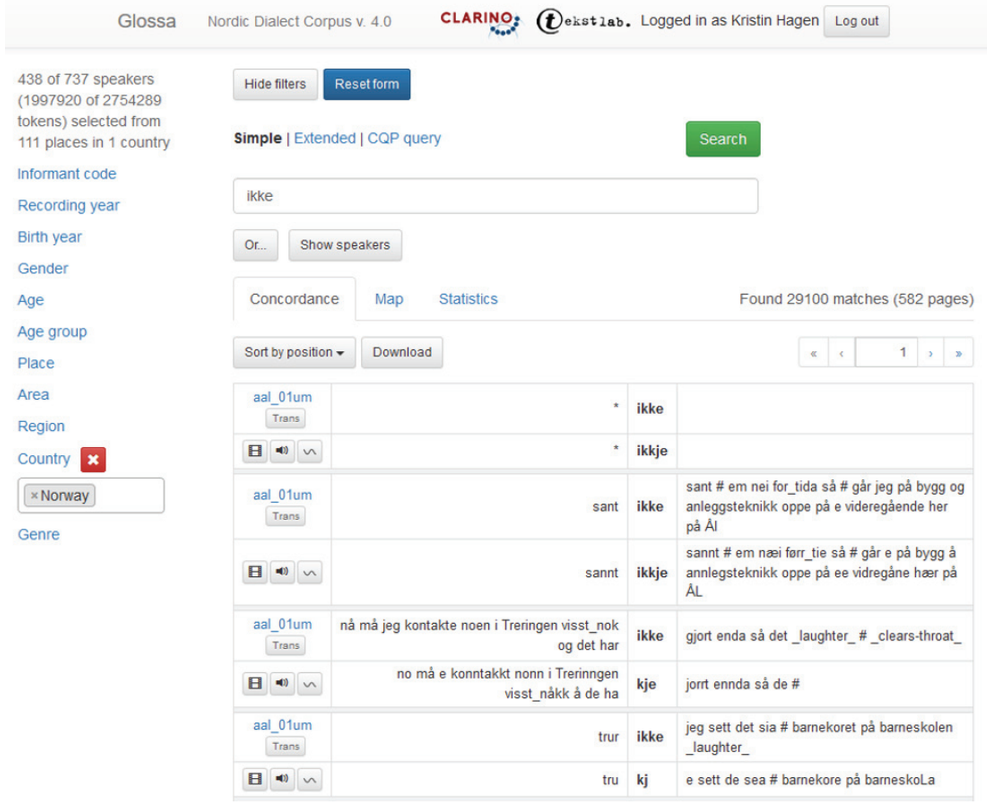


Figure 9: A simple search in the Norwegian part of the dialect corpus for *ikke* (“not”) returns 29,100 matches. The Norwegian and Övdalian parts of the corpus have two aligned transcriptions, one phonetic and one orthographic. Both are included in the concordance view shown in the figure. (The other languages have one orthographic transcription for each recording.)

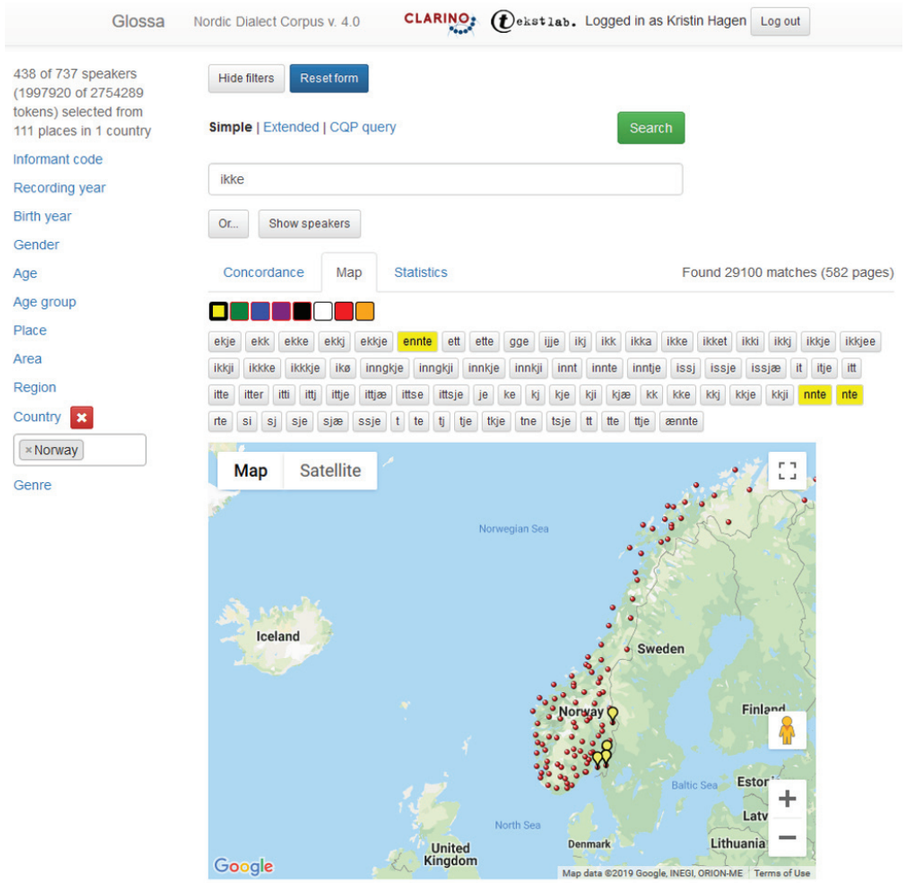


Figure 10: The map view can show the distribution of the phonetic variants of the search, with this figure showing all the variants of the Norwegian word *ikke* ("not"). By clicking on a colour and then on a variant, you can see the distribution of the chosen variant on the map. In the figure above, the dialectal variants *ennte, ente, nnte,* and *nte* are shown in yellow.

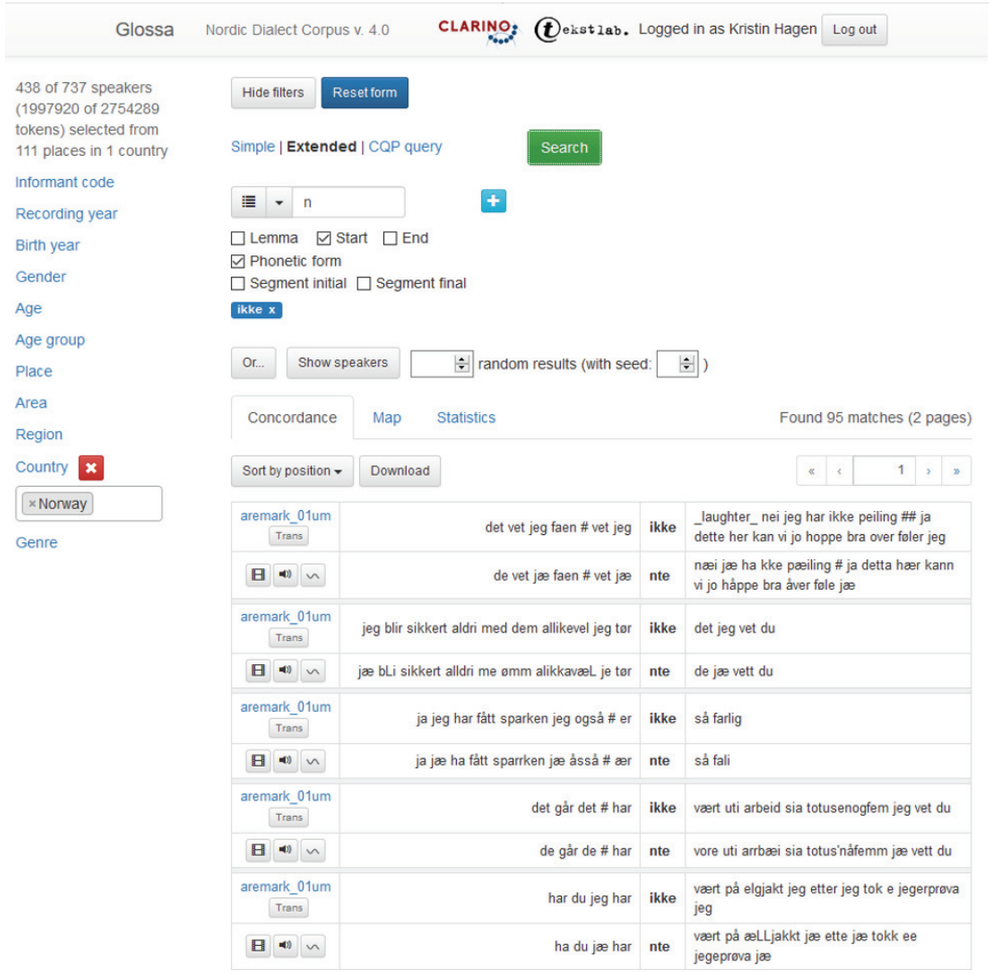


Figure 11: The search interface of the corpus is easy to use, with user-friendly buttons and menus, while the search possibilities are still advanced. The search shown above is an *Extended* search for words starting with n– in phonetic form, where the orthographic word form is *ikke.* The orthographic form is chosen by clicking on the menu button to the left of the search word and filling in the box at the bottom, see Figure 12.

Figure 12: The parts-of-speech menu in NDC with the *Specify* or *Exclude* box at the bottom. All transcriptions in NDC are morphologically tagged and searchable for all languages.

NDC has been frequently used in a wide range of research areas, such as phonology, morphology, syntax and lexicography. Janne Johannessen and Kristin Hagen published a monograph called *Språk i Norge og nabolanda: Ny forskning om talespråk* ("Language in Norway and Neighbouring Countries: New Research on Spoken Language"), which presents linguistic research carried out on the basis of the corpus. The monograph includes topics such as the syntactic structure of adjectival complements in Norwegian dialects, the use of possessive markers, the order of particles and objects in Norwegian and Swedish dialects, and the syntax of questions in Norwegian dialects.

The open access journal, *Nordic Atlas of Language Structure (NALS) Journal,* actively encourages linguists to publish empirical research on geographical linguistic variation that is observed on the basis of data from NDC, as well as from other dialect resources. For instance, Vangsnes (2014) used NDC to explore variation in the syntax of noun phrases, focusing on the marking of definiteness and the syntactic features of possessors, while Garbacz (2014) explored the non-standard use of definite articles in indefinite contexts in Swedish, Fenno-Swedish and Norwegian dialects. Finally, the NDC corpus has been used as the empirical basis for a number of MA and PhD theses at universities in Nordic countries.

References:

Garbacz, P. 2014. Definite articles in indefinite contexts. *The Nordic Atlas of Language Structures (NALS) Journal* 1 (1): 87–93.

Johannessen, J.B., and Hagen, K. (eds.). 2014. *Språk i Norge og nabolanda. Ny forskning om talespråk.* Oslo: Novus forlag.

Vangsnes, Ø. A. 2014. Noun Phrases. *The Nordic Atlas of Language Structures (NALS) Journal* 1 (1): 4–9.

Event | **CLARINO's Involvement in the University Curriculum**

Written by **Koenraad de Smedt**

Over the past few years, the CLARINO Bergen Centre has made significant efforts to engage and train potential users of the CLARINO and the wider CLARIN infrastructure in the context of higher education. This includes both one-time teaching events and active involvement in regular university courses.

Staff of the CLARINO Bergen Centre have participated in several PhD researcher training courses, in particular those organized by the Norwegian Graduate Researcher School in Linguistics and Philology (LingPhil). One of these was a course at a LingPhil summer school in Northern in 2016. The course was focused on data management and spread out over a whole week, in daily two-hour blocks. The course, given by Koenraad De Smedt and Gunn Inger Lyse Samdal, presented the creation and management of research data, repositories and standards, and documentation and metadata, as well as showcased how the data can be found and used. Examples based on resources, tools and recommendations from CLARINO and CLARIN were used throughout the course. For instance, CLARIN licences and the CLARIN "laundry tag" system of categorizing licences were explained. Examples from the Trolling repository were used to illustrate good citation practices and examples from the ASK corpus at the CLARINO Bergen Centre to illustrate principles and standards for annotation.

Teaching turned out to be highly interactive, with questions coming from lecturers as well as students. Besides the lectures, the course also featured a lot of individual and group exercises. Students practiced, for instance, how to choose a licence type and document metadata. Since this course was part of the LingPhil summer school, data management was not seen as an isolated activity, but linked to other aspects of linguistic methodology.

In recent years, the use of data and tools in general has also become well integrated in several regular courses of the bachelor's and master's programmes in Linguistics at the University of Bergen. An introductory course in language and computers demonstrates corpus search with the Corpuscle system, while syntax courses have used online parsing with XLE-Web and materials from treebanks in INESS, part of the CLARIN Knowledge Centre for Treebanking, which was also featured in Tour de CLARIN.[4] Both web services are provided by CLARINO. Data from INESS has also been used in a master's course on computational language models to show how empirical corpus data can strengthen or challenge hypotheses about grammar. It has turned out that students quickly learn to use the system in projects for term papers and master's theses. Victoria Troland's master's thesis, for instance, used INESS to extract syntactic markers from syntactically analysed Norwegian novels, and subsequently used these markers as a basis for an author identification model.

In our experience, students learn to use the infrastructure best when such use is well integrated in the regular curriculum. We therefore intend to introduce materials and methods based on CLARINO and CLARIN in even more courses. We will also repeat the PhD researcher training course in 2020, albeit in a more compact form.

Reference:

Troland, V. 2015. *Hvem er forfatteren? - Stilometriske undersøkelser av norske prosatekster*, master's thesis.

Figure 13: Students at the LingPhil summer school 2016 in Northern Norway did group exercises under the supervision of Koenraad De Smedt and Gunn Inger Lyse Samdal.

[4] https://www.clarin.eu/blog/clarin-knowledge-centre-treebanking

Interview | **Yvonne van Baal**

Yvonne van Baal is a PhD student in linguistics at the University of Oslo.[5] In her research on definiteness marking on the noun phrase in American Norwegian, she has successfully used resources and tools developed at the CLARINO Text Laboratory. [Photo by: Nadia Frantsen/UiO]

Could you briefly describe your research background and your current position? How did you get involved with Text Laboratory?

<

Currently, I am a PhD Candidate at the University of Oslo (Department of Linguistics & Scandinavian Studies), and I will defend my dissertation in February. My research interests are bilingualism, language acquisition and language variation. All these topics come together in the field of heritage languages, which is the topic of my dissertation. I am affiliated to the Text Laboratory, and Janne Bondi Johannessen, the leader of the TextLab, is one of the supervisors of my PhD project. The TextLab has been very important for me because of their expertise with data collection, data storage, and corpora.

>

[5] This interview took place in January 2020. Yvonne successfully defended her PhD thesis in Spring 2020 and is now a postdoctoral researcher at the Department of Language and Literature and a member of the AcqVA research group at the Norwegian University of Science and Technology in Trondheim, Norway.

Your PhD project focuses on a (morpho)-syntactic topic, which is definiteness marking in American Norwegian, a heritage language. Could you briefly present your PhD work, in terms of its aims and results? What makes American Norwegian an interesting language to study?

<

American Norwegian is a heritage variety of Norwegian that is spoken in the Midwest of the US by descendants of Norwegian immigrants. Heritage languages are interesting for linguistic research because they are minority languages learned and used at home, while the larger, national society uses another language. As such, they provide unique insights into the roles that language acquisition as well as language use play in a speaker's language competence. In addition, they can provide interesting perspectives on language variation, both within and across speakers, and on language change.

The main aim of my project was to investigate how definiteness is marked in American Norwegian, specifically in relation to the use of the so-called "compositional definiteness" structures. This refers to the fact that a semantically definite noun phrase in Norwegian contains both a prenominal determiner and a definite suffix on the noun, but only when the noun phrase also contains an adjectival modifier. Let's exemplify this with the noun *hus* ("house"). In the Norwegian spoken in Norway, the equivalent of the definite noun phrase *the house* would be *huset*, where definiteness is grammatically marked by the suffix *-et* rather than by an article like in English. However, a modified definite noun phrase such as *the old house* would be *det gamle huset*, because the use of the modifier *gamle* ("old") requires the additional use of the definitive determiner *det*.

In my PhD work, I have found that definiteness marking is in general very similar to the Norwegian spoken in Norway, while compositional definiteness is vulnerable to restructuring. This means that the prenominal determiner is often omitted, but the suffix is very stable – speakers of American Norwegian often use structures like *gamle huset*, which would not be used in homeland Norwegian.

I argue that this language change cannot be caused by transfer from English. If American Norwegian would have become more like English, the speakers would have said *det gamle hus* (like "the old house"), but they do exactly the opposite. Interestingly, it turns out that children who grow up in Norway often omit the determiner while acquiring compositional definiteness. They too say *gamle huset*! But acquisition in the heritage context is somewhat different from acquisition in a monolingual context. I therefore argue that it is the acquisitional context that caused this difference between American Norwegian and homeland Norwegian.

>

Your work relies on spontaneous speech data from the Corpus of American Nordic Speech,[6] which is made available through Text Laboratory. How have you used the corpus in your PhD work to study definiteness marking? What kind of new empirical results were you able to achieve on the basis of this corpus?

<

The Corpus of American Nordic Speech (CANS) has already been used by other researchers to study definiteness marking, in Anderssen, Lundquist and Westergaard (2018), who find that the determiner is often omitted, and compare this with the use of pre- and post-nominal possessives. I complement their research with experimental data (see below). I did, however, use the CANS corpus to extract frequency lists of lexical items.

I used the Nordic Dialect Corpus (NDC). This is a corpus with spoken conversations from many different dialects in Scandinavia, and is available through the same search interface (Glossa) as CANS. I used this corpus to study those Norwegian dialects that are spoken in the regions where the ancestors of the American Norwegians came from. The Nordic Dialect Corpus was extremely important for my research, because the dialectal data it offers allowed me to establish a proper baseline for comparison with the language spoken by the heritage speakers in the United States. A comparison with Bokmål or Nynorsk Norwegian would be unfair in this case, because most American Norwegian speakers are not familiar with the written language used in Norway.

>

Why is it important that the corpus consists of speech data instead of written materials? Could you discuss how you have complemented the corpus data with experimental data that you have collected during fieldwork in the United States?

<

The only data source that is available for American Norwegian is spoken data. These speakers grew up speaking Norwegian at home, but the school system in the US is of course completely in English. As a result, American Norwegians are only literate in English, and not in Norwegian (with a few exceptions); we therefore have to rely on spoken speech data.

[6] http://tekstlab.uio.no/norskiamerika/english/corpus.html

CANS is a great resource for research on American Norwegian: it is the only corpus of this language, contains the speech of many speakers, and is easily searchable. However, as with most corpora that consist of spontaneous conversations, it has some limitations. One of these is that infrequent grammatical constructions are difficult to study, and compositional definiteness is one of these phenomena; in the corpus one only finds a few instances of the construction for each speaker. I therefore complemented the findings from CANS with elicitation experiments that I conducted during fieldwork trips to the Midwestern United States in 2016 and 2018. With the experiments, which included a picture-aided elicitation task and a translation of a short story from English into Norwegian, I was able to obtain many additional phrases that in Norwegian require compositional definiteness. By doing so, I could investigate how each speaker uses this construction.

>

Both the Corpus of American Norwegian and the Nordic Dialect Corpus are available through the Glossa search system developed by the Text Laboratory. How have you used the system in your work? Could you highlight any features in Glossa that were especially indispensable for your research purposes?

<

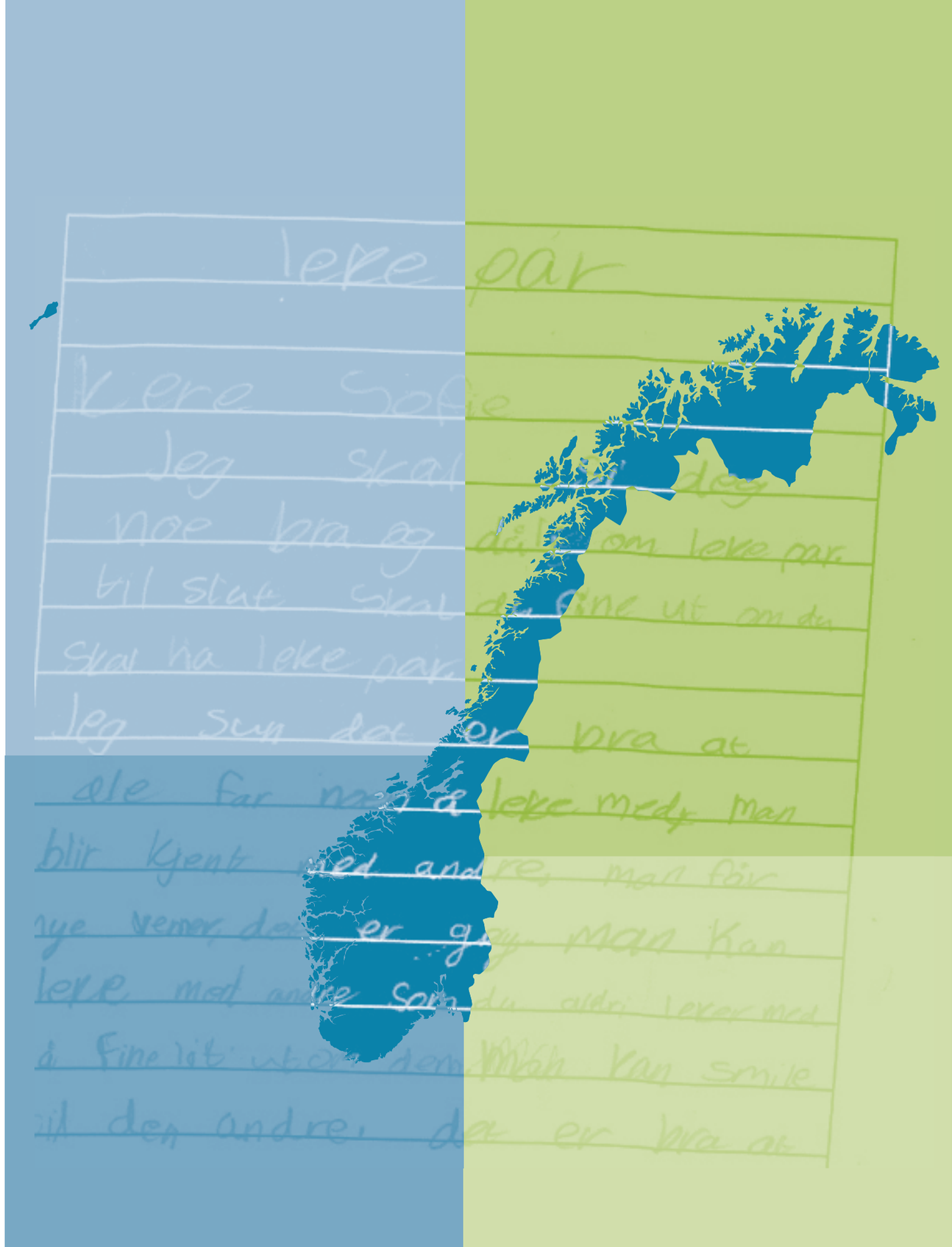
I found Glossa to be a very user-friendly system that is tailored to researchers who do not have a lot of experience with making complex search queries. Unlike most search interfaces which require familiarity with the query language syntax, the main advantage of Glossa is that it allows you to specify detailed morphosyntactic characteristics of the lemma that you are looking for by simply selecting them in a drop-down menu. I was thereby able to observe the use of the compositional definiteness construction in the Corpus of American Nordic Speech and the Nordic Dialect Corpus by narrowing the Glossa query to those strings that include a noun marked as definite and immediately preceded by an adjective and possibly a determiner. Furthermore, Glossa allowed me to search for important extra-linguistic metadata. For instance, in the Nordic Dialect Corpus, I could restrict the search to Norwegian dialects spoken in a specific location and only include those dialects spoken by the ancestors of our American Norwegian speakers.

>

How does the Text Laboratory facilitate the study of heritage languages?

The TextLab has created the technological foundation for the study of American Norwegian. Basically, all the data that we have on American Norwegian is stored, transcribed, and tagged at the TextLab, where they have also built the only corpus available for this heritage language. Such preservation is important, because the current speakers of American Norwegian are all elderly, and the final generation of speakers. Their children do not speak the language, which means that building and preserving collections of their speech is an urgent task for infrastructures like the TextLab and by extension CLARINO. There are other minority languages that are in a situation like this.

If these languages are not recorded, valuable data for linguists are lost forever. In order to further facilitate the study of endangered languages, TextLab recently launched the LIA-corpus, which contains a lot of old recordings of Norwegian dialects, and also includes recordings of the Sami languages.

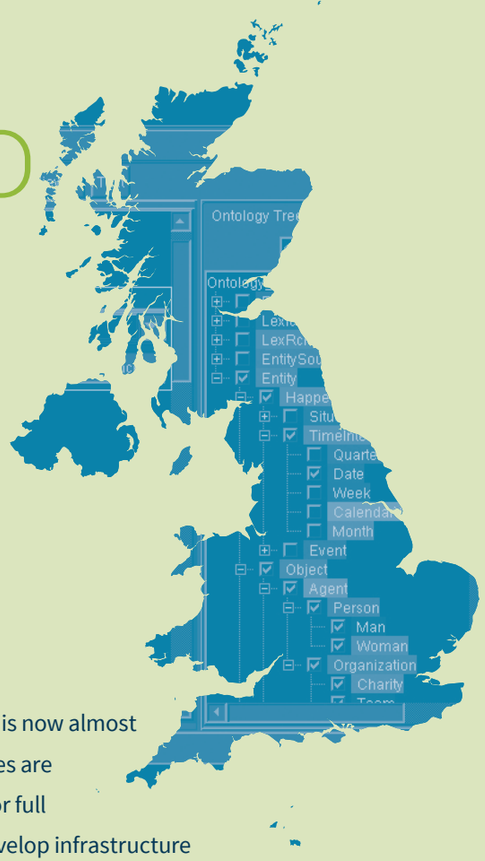


References:

van Baal, Y. 2020. *Compositional Definiteness in American Heritage Norwegian.* Doctoral dissertation, University of Oslo, Norway.

Anderssen, M., Lundquist, B., and Westergaard, M. 2018. Cross-linguistic similarities and differences in bilingual acquisition and attrition: Possessives and double definiteness in Norwegian heritage language. *Bilingualism: Language and Cognition* 24 (4): 748–764.

THE UNITED KINGDOM



Introduction

Written by **Martin Wynne**

The UK has been an observer of CLARIN since 2015 and is now almost half-way through its second three-year period. Countries are admitted as observers in order to prepare a proposal for full membership, to build a national consortium, and to develop infrastructure at the national level, and these are the tasks that are currently the focus for CLARIN in the UK. CLARIN-UK is a loose consortium of researchers and providers of language resources and tools, working in tandem with the Arts and Humanities Research Council, part of the recently founded body UK Research and Innovation (UKRI).

The CLARIN-UK consortium currently consists of eleven centres,[7] which have stepped forward to express an interest and commitment to being part of CLARIN. Other centres are welcome to get in touch and participate in meetings and other activities. The current members of the UK CLARIN consortium are:

- Bodleian Libraries (University of Oxford)
- British Library
- Centre for Corpus Research (University of Birmingham)
- Centre for Translation Studies (University of Leeds)
- Endangered Languages Archive (SOAS University of London)
- Centre for Corpus Approaches to the Social Sciences and UCREL (Lancaster University)
- National Centre for Text Mining (University of Manchester)
- Natural Language Processing Group (University of Sheffield)
- Research Group in Computational Linguistics (University of Wolverhampton)
- School of Critical Studies (University of Glasgow)
- School of Humanities (Coventry University)

The main criteria for inclusion of institutions in CLARIN-UK are that they have a strong relation to digital language data, tools, or research, and that there is a commitment to sharing and connecting data and tools to support research. The particular strengths of digital language research in the UK are reflected in the centres, resources and events featured on the CLARIN-UK website. Some of the most widely used resources are the Spoken BNC2014, the Historical Thesaurus of English, the CliC interface for the works of Dickens and other literary corpora, and the CQPweb concordancer at Lancaster, while prominent CLARIN-UK tools include GATE, which is an open source software that performs a wide range of computational tasks, CLAWS, which is a powerful part-of-speech tagger for English, and Wmatrix, which is a corpus analysis tool that among other functions provides a web interface for the CLAWS tagger.

The numerous training events include annual summer schools in corpus linguistics and Digital Humanities, held in Birmingham and Lancaster, and training in language documentation offered by SOAS. UK members have co-organized workshops on Oral History, NLP for Historical Texts, and Analysing Social Media in East Asian Studies, and participated in workshops on language resources in teaching, and dealing with the GDPR, among other topics.

The Oxford Text Archive (OTA) is registered as a CLARIN-certified repository, using the CLARIN single sign-on system and offering language resources to the Virtual Language Observatory. The OTA makes available more than 60,000 digital resources, including the British National Corpus and a wide variety of Old English, Middle English (such as the first printed edition of Geoffrey Chaucer's famous *Canterbury Tales* from 1476), and Modern English historical texts, digitized as part of Early English Books Online (EEBO).

SOAS University of London, whose Endangered Text Archives are involved in CLARIN UK activities, is part of the CLARIN Knowledge Centre for linguistic diversity and language documentation (CKLD). Such centres in the UK play an important role as part of CLARIN's distributed infrastructure. The UK is currently developing a national research infrastructure roadmap, and CLARIN-UK is contributing to the consultation about both requirements and existing

[7] https://www.clarin.ac.uk/

infrastructure services. CLARIN-UK is featured on InfraPortal, the UK's Research and Innovation Infrastructure Portal.

Participation in CLARIN is not dependent on EU membership, so Brexit does not represent a direct threat to our activities and plans. On the contrary, CLARIN, along with other European research infrastructures, offers an excellent opportunity for continued and growing collaboration and cooperation with our European partners.

Figure 14: Martin Wynne, the national coordinator of the UK consortium.

Tool | **GATE Services**

Written by **Diana G. Maynard**

GATE is a widely used, established open-source NLP infrastructure that provides a framework and numerous essential components (plugins) for all kinds of NLP and text processing tasks.[8] Developed at the University of Sheffield, which is a partner in CLARIN-UK, it is now 20 years old and has a research team of 16 people, as well as a vibrant community of users, ensuring its continuous development and usage in a wide variety of scenarios and domains. The components, some of which are available in a variety of languages, include:

- pre-processing tools (e.g., tokenization, lemmatization, normalization);
- language processing tools (e.g., part-of-speech tagging, parsing, chunking, morphological analysis);
- domain- and task-specific NLP tools (e.g., named-entity recognition and linking, gene tagging, recognition of legal terminology, biomedical processing, social media analysis);
- NLP development tools (machine learning algorithms, a linguistic pattern-matching, rule engine, performance evaluation tools).

Beyond making these NLP tools openly available, GATE also provides:

1. GATE Developer – a graphical interface for developing and testing new NLP tools and applications;
2. GATE Cloud – a cloud-based NLP platform-as-a-service, for seamless service-based deployment of GATE NLP tools and applications;
3. GATE Mimir – a highly scalable semantic indexing and search platform;
4. GATE Teamware – a collaborative, web-based document annotation tool.

With these NLP tools and services, even researchers without coding experience can easily use, adapt, or build an NLP system to analyse text. Thanks to its open source nature, GATE users also benefit from tools and applications that are provided by third-party GATE users (typically other academic researchers adapting GATE tools to other languages and tasks) and shared via public repositories such as GitHub. In particular, the GATE Cloud platform offers 69 different services covering many languages and domains, providing an easy way for people to try out a number of applications on sample text and to run them as a web service over more substantial datasets,

[8] https://gate.ac.uk/

without having to deal with any integration issues, by providing a common API across disparate services. Many of these services have already been integrated in D4Science and are being currently integrated in the European Language Grid and RISIS platforms, as well as the CLARIN Language Resources Switchboard.

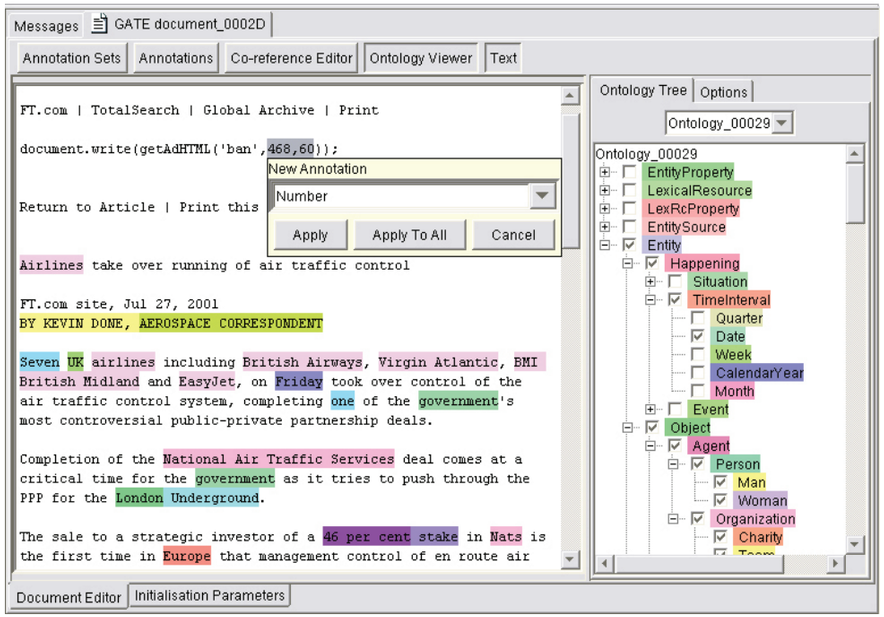


Figure 15: Ontology viewer in GATE developer, highlighting named-entity relations within an annotated English text.

The GATE development team dedicates significant resources to supporting and growing the GATE user community through regular and bespoke training courses, open access training materials/documentation, and an open user mailing list, as well as offering consulting services to help the development of new NLP applications in a wide variety of sectors. At the CLARIN-PLUS workshop "Creation and Use of Social Media Resources", Diana Maynard, who is one of the developers of GATE at the University of Sheffield, held a demo session in which she showed how open-source GATE tools, such as the TwitIE named-entity recognizer and a Twitter-based sentiment analyser, can be applied to analyse social media texts in various languages.

Large-scale text analysis can be carried out with GATE tools to gain valuable quantitative insights from large volumes of social media content, helping to answer important open questions. For example, one important strand of work has looked at how social media can be harnessed more effectively during crises and natural disasters, while another has looked at political debates. The past few years have heralded the age of ubiquitous disinformation – aka fake news – which poses serious questions over the role of social media and the internet in modern democratic societies. Topics and examples abound, ranging from the UK "Brexit" referendum and the 2016 US presidential election to medical misinformation (e.g., miraculous cancer cures). For example, how do political parties, candidates, and voters engage online in the run up to elections and referenda? How polarized are these discussions and how prevalent are abusive comments? What is the role of disinformation and bots during elections? Can they influence the results?

In addition, the explosion of free text in healthcare (such as electronic health records and research papers) creates important opportunities that can benefit from NLP and text mining in the biomedical domain. Examples include extracting patients' background data (e.g., occupation, HIV stats, prescription) from their records, or labelling protein, DNA/RNA and cell types from the biomedical literature. GATE tools have been successfully applied to the following tasks in biomedicine (amongst others):

- the extraction of medical terms in the text and linking them with UMLS concepts;
- automatic extraction of drug names and dosages from prescriptions;
- the expansion, annotation and co-reference of biomedical abbreviations and acronyms;
- the recognition of organisms in biomedical literature.

Reference:

Cunningham, H., Maynard, D., Bontcheva, K., et al. 2011. *Developing Processing Components with GATE Version 6 (a User Guide)*. The University of Sheffield, Department of Computer Science.

Resource | **Historical Thesaurus of English**

Written by **Fraser Dallachy** and **Marc Alexander**

The Historical Thesaurus of English is an invaluable resource for research into the semantics of English, from the study of individual concepts up to a perspective on the language as a whole, from its beginnings to the present day.[9] It contains every sense of every word in the language as recorded by the Oxford English Dictionary and other sources, including Old English dictionaries, sorted into semantic categories which are themselves ordered into a comprehensive hierarchy of ideas. The first edition was developed by research staff and students at the University of Glasgow over the course of half a century, beginning in 1964 and reaching publication in 2009. The Thesaurus is now freely available for consultation through the University of Glasgow's website and accessible via the CLARIN-UK website.

Since completion, Thesaurus data has been explored by a number of daughter projects, most notably the Mapping Metaphor project, also based at the University of Glasgow. Mapping Metaphor looked for evidence of systemic repurposing of words from one semantic field into another – a phenomenon known as a conceptual metaphor – such as the use of words related to travelling ("It's been a long *road*" or "We had a *bumpy* start") to describe life experiences. By comparing the contents of every individual category in the Historical Thesaurus against all other categories, the project mapped those areas of the English vocabulary which are strongly connected by such metaphorical borrowing of words. These are displayed on the Metaphor Map of English, which was awarded "Best DH data visualization" in the 2015 DH Awards.

The organization of lexis into meaning categories forms an ideal knowledge base for semantic annotation software, which attempts to tag words in text with a label representing their meaning. Such semantically annotated text allows the use of concepts as search terms (e.g., the concept HAPPINESS rather than the word *happiness*) and can feed into deeper analysis of the structuring of information within sentences and texts. This use of the data was explored by a multi-institutional team including NLP experts from the University of Lancaster whose previous work included the UCREL Semantic Analysis System (USAS) and the VARD spelling normalization tool.

[9] https://ht.ac.uk/

The project developed the Historical Thesaurus Semantic Tagger and the release of two semantically annotated linguistic corpora, Semantic Hansard and Semantic EEBO, both freely accessible through English-corpora.org (registration required). The Hansard Corpus contains records of debates in the Houses of Commons and Lords of the British parliament from 1803 to 2005 CE (approximately 1.6 billion words) and is the largest parliamentary corpus in CLARIN Resource Families, while EEBO (Early English Books Online) contains open-source transcriptions of early modern (roughly 1470–1700 CE) printed material (approximately 755 million words in 25,000 texts).

The final major project to use Thesaurus data thus far is Linguistic DNA, which primarily sought to explore regular word groupings in EEBO texts as evidence for the emergence and development of concepts in the early modern period. Thesaurus data was analysed for evidence that particular semantic fields experienced sudden rises, falls, or other remarkable behaviour in the focal period, and a tool for exploring these is under development, with a test version available through the Thesaurus website.

The Historical Thesaurus is a rich resource whose exploration has really only just begun. There are a number of other projects working with the data and that of its sister project, A Thesaurus of Old English, with highly anticipated results. A second edition of the Thesaurus, incorporating data from the 3rd edition of the Oxford English Dictionary, is due to launch by the end of 2020 and is currently keeping the editorial team very busy. Nonetheless, researchers interested in using Thesaurus data are encouraged to get in touch.

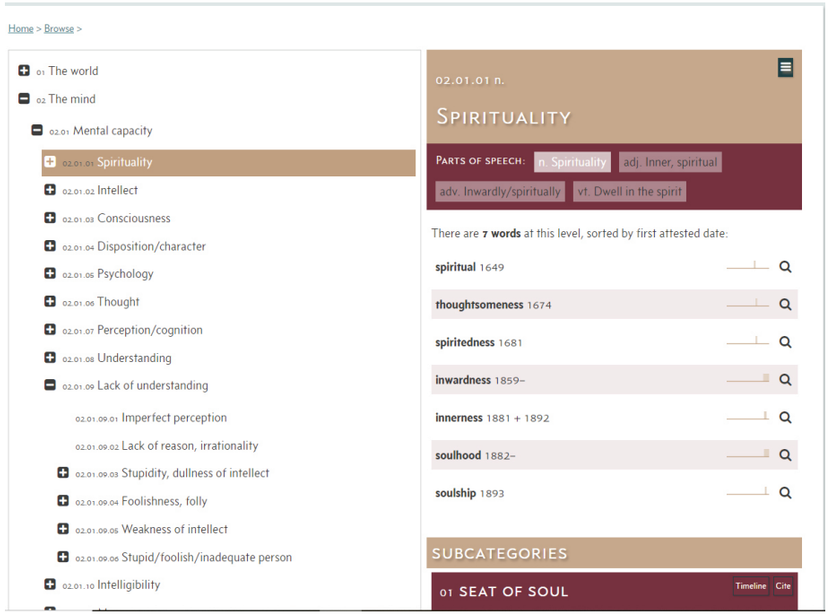


Figure 16: The Historical Thesaurus entry for the noun *spirituality,* showing related concepts sorted by date of attestation.

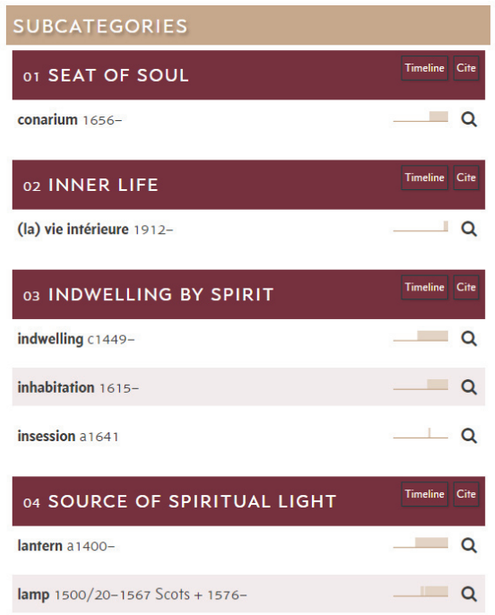


Figure 17: The list of concepts related to *spirituality.*

Event | **Lancaster Summer Schools in Corpus Linguistics: Behind the Scenes**

Written by **Vaclav Brezina** and **Dana Gablasova**

Lancaster Summer Schools in Corpus Linguistics are a free annual event held at Lancaster University, UK, at the end of June.[10] Co-organized by the ESRC Centre for Corpus Approaches to Social Science, which is a CLARIN-UK member, and the Department of Linguistics and English Language, the summer schools offer a week of intensive training in corpus linguistics, in which lectures are combined with practical, hands-on sessions in computer labs. The summer schools attract participants from all around the world; since 2013, more than 1,000 participants from over 30 countries have attended the summer training programme. In 2017, Darja Fišer, the director of User Involvement at CLARIN ERIC, gave a presentation on CLARIN ERIC at a plenary session of the summer schools.

Drawing on particular strengths of the department and research centre, we currently offer three streams reflecting major areas in the field of corpus linguistics and its applications:

- corpus linguistics for analysis of language, discourse and society;
- corpus linguistics for language learning, teaching and testing;
- statistics and data visualization for corpus linguistics.

In this text, we share our experience as organizers of the event. Dana is the main organizer, coordinating the academic programme and teaching in all three streams; she is also the convenor of the stream for language learning, teaching and testing. Vaclav is the convenor of the stream focusing on statistics and data visualization in corpus linguistics.

Figure 18: A lecture on corpus linguistics and health communication (Elena Semino).

[10] http://wp.lancs.ac.uk/corpussummerschools/

Figure 19: A computer lab session during Lancaster Summer schools in 2019 (Vaclav Brezina).

Figure 20: A lecture on the language of Shakespeare from the corpus perspective during Lancaster Summer Schools in 2019 (Jonathan Culpeper).

This year brought a special challenge across the higher education sector and society in general. Despite the COVID-19 crisis, we decided to go ahead with the Lancaster Summer Schools 2020 in an online format. This involved pre-recording of lectures (the corpus statistics stream offered a full series of seven lectures) and preparing materials and events to be available online. We used Lancaster University's open learning MOODLE environment to deliver lectures and exercises asynchronously; in addition, several live events were organized via MS Teams to give the participants the opportunity of synchronous interactions via Q&A. One of the highlights of the online version of the summer schools this year was a free webinar on corpus analysis of linguistic data. As part of the webinar, we released #LancsBox v. 5, a free software tool for the analysis and visualization of corpora. The new version includes the Wizard tool, which automatically analyses corpora and produces research reports. #LancsBox is one of the corpus tools listed on the CLARIN-UK website.

While moving resources online involved considerable effort, it was very rewarding that this format allowed us to share the materials and training with a large audience of those interested in learning about corpora and corpus methods; we were delighted that more than 5,000 people accessed the resources during the online summer schools in June 2020. As a result, we decided to keep the resource page available after the end of the event and continue adding resources to it.[11]
If you are interested in corpus linguistics and would like to visit Lancaster University, do not miss the application process, which usually opens in early January. It would be wonderful to see you in Lancaster at one of our future training events, or you can keep in touch with us online.

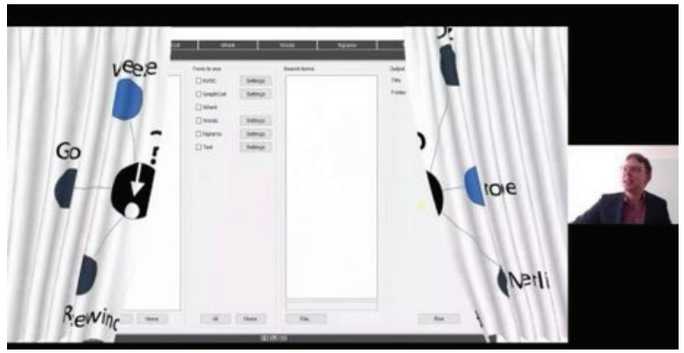


Figure 21: Unveiling the Wizard tool in #LancsBox v. 5.

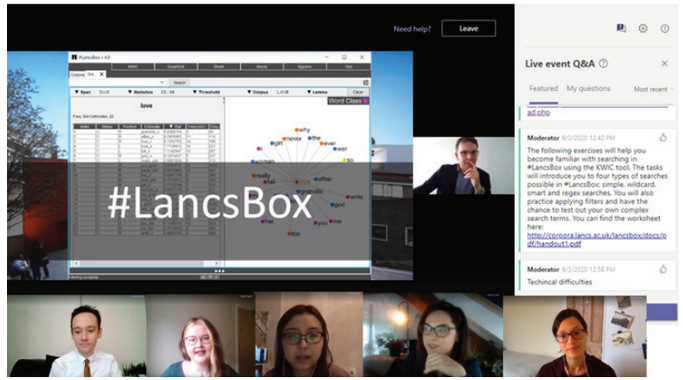


Figure 22: A snapshot from a webinar introducing a brand new version of #LancsBox.

[11] https://www.lancs.ac.uk/corpussummerschools/online-resources/

Interview | **Michaela Mahlberg**

Michaela Mahlberg is Chair of Corpus Linguistics at the Department of English Language and Linguistics at the University of Birmingham. She is Principal Investigator of the project that has developed the CLiC search system, which is one of the flagship resources of CLARIN-UK.

Please introduce yourself. What is your research background, and what inspired you to approach literature from a corpus stylistic perspective, both in research and teaching?

<

I did a degree in English and Mathematics at Bonn University, and then went on to complete a PhD in English Linguistics at the University of Saarbrücken. I got into using corpus linguistics for the study of literature through my interest in Charles Dickens, my favourite author. From a corpus linguistics point of view, Dickens is quite fascinating, too. He is a master of using language to great effect, and his use of repetition in particular has often been commented on. Repetition, and hence frequency, is obviously right up our street as corpus linguists. Dickens also seems to have been especially aware of the typical language use and patterns that are common in the language in general. An observation in *David Copperfield* almost sounds like a corpus linguistic comment: "conventional phrases are a sort of fireworks, easily let off, and liable to take a great variety of shapes and colours not at all suggested by their original form". In practical terms, an important catalyst for my focus on literature was a workshop that Martin Wynne, who is now the coordinator of CLARIN-UK, organized in Oxford – more than 14 years ago – to look at "Corpus Approaches to the Language of Literature". This opportunity really made me think through the fundamental principles of what it means to study literature with corpus methods. These kinds of questions have kept me busy ever since!

>

What were the biggest bottlenecks in corpus-assisted approaches to study literary texts when you first started, and how has the field evolved since?

<

Examples from literary texts have always been used in corpus linguistics, but initially mainly as examples – to illustrate, for instance, the difference between types and tokens – rather than with a focus on the literariness of these texts. Equally, general reference corpora would typically contain samples from fiction – not necessarily full texts but only text samples. One reason for text samples can be copyright restrictions, but another was the belief that text samples are sufficient to study the phenomena that are of interest to corpus linguists. Overall, fiction was mainly treated as a register to compare other registers to. Notably, researchers from literary stylistics have contributed to demonstrating that literary qualities and features of individual texts are worth investigating with corpus methods. There is now also increasing interest from literary scholars. This exchange across disciplines is rather important, so that corpus linguists can demonstrate the benefits of methods but also learn about how literature is approached in other fields.

>

What are the main challenges today?

<

Today the key challenge is to bring developments in corpus linguistics and Digital Humanities better together. It is amazing how much is shared between the two fields. But seemingly, researchers in the two fields are not aware of these similarities. If you look at the research literature, there is very little in terms of cross-referencing and separate terminology is being developed that veils methodological similarities, especially at this point where technology is developing much faster than it has ever done.

>

Could you briefly introduce the Corpus Linguistics in Context (CLiC) search system,[12] which is one of the flagship resources of CLARIN-UK? What distinguishes CLiC from other well-known corpus concordancers?

<

CLiC was designed with users in mind who want to focus on the literary properties of texts, so that the concordance function can also be seen as an aid for close reading and engagement with a text. We aimed to consider how literary scholars, English teachers or pupils in schools might find it useful to draw on standard corpus methods. So for CLiC, the text view, which shows

[12] https://www.clarin.ac.uk/clic

how the search result appears in the running text, is equally important to the concordance view. Moreover, for the study of concordances there is a KWICgrouper function to help users sort concordance lines according to specific context words. It is further possible to add user-defined "tags" to a concordance analysis to support the classification of lines in an easy format. The most distinctive feature of CLiC, however, is that the corpora it accesses have been annotated for direct speech and narration so that concordance searches can be run for specific subsections of texts. The main distinction is between "quotes", i.e., text within quotation marks, and "non-quotes", i.e., text outside of quotation marks, which roughly equates to direct speech and narration – as the corpora mainly contain fiction from the 19th century, where direct speech still tends to be prevalent. It is also possible to focus searches on "suspensions", i.e., stretches of narration that interrupt the speech of characters. The ability to search fiction in this way is crucial to study textual features that are characteristic of this specific register. In fiction, different discourse levels come together. If the voice of the narrator and the speech of fictional characters is just treated the same in the textual analysis, important information will be missed.

>

And how does CLiC serve the academic community from a research infrastructure perspective?

<

The CLiC web application is freely accessible – without a need to log in. Importantly, we made the corpora as well as the code for CLiC openly available through GitHub. We have also created extensive documentation to further support open research and reproducibility. A good example of the effectiveness of this approach is the way in which the Corpus of African American Writers 1892–1912 (AAW) came to be added to the CLiC corpora. Claiborne Rice and Nicholas J. Rosato from the University of Louisiana at Lafayette had come across our documentation for the quote and non-quote annotation and compiled this corpus, which we then jointly incorporated into CLiC.

>

In your recent work, you and your colleagues have used CLiC to study speech-bundles in 19th century English novels, primarily Charles Dickens's novels. How did you use CLiC to extract and study the speech bundles? Could you briefly present the main aims of the research, and how does fictional speech in the CLiC corpora relate to real spoken language?

<

In our study, we generated 5-grams for the quotes subcorpus of the Dickens novels corpus. Lexical bundles are n-grams that are highly frequent, as determined by specified frequency thresholds. So far, lexical bundles in fiction have mainly been looked at to compare fiction as a register against other registers. In such comparisons, lexical bundles are considered across the whole texts. In our study, we specifically focus on frequent 5-grams in the quotes subcorpus. We compare Dickens to other 19th century fiction, as well as to the BNC1994, which is also available through the Oxford Text Archive CLARIN-UK repository. This allows us to identify fictional speech-bundles, i.e., bundles that have particular functions in creating fictional worlds. Such bundles include those that generally appear in fictional speech across a range of authors, as well as bundles that reflect idiosyncratic authorial features. Most interesting are the bundles that are shared between fictional and real speech. It has been a common view that fictional speech is really rather different from real spoken language. But speech-bundles show that both in fiction and real life people say things like *it seems to me that*, *what do you think of* or *and all that sort of.* Such phrases probably receive less attention in the study of literature precisely because they are generally common in everyday speech.

>

Why is it important to contrastively compare fictional corpora with non-fictional corpora in the context of a Digital Humanities approach to literary theory?

<

Patterns that are shared across the language of fiction and non-fiction are an important pointer to the link between fiction and the real world. This link works in two ways. For instance, when fictional people use phrases like real people do, these phrases trigger readerly responses that draw on the reader's linguistic background knowledge of how people speak. So information is read into a character. On the other hand, fiction can portray real world phenomena in specific ways and it can equally affect the reader's perception of the real world. Approaches in Digital Humanities that focus on the study of literary and cultural history provide such big picture views of fictional worlds. Such approaches are very similar to corpus studies of non-fiction texts that identify specific discourses.

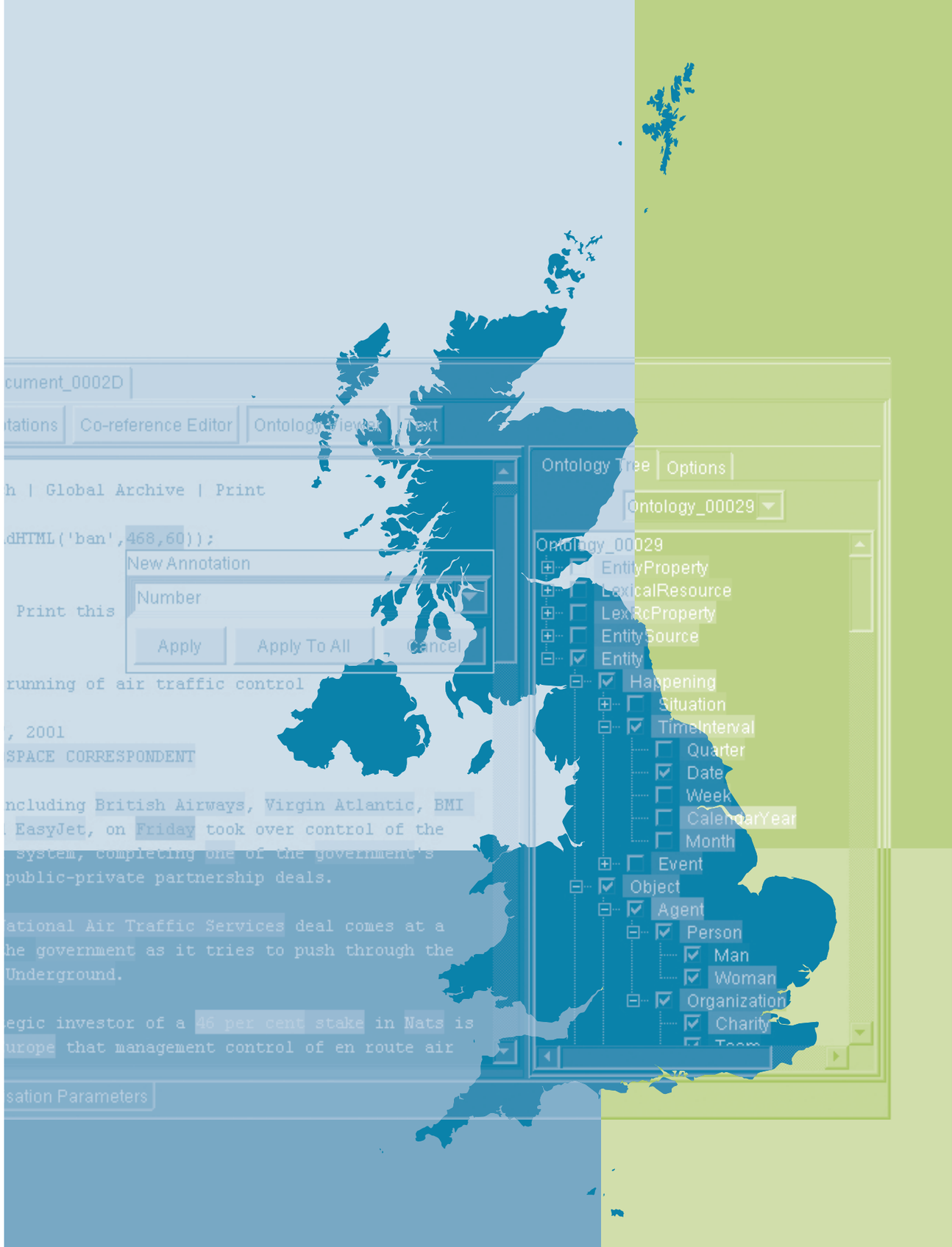
>

How is CLiC supporting the new generations of researchers, and how is it advancing the state of the art in research methods? What in your opinion are the key next steps for the CLARIN research infrastructure in the coming years in order to best serve its user base?

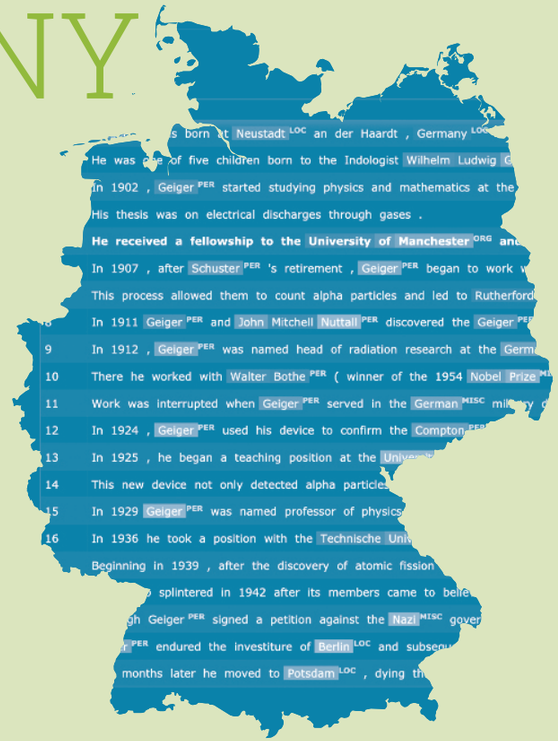
In addition to supporting open research and replicability, as I described above, we have also been running a CLiC blog for a couple of years now. On this blog, researchers and educators present examples of how they use CLiC or discuss related topics, methods and resources. The aim of the blog is to facilitate the sharing of ideas as well as encourage new ways of thinking about digital resources. Most contributors are early career researchers, but we also get blog posts from teachers – who in a way thus start very early with supporting new generations of researchers.

CLARIN already provides amazing resources and opportunities. I find the Federated Content Search – the ability to search across a range of CLARIN resources at the same time – particularly fascinating! In the coming years, CLARIN can probably achieve even more in driving forward standards of interoperability.

There is also still a lot of potential to set agendas with funders and support them in understanding the infrastructural needs of digital projects, as well as what's needed to ensure the sustainability of infrastructures once a funded period ends.



GERMANY



Introduction

Written by **Erhard Hinrichs**

The national consortium CLARIN-D has been contributing to the European CLARIN research infrastructure since 2008.[13] The institutions that are involved in the CLARIN-D consortium comprise members of the Leibniz Association, the German Academies of Arts and Sciences, and leading research universities with a focus on digital methods for Humanities research. CLARIN-D institutions offer research data and associated data services to support all stages of the research data lifecycle for scholars in the Humanities and in select disciplines of the Behavioural and Social Sciences, particularly in Psychology, Political Science and Sociology.
The National Coordinator of CLARIN-D is Professor Erhard Hinrichs.

The backbone of the CLARIN-D infrastructure is an open and extendable network of certified data and service centres with complementary areas of expertise. Currently, there are eight data centres certified by the Core Trust Seal assessment for trustworthy data repositories. They are located at the following institutions:

[13] https://www.clarin-d.net/en/

- Bavarian Archive for Speech Signals, Ludwig-Maximilian University of Munich
- Berlin-Brandenburg Academy of Sciences and Humanities, Berlin
- Leibniz Institute for the German Language, Mannheim
- Department of Linguistics, Tübingen University
- Hamburg Centre for Language Corpora, University of Hamburg
- Department of Computer Science, University of Leipzig
- English Linguistics and Translation Science, Saarland University, Saarbrücken
- Institute for Natural Language Processing, University of Stuttgart

In addition, four associated data centres contribute data and services to CLARIN-D. They are located at Cologne University, the University of Duisburg-Essen, Frankfurt University, and at the Georg Eckart Institute for International Textbook Research. CLARIN-D institutions are also contributing their expertise to the CLARIN Knowledge-Centre for Linguistic Diversity and Language Documentation.

From the very beginning, CLARIN-D has put strong emphasis on close and continued interaction between infrastructure users and infrastructure providers. In order to foster this, CLARIN-D working groups in different disciplines of the Humanities and of the Behavioural and Social Sciences were established. The main objective of these working groups is to aggregate and prioritize the data and service needs of the user communities that they represent, and to help promote the use of CLARIN-D data and services in research and teaching. This agenda is greatly facilitated by a number of representative use cases that have been jointly developed by scholars from CLARIN-D working groups and members of the CLARIN-D data centres. Whenever possible, CLARIN-D offers its data and services via easy-to-use web portals and web applications, such as WebLicht (automatic annotation and workflow engine), WebAnno (manual annotation), WebMAUS (signal and transcript alignment), the CLARIN Helpdesk, and the CLARIN Legal Helpdesk, to name some examples. This strategy obviates the need for users to have to download and install software on their own computers. This seems particularly important for humanities scholars, who often lack the necessary programming skills or technical expertise.

The CLARIN-D consortium actively contributes to community building and educational activities for members of its user communities. The CLARIN-D consortium regularly contributes courses to the annual European Summer University in Digital Humanities, which has been organized by Elisabeth Burr at Leipzig University since 2009. CLARIN-D has also built strong ties to DARIAH-DE and is currently engaged in CLARIAH-DE, which aims at unifying the data and services offered by the two research infrastructures.

Figure 23: The members of CLARIN-D (starting upper-left corner): Erhard Hinrichs, Christian Thomas, Elke Teich, Marie Annisius and Ulrike Fußbahn, Antonina Werthmann, Gerhard Heyer, Christoph Draxler, Kristin Bührig, Nathalie Walker, Andreas Witt, Andreas Nolda, Thomas Eckart, Dirk Goldhahn, André Blessing, Melanie Grumt Suárez, Bernhard Fisseni, Florian Schiel, Alexander Geyken, Jutta Bopp, Felix Helfer, Thorsten Trippel, Lydia Müller, Piotr Bański, and Jonas Kuhn.

Tool | **WebLicht and WebMAUS**

Written by **Marie Hinrichs** and **Christoph Draxler,** edited by **Nathalie Walker**

WebLicht ("Web-based Linguistic Chaining Tool") is an environment for building and executing chains of natural language processing tools, with integrated capabilities for visualizing and searching the resulting annotations.[14] It is hosted by the CLARIN centre at the University of Tübingen.

One of the main goals of WebLicht is to make a wide range of text processing tools, such as tokenizers, part-of-speech taggers and syntactic parsers, easily accessible to researchers in the humanities and Social Sciences. WebLicht's annotation tools can be invoked via any web browser, without the need for local software installation or any prior familiarity with the tools. Researchers can select predefined processing chains, called "Easy Chains", that have been created for the most common annotations and languages. However, custom processing chains can also be easily generated.
The user is guided through each tool choice, where only tools that are valid for the current annotation task in the processing chain are made available for selection. This is made possible by detailed metadata about the input requirements and output annotations of each tool, and ensures that custom processing chains are always valid. CLARIN-D has also prepared a set of illustrative use cases and annotation examples which showcase how new users can get started with the tool.

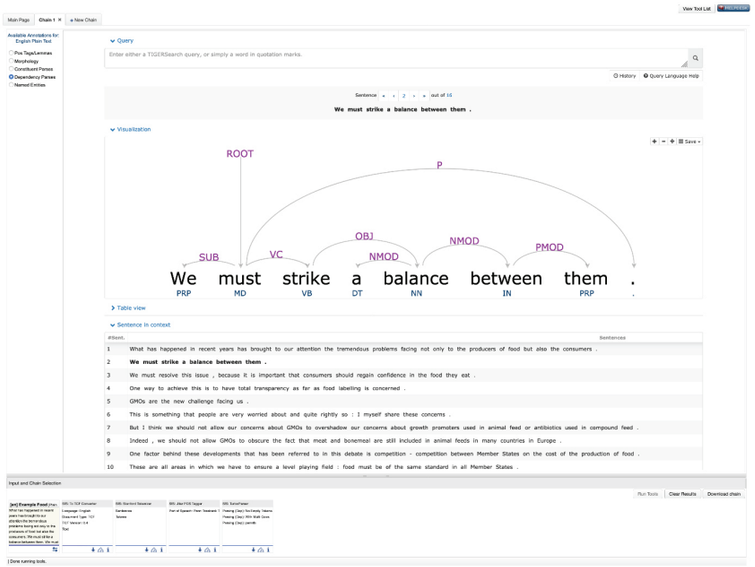


Figure 24: Dependency Parsing in WebLicht.

[14] https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page

WebLicht is tightly integrated into the CLARIN infrastructure. It uses information from the Centre Registry to harvest tool metadata from all CLARIN centre repositories. The tool metadata from the Centre Registry is automatically harvested several times each day, ensuring that all tool information is up to date. WebLicht also supports logging in with CLARIN Federated Identity, which allows researchers to log in through their academic institutions and makes the service available to researchers from thousands of institutions.

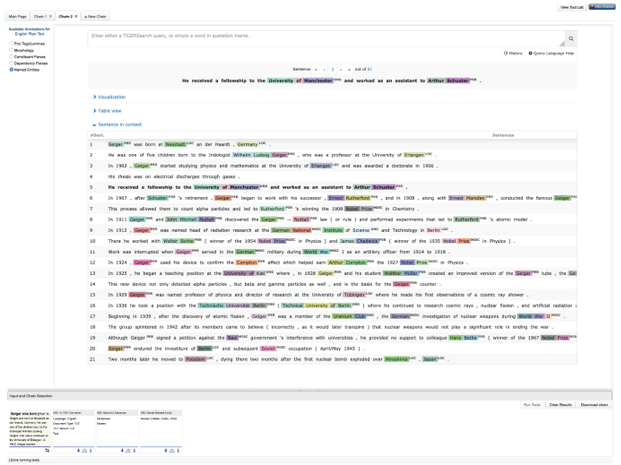


Figure 25: Named-Entity Recognition in WebLicht.

Another highly relevant and closely related CLARIN-D tool is WebMAUS, a web service for automatic word and phoneme alignment,[15] developed by the Bavarian Archive for Speech Signals. It is part of the suite of BAS web tools for speech processing and provides word and phoneme alignment for more than 25 languages, including several dialects, and even a language independent alignment mode based on phonemic transcripts.

Most aligners are based on forced alignment, i.e., they map a given sequence of phonemes to a signal file. WebMAUS takes a different approach: from a set of pronunciation rules and a language model it generates a large number of phoneme sequences, and then returns the sequence that best matches the signal file. Thus, WebMAUS captures phonetic variation caused by e.g. coarticulation, regional variation or speaking style. In inter-rater comparisons, WebMAUS achieves up to 95% of human transcriber performance, and in an evaluation of automatic aligners for Swiss parliamentary speech, MAUS outperformed the other aligners in terms of boundary precision.

[15] https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface

Originally developed for phonetic analysis of speech, WebMAUS has seen growing interest from communities as diverse as those working in speech technology development, language documentation, and research in oral history. Each new application area has led to important extensions and improvements of the service. For example, many widely used annotation tools for oral corpora, such as the Emu Speech Database System, ELAN, EXMARaLDA, and Octra, integrate access to WebMAUS, which greatly facilitates their transcription tasks.

Work on WebMAUS continues at BAS, and CLARIN-D is actively and closely collaborating with speech researchers and potential users all over the world. Recently, the first tone language – Thai – has been added, as well as six different Swiss German dialects. The CLARIN BAS team also encourages anyone who works with a language not yet covered by WebMAUS to get in touch so it can be added to the service.

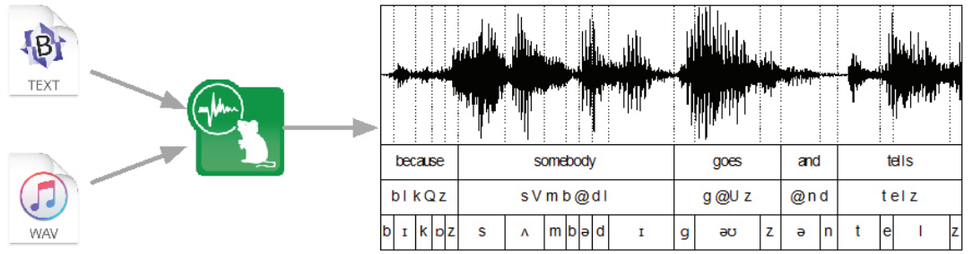


Figure 26: Schematic description of WebMAUS input and resulting multi-level time-aligned transcript. Note that the sequence “and tells” is produced as [a n t e l z].

References:

Dima, E., Hinrichs, E., Hinrichs, M., Kislev, A., Trippel, T., and Zastrow, T. 2012. Integration of WebLicht into the CLARIN Infrastructure. In *Proceedings of the Joint CLARIN-D/DARIAH Workshop at Digital Humanities Conference 2012: Service-oriented Architectures (SOAs) for the Humanities: Solutions and Impacts*, 17–23.

Hinrichs, E., Hinrichs, M., and Zastrow, T. 2010. WebLicht: Web-Based LRT Services for German. In *Proceedings of the Systems Demonstrations at the 48th Annual Meeting of the Association for Computational Linguistics (ACL-2010)*, 25–29.

Kisler, T., Schiel, F., and Sloetjes, H. 2012. Signal processing via web services: the use case WebMAUS. In *Proceedings of Digital Humanities Conference 2012*, 30–34.

Kisler, T., Reichel, U., and Schiel, F. 2017. Multilingual processing of speech via web services. *Computer Speech and Language* 45: 326–347.

Resource | **German Reference Corpus (DeReKo) and German Text Archive (DTA)**

Written by **Marc Kupietz** and **Christian Thomas,** edited by **Nathalie Walker**

German reference resources are flagships in the CLARIN-D consortium, with the German Reference Corpus DeReKo for more modern language data and the German Text Archive DTA for historic language data.

One of CLARIN-D's most important resources is the German Reference Corpus DeReKo,[16] which has been built and maintained at the Leibniz Institute for the German Language (IDS), now a certified CLARIN B-centre, since its foundation in the mid-1960s. It continuously samples the contemporary German language use from around 1950 onwards in a stratified fashion, and thus serves primarily as an empirical basis for synchronous German linguistics.

The DeReKo archive currently contains almost 47 billion words (with a growth rate of three billion words per year) from a variety of genres, ranging from newspaper texts from all areas of German-speaking countries, over fiction and specialized texts, and to Usenet news and Wikipedia talk pages. The entire archive is equipped with several morphosyntactic (see Area D in Figure 27) as well as dependency (see Area E) and constituency annotations (currently only Stanford CoreNLP, not shown in the figure).

[16] https://www1.ids-mannheim.de/kl/projekte/korpora/

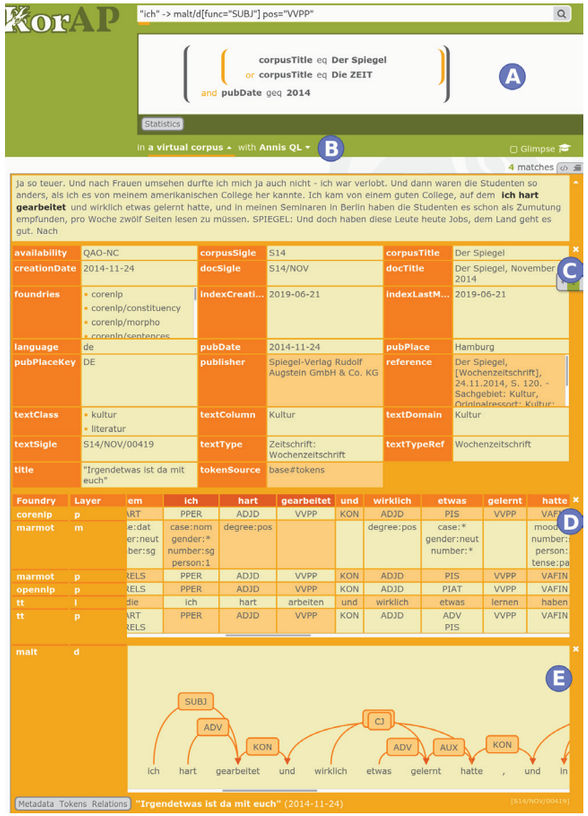


Figure 27: ANNIS query on the Malt annotations of a DeReKo virtual corpus using KorAP, showing the virtual corpus builder (A), the query language selector (B) and expanded views for the metadata (C), token annotations (D) and Malt dependency annotation of a search result.

One of the distinctive features of DeReKo is that it invites users to compile their own stratified virtual subcorpora on the basis of extra-textual metadata (see area C in Figure 27 for a subset) using, for example, KorAP's virtual corpus builder (Area A). This enables samples that are as representative as possible with regard to specific linguistic research questions on, for instance, the diachronic differences between variants of German and specific language domains, like Austrian German used in the newspapers of the 1990s, or German in computer mediated communication. Currently, more than 50,000 linguists use DeReKo free of charge via the analysis platform COSMAS II and the open source analysis platform KorAP, both of which are also available through the IDS Mannheim centre. Since September 2019, DeReKo can also be used via a library for the programming language R, making it easy to perform and visualize quantitative analyses in a reproducible fashion.

Being part of the IDS CLARIN B-centre in Mannheim, DeReKo has been integrated into the CLARIN infrastructure from the outset. DeReKo uses and implements many of the standards and best practices developed within CLARIN and can be accessed partially via CLARIN's Federated Content Search (CLARIN-FCS) and WebLicht.

Users interested in language material from before 1950 will find another resource provided within CLARIN-D extremely useful: the Deutsches Textarchiv ("German Text Archive", DTA).[17] Hosted by the CLARIN centre at the Berlin-Brandenburg Academy of Sciences and Humanities, the DTA is the largest single corpus of historical New High German covering the period from the 16th to the early 20th centuries, comprising more than 350 million tokens in 1.34 million digitized pages. Focusing mostly on (digitized) printed material, the DTA also includes a growing number of hand-written documents. Specialty subcorpora include historical newspapers and other periodicals. The DTA as a whole covers a rich variety of fiction and non-fiction texts, the latter including academic as well as non-academic writing.

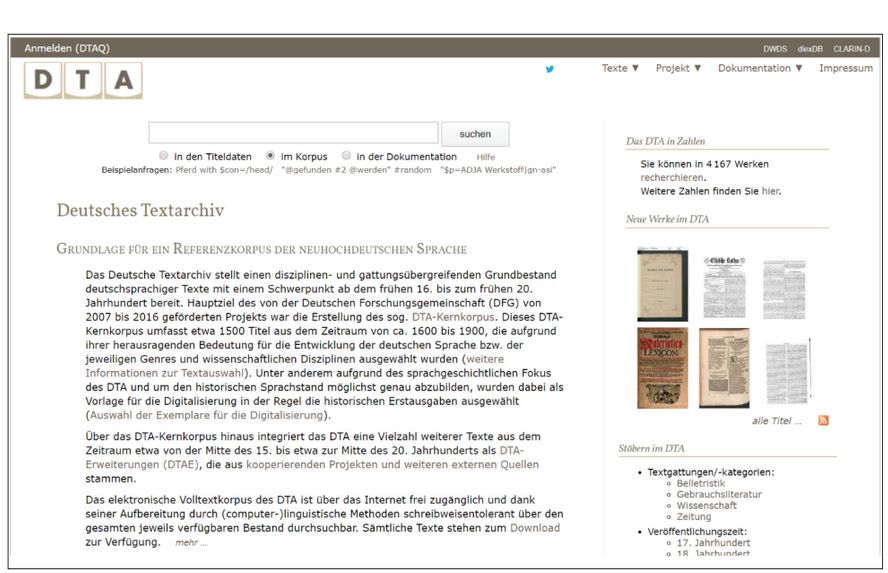


Figure 28: Landing page of the DTA.

[17] http://www.deutschestextarchiv.de/

The DTA is composed of the DTA-Kernkorpus (DTAK, "DTA Core Corpus") with approximately 1,500 first editions from the 16th to the 19th centuries. Additionally, the DTA-Erweiterungen (DTAE, "DTA Extensions") module contains specialty corpora and individual texts which have been curated in the context of CLARIN-D and other projects. The full-text sources provided by digitization projects and other discipline-specific initiatives have been (manually or semi-automatically) converted to a TEI-compatible XML format conforming to the DTA-Basisformat (DTABf, "DTA Base Format") guidelines, including extensive metadata on the original sources and data preparation. OCR texts in the DTA Core Corpus – as well as numerous additional text resources – have been manually corrected. A continuous quality assurance process is made possible by the collaborative web-based platform DTAQ, with around 2000 currently registered users. All DTA corpora are prepared for user consumption by automated computational linguistic analysis methods, including not only PoS-tagging and lemmatization, but also – among other approaches – the orthographic normalization of historical spelling variants, allowing users to formulate queries in modern orthography. Like DeReKo, the DTA is fully integrated into the CLARIN infrastructure, and can for instance be accessed via VLO, Federated Content Search, Language Resource Switchboard, and WebLicht. Among many other resources available from the CLARIN-D community, the DeReKo and the DTA datasets are flagships with tens of thousands of users.

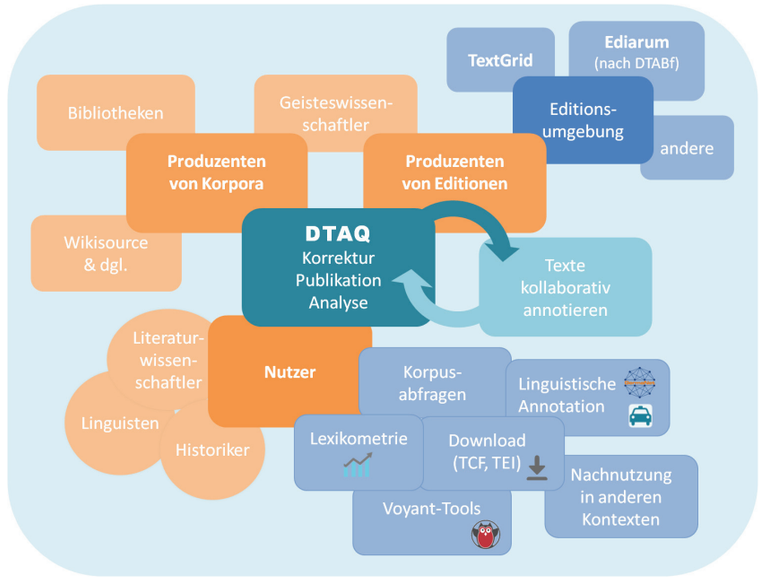


Figure 29: The Deutsches Textarchiv / German Text Archive: an integrated research platform. (Illustration from Geyken et al. 2018.)

References:

Bański, P., Bingel, J., Diewald, N., Frick, E., Hanl, M., Kupietz, M., Pęzik, P., Schnober, C., and Witt, A. 2013. KorAP: the new corpus analysis platform at IDS Mannheim. In *Human Language Technologies as a Challenge for Computer Science and Linguistics. Proceedings of the 6th Language and Technology Conference*, 586–587.

Boening, M., and Haaf, S. 2019. Aggregating resources in CLARIN: FAIR corpora of historical newspapers in the German Text Archive. In *Proceedings of CLARIN Annual Conference 2019*, 124–128.

Fischer, F., Haaf, S., and Hug, M. 2019. The best of three worlds: Mutual enhancement of corpora of dramatic texts (GerDraCor, German Text Archive, TextGrid Repository). In *Proceedings of CLARIN Annual Conference 2019*, 97–103.

Geyken, A., Boenig, M., Haaf, S., Jurish, B., Thomas, C., and Wiegand, F. 2018. Das Deutsche Textarchiv als Forschungsplattform für historische Daten in CLARIN. In *Digitale Infrastrukturen für die germanistische Forschung*, 219–248.

Jurish, B., and Nieländer, M. 2019. Using DiaCollo for historical research. In *Proceedings of CLARIN Annual Conference 2019*, 40–43.

Kupietz, M., Belica, C., Keibel, H., and Witt, A. 2010. The German Reference Corpus DeReKo: A primordial sample for linguistic research. In *Proceedings of the seventh conference on International Language Resources and Evaluation (LREC 2010)*, 1848–1854.

Kupietz, M., Lüngen, H., Kamocki, P., and Witt, A. 2018. The German Reference Corpus DeReKo: New Developments – New Opportunities. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 4353–4360.

Event | **An Internship at CLARIN-D**

Written by **Nathalie Walker**

One central focus of CLARIN-D is the support of early-stage researchers in the Digital Humanities. Apart from our active participation in the ESU, the European Summer School in Digital Humanities, which takes place each year in Leipzig, we engage in a number of User Involvement Activities on a regular basis: we offer information booths at various conferences and workshops, special courses and conventions for PhD candidates, scholarships for researchers to participate in conferences or workshops, use cases which showcase the use of tools and services, and internships at our eight different CLARIN-D centres.

Eva Huber, who is currently enrolled in the MA programme "Computational Linguistics" at the University of Tübingen, is a perfect example to show how early-stage researchers can profit from CLARIN-D's offers. After completing a BA in English Linguistics at the University of Manchester, Eva came to the University of Tübingen for a two-month-internship at the Tübingen CLARIN-D centre. She had always been fascinated by the concept of research infrastructures, so the internship was a great opportunity to look behind the scenes of such an infrastructure and to apply her knowledge and skills in her work with GermaNet,[18] a machine-readable lexical-semantic network for German maintained by CLARIN-D. During her internship, she integrated Swiss Standard German terms into GermaNet, with which she expanded her knowledge on computational linguistics and gained her first hands-on programming experience after her more theory-based BA. This led to her decision to pursue an MA with a greater focus on the computational side of linguistics. Now that her internship is completed, Eva has also been able to continue her work on GermaNet and other CLARIN-D projects through her job as a research assistant in CLARIN-D.

Figure 30: Eva Huber.

[18] http://www.sfs.uni-tuebingen.de/GermaNet/

The internship also allowed Eva to participate in a number of CLARIN events: at the ESU 2018 in Leipzig, for instance, Eva took part in two workshops offered by members of CLARIN-D. Nils Reiter and Sarah Schulz's workshop on "Reflected Text Analysis in the Digital Humanities" as well as Tommi A. Pirinen's "The Humanities Scholar's Perspective on Rule-based Machine Translation" expanded Eva's skills in programming and convinced her that her future lay in the realm of the Digital Humanities. User Involvement activities such as the workshops at the ESU are a regular CLARIN-D feature and allow early-stage researchers to not only deepen their knowledge in various relevant fields, but also get to know other researchers, both young and senior, and thus expand their networks significantly.

Figure 31: Participants of the ESU 2018.

Attending the CLARIN Annual Conference in Pisa in 2018 has been one of the personal highlights of Eva's university career so far: she enjoyed meeting other researchers from different backgrounds and got to know the internal aspects of research infrastructures on a more practical and hands-on level. Her internship also led to her participation in the 10th Global WordNet Conference 2019 in Wrocław, where she presented her first paper together with CLARIN-D national coordinator Erhard Hinrichs on "Including Swiss Standard German in GermaNet".

Interview | **Eva Gredel** and **Yana Strakatova**

Eva Gredel is a postdoctoral researcher, and currently works as a substitute professor at the Chair of German Linguistics of the University of Mannheim.

Yana Strakatova is a PhD student at the University of Tübingen who works in the project "MoCo – Lexical-Semantic Modelling of Collocations".

Could you tell us a little bit about yourselves, your academic background and current work?

<

Eva Gredel (EG): My name is Eva Gredel and I am a postdoctoral researcher at the Chair of German Linguistics of the University of Mannheim. After my first degree in German, Romance and Media and Communication Studies, I received my doctorate at the University of Mannheim in 2014. In my doctoral thesis I evaluated corpus linguistic methods for discourse analysis. Currently, I am a substitute professor at the Chair of German Linguistics in Mannheim, and responsible for the coordination of the linguistic courses offered in German Studies. In Mannheim, teaching in linguistics is not only carried out by lecturers of the university, but also by colleagues from the Mannheim Leibniz Institute for the German Language (IDS), which is a CLARIN-D centre.

Yana Strakatova (YS): My name is Yana Strakatova, and I received my first degree in Russia at the Vladimir State University, where I studied English and German and received a diploma in teaching these languages. My education in the field of linguistics really started in Germany at the University of Dresden, where I completed the master's programme in European Languages. The syllabus had a few courses in corpus linguistics and programming that got me interested in computational

linguistics. My master's thesis was interdisciplinary: it was situated at the intersection of neonatology, psychology, and linguistics. I focused on the linguistics part and studied the language of birth stories using corpora and statistics. The variety of research directions in the field of corpus and computational linguistics motivated me to pursue a PhD. I am currently in the last year of my doctoral studies at the University of Tübingen, where I work on the project MoCo – Lexical-Semantic Modelling of Collocations funded by the German Research Foundation (DFG).

>

How did you hear about CLARIN-D and how did you get involved?

<

EG: Linguistics at the University of Mannheim has a corpus linguistic focus in research and teaching. In teaching, the lecturers use the CLARIN corpus infrastructure that is developed at the IDS in Mannheim. I therefore came into contact with CLARIN resources and the CLARIN infrastructure at a very early stage of my own studies, and used them for empirical language analysis at all qualification levels.

In my PhD thesis, I then evaluated the usability of various CLARIN resources for discourse linguistic studies in the tradition of Foucault. In my postdoctoral project, I now extend discourse-linguistic models in such a way that they can be used to analyse digital discourses, such as those on Wikipedia. In 2016, 14 colleagues and I applied for a project with the German Research Foundation (DFG) in which corpus linguistic approaches to digital discourses also play a central role. In the context of the network, I was invited to participate in CLARIN-D working group 1, "German Philology". The task of this working group is to set up documentation for applications of digital data and tools for research questions (e.g., screencasts with use cases).

>

How has CLARIN-D influenced your way of working? How does your research benefit from the CLARIN-D infrastructure?

<

EG: The CLARIN-D infrastructure has always been a good starting point for my own study projects and qualification work to analyse language empirically. In my PhD thesis, I investigated newspaper language for metaphorical patterns in media discourses with a focus on the discourse object "virus" – a topic that has just taken on a whole new topicality and relevance – and evaluated, on a methodological level, various CLARIN-D resources with regard to their potential for such questions.

YS: CLARIN-D provides resources and tools for researchers with different backgrounds and research goals. I have never studied computer science and find myself somewhere in between computational and theoretical linguistics. The tools that assist me in my PhD project have been created in a way that anyone can benefit from using them by adjusting them to their needs. If there is a word or phrase that seems ambiguous and causes a lot of discussion, the user-friendly interface can give a quick answer to resolve the issue. I can, moreover, make use of the raw data and obtain the statistics or certain combinations and patterns I am interested in.

Moreover, CLARIN-D provides an opportunity to store and share the new data collected and annotated in the research process. One of the crucial steps in my PhD was creating a gold standard dataset of collocations that we could use for different experiments. Its creation was based on the tools and resources already integrated in the CLARIN-D infrastructure, and now this dataset (codenamed "GerCo") can be used by other researchers interested in the topic.

CLARIN-D is not just a collection of resources, it is also a community that offers great opportunities for exchanging experience and knowledge.

In the first year of my PhD, I took part in the CLARIN Annual Conference in Pisa. If I am not mistaken, it was the first time there was a PhD students' poster session at the conference and it was a success. I presented some experiments I conducted for our project and got a lot of useful feedback from more experienced researchers.

>

Which CLARIN-D tools and corpora have you used and how did you integrate them into your existing research?

<

EG: Several CLARIN resources play a central role in my research projects and publications. In my PhD thesis, I evaluated corpora that are available via the German Reference Corpus (DeReKo) and the Digital Dictionary of the German Language (DWDS) which contain newspaper texts, while the focus of my postdoctoral projects has shifted to corpora of computer-mediated communication (CMC corpora). In particular the Wikipedia corpora in DeReKo, provided by the IDS, are the basis of my current research projects. In case studies, I research how Wikipedia can be analysed as a discursive space. Not only are the encyclopaedic texts of Wikipedia (article pages) relevant here, but also the hypertextually linked pages (talk pages), where Wikipedia authors exchange ideas about the collaborative text production. In particular, the subcorpora of talk pages provide empirical access to internet-based communication and, for example, to Wikipedia-specific net jargon, which is characterized by numerous word formation products.

YS: In the MoCo project, I investigate the semantic properties of collocations, i.e., combinations of two words that co-occur together more often than by chance, such as “black coffee”. For now, we focus on the German language since there are several lexical resources that can serve as an empirical basis for our research. The Digital Dictionary of the German Language (DWDS), a digital resource developed at the CLARIN-D centre BBAW (Berlin-Brandenburg Academy of Sciences and Humanities), contains numerous German corpora, statistical data based on these, detailed dictionary entries with grammatical, etymological and semantic information and examples from the corpora. We work a lot with polysemous words and use the DWDS dictionary for word sense disambiguation. We used Wortprofil to collect the data for our research, a SketchEngine-like application of the DWDS. Wortprofil provides ranked lists of co-occurrences based on the statistical data from the DWDS corpora. This kind of automatic preprocessing made it feasible to collect and study a lot of data rather quickly. Another CLARIN-D resource that is a part of my daily work is the German wordnet GermaNet. This provides information about the lexical and conceptual relations between words. All the entries in GermaNet are manually created by experts in the field, and thus are very reliable.

>

To what extent are these corpora and tools important for your research, and what is specific to them?

<

EG: The above-mentioned Wikipedia corpora of the IDS serve as a database for empirical case studies; for example, I use them to analyse the distribution of word formation products in the digital discourse of Wikipedia. A feature of these corpora is that the different types of talk pages of Wikipedia are characterized by CMC features, and these different types of pages are each available as subcorpora. The availability of subcorpora containing article talk pages, user talk pages and redundancy talk pages allows for extensive analysis of language use in digital discourses on Wikipedia. Another special feature of the available Wikipedia corpora is that the Leibniz Institute for the German Language Wikipedia provides corpora in eight additional languages (including English, French, and Spanish). This makes it possible to evaluate Wikipedia texts linguistically, even in contrastive studies.

YS: All the resources I described above were not only used at a certain stage in the development of my PhD project: I use them every day to resolve new issues and confirm new hypotheses. Those resources provide not only the empirical data for my research, but also serve as the basis for developing a theoretical framework for describing this data. I get the statistics about the use of natural language from the corpora, and all the lexical and semantic information comes from the work of many experts in lexicography, lexical semantics, and syntax. I know I can rely on this data and it allows me to make quick progress in my own research and make new discoveries.

>

What are the methodological and technical challenges that you face in your particular field?

<

EG: One challenge is certainly the enormous complexity of discourses on Wikipedia, which arises due to their non-linearity – specifically with regard to links. Thus, there are numerous hyperlinks between the above-mentioned types of talk pages that are relevant for analysing digital discourses. However, by splitting the Wikipedia data into subcorpora according to the different types of talk pages – which makes sense in many places, both technically and in terms of content – these links and thus the complexity of the discourses are to a certain extent lost.

YS: In computational linguistics, we need large amounts of text data for training and testing different models. Getting access to the required data can sometimes be very challenging, mainly due to the strict copyright laws. Another challenge is that language is not static; rather, it is constantly changing, so there can never be a resource that can be considered complete and account for all the variety there is in a language.

>

Which CLARIN-D resources, tools and services would you recommend to your colleagues?

<

EG: I can unreservedly recommend the CLARIN resources and infrastructure described above for language analysis to my colleagues.

YS: Apart from the resources and tools that I have already described, I would definitely recommend taking a look at the variety of tools provided by WebLicht. It can assist researchers in such tasks as tokenizing, PoS-tagging, and parsing text corpora.

>

Do you integrate CLARIN-D tools and resources into your university teaching? If so, how exactly do you do this?

<

EG: CLARIN-D tools and resources play a very important role in my university teaching. In my seminars, I introduce students to empirical work with corpora and show how, for example, morphological analyses can be performed using corpora. Screencasts with use cases, which CLARIN-D has made publicly available on YouTube, for instance, have proven to be very helpful. Students and learners in general can then follow in detail how individual CLARIN tools and resources can be used for corpus linguistic studies. The spectrum of possible studies ranges from grammatical and morphological to semantic questions. Students greatly appreciate the empirical work using corpora and achieve impressive results. For example, it is very common to investigate neologisms using CLARIN resources.

>

What was the latest course that you taught with CLARIN-D resources and tools, and what were the students' reactions?

<

EG: Apart from my current course "Media Linguistics" in the spring term 2020, my last course in which I accessed corpora with students was the seminar "Digital Discourses" in the autumn term 2019. In this seminar, the students were very much involved in the corpus linguistic analysis of language using CLARIN resources. Some very good student research papers were produced.

>

What would you recommend to students who are interested in the Digital Humanities?

<

EG: For students interested in the Digital Humanities it is, of course, first of all useful to get a good overview of existing resources and infrastructures, and CLARIN's offers and services provide an this to a very high level. Secondly, it is important to reflect on the specific characteristics of the available language material and how the respective infrastructure deals with these (for example, the availability of Wikipedia subcorpora for the different types of (talk) pages). In the case of Wikipedia corpora, for example, it is then relevant to understand that language on the article pages of Wikipedia is fundamentally different from the language on the talk pages. CMC features of Wikipedia texts can only be analysed very limitedly using the subcorpora of article pages, which are characterized by a special encyclopaedic style. For questions in this area, it is useful to use the subcorpora for talk pages, where Wikipedia authors engage in social interaction. Finally, it is then relevant to match good research questions with the appropriate resources.

>

What is your vision for CLARIN-D and the Digital Humanities 10 years from now?

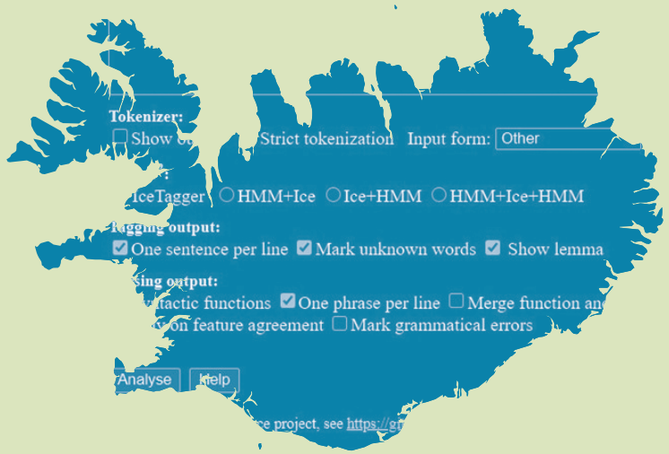
<

EG: One great wish would certainly be to continue working on the idea of building multimodal corpora. With Wikipedia as the object of investigation, for example, the relevance of image material becomes clear in many areas, and it would often be appropriate to take the multimodality of the data material into account. The further expansion of CMC corpora that allow contrastive analyses would also be a great development.

YS: I would like to a see a uniform user-friendly platform connecting all the resources and tools and providing free access for all researchers. I am sure that a lot of researchers would benefit immensely from a well-structured environment with clearly defined pipelines for the basic tasks. In that way, one would have more time for more complicated tasks.

>

ICELAND



Introduction

Written by **Eiríkur Rögnvaldsson**

Iceland first joined CLARIN ERIC as an observer in November 2018, but after a new law was passed by the Icelandic Parliament on European Research Infrastructure Consortia in 2019, it was able to apply for full membership, which was approved in February 2020.[19] The membership agreement was signed on March 10, 2020. The Icelandic consortium is led by the Árni Magnússon Institute for Icelandic Studies, and Professor Emeritus Eiríkur Rögnvaldsson is the National Coordinator.

The following institutions participate in the CLARIN-IS consortium and have signed a memorandum of partnership:

- The Árni Magnússon Institute for Icelandic Studies (leading partner)
- The University of Iceland
- Reykjavík University
- National and University Library of Iceland
- National Archives of Iceland
- The Icelandic Language Council
- The Icelandic National Broadcasting Service
- Almannarómur – Consortium for Language Technology

[19] https://clarin.is/en/

The CLARIN-IS office is based in Reykjavík, where it shares the premises with the Árni Magnússon Institute's Language Technology Unit, with which it cooperates closely. The staff consists of the National Coordinator who works part-time, and a computer scientist, Samúel Þórisson, who works full time.

In the first year of our CLARIN ERIC membership, the main tasks of CLARIN-IS have been to establish a National Consortium and to build a Metadata Providing Centre (CLARIN C-centre) which hosts metadata for Icelandic language resources and makes them available through the Virtual Language Observatory. Furthermore, we have now established a repository which already hosts a number of tools and resources.

In connection with Iceland's participation in the META-NORD project from 2011–2013, the Árni Magnússon Institute established a local website, Málföng ("Language Resources"), where the institute's language resources and tools were stored and made accessible. The website also contains links to several resources and tools owned by others. Most of the institute's tools and resources have now been made available through the CLARIN-IS website, and we are in the process of preparing them for archiving in our repository by adapting them to standards, writing metadata, and so on.

A number of our resources are already widely used, both by researchers and the general public. For instance, the Database of Modern Icelandic Inflection (DMII) is a multipurpose linguistic resource which contains inflectional paradigms, with a vocabulary of 300,000 lemmas and 6.5 million inflectional forms. The online version is very popular among the general public. The Saga Corpus contains 49 Old Icelandic narrative texts, approx. 1.7 million words in total. The spelling has been normalized to Modern Icelandic spelling, and some inflectional endings have been changed to Modern Icelandic form. The Icelandic Gigaword Corpus (IGC) is a tagged and lemmatized corpus of Modern Icelandic containing approximately 1,550 million running words of text. Each running word is accompanied by a morphosyntactic tag and lemma, and each text is accompanied by bibliographic information. Both the Saga Corpus and the Gigaword Corpus can be queried online, using the Korp corpus tool developed by Språkbanken in Gothenburg.

We expect our repository to expand considerably in the coming years, especially with resources and tools developed within the Icelandic National Language Technology Programme which started in 2019 and will run for five years. The Ministry of Education, Science and Culture, which funds the programme, demands that all its deliverables be submitted to CLARIN-IS and made accessible under maximally open licences. Thus, one of our main tasks in the next few years will be to validate these tools and resources, archive them, and make them openly available.

Figure 32: CLARIN-IS staff – Eiríkur Rögnvaldsson (right) and Samúel Þórisson (left).

Tool | **IceNLP**

Written by **Eiríkur Rögnvaldsson** and **Hrafn Loftsson**

IceNLP is an open source toolkit for processing and analysing Icelandic text. It can be downloaded from the CLARIN-IS repository under the GNU General Public Licence, version 2.[20] An online version is also available on the Reykjavík University website. IceNLP was originally developed by Hrafn Loftsson, Associate Professor in Computer Science at Reykjavík University, as a part of his PhD studies during the years 2004–2007. Since then, students at Reykjavík University and the University of Iceland have helped in developing individual components. IceNLP is written as a collection of Java classes. Its main modules are the following:

- A tokenizer. This module performs both word tokenization and sentence segmentation.
- A morphological analyser (IceMorphy; Loftsson 2008). The program provides the tag profile (the ambiguity class) for known words by looking up words in its dictionary. The tag profile for unknown words is guessed by applying rules based on suffixes and endings.
- A linguistic rule-based PoS tagger (IceTagger; Loftsson 2008). The tagger produces disambiguated morphosyntactic tags. It uses IceMorphy for morphological analysis and applies both local rules and heuristics for disambiguation.
- A statistical PoS tagger (TriTagger). This trigram tagger is a re-implementation of Brandt's well-known HMM tagger (TnT).
- A lemmatizer (Lemmald; Ingason et al. 2008). The method used combines a data-driven method with linguistic knowledge to maximize accuracy.
- A shallow parser (IceParser; Loftsson and Rögnvaldsson 2007). The parser marks both constituent structure and syntactic functions using a cascade of finite-state transducers.

Most of these modules were the first of their kind to be developed for Icelandic. Individual components of IceNLP can be run independently, or the Java clusters in question connected directly to software that is being developed.

Hrafn Loftsson and his students and collaborators have used IceNLP in a number of projects and publications; for instance, in developing intelligent computer-assisted language learning applications (ICALL; Volodina et al. 2012). They have also used the toolkit for experimenting with the Apertium machine translation system where they replaced some of the Apertium original modules with modules from IceNLP (Brandt et al. 2011).

[20] http://hdl.handle.net/20.500.12537/8

Other researchers and developers have also made extensive use of IceNLP. It was used for preprocessing texts in the manually annotated one-million-word Icelandic Parsed Historical Corpus (IcePaHC; Rögnvaldsson et al. 2012), which is accessible via CLARIN-IS and has proven crucial both for studying Icelandic diachronic syntax and both developing and training parsers for Icelandic. IceNLP was also used for tokenizing, lemmatizing and tagging the Tagged Icelandic Corpus (Mörkuð íslensk málheild, MÍM; Helgadóttir et al. 2012), a 25 million-word balanced tagged corpus of Modern Icelandic that is also accessible via CLARIN-IS. Furthermore, IceNLP was an essential tool in the development of the first Icelandic Frequency Dictionary of children's speech (Einarsdóttir et al. 2019).

IceNLP is the only toolkit that comprises several different tools for analysing Icelandic text, and also the only one that is available online (Figure 33). Thus, it is a very important tool for researchers working on Icelandic, especially those who are not very technically oriented.

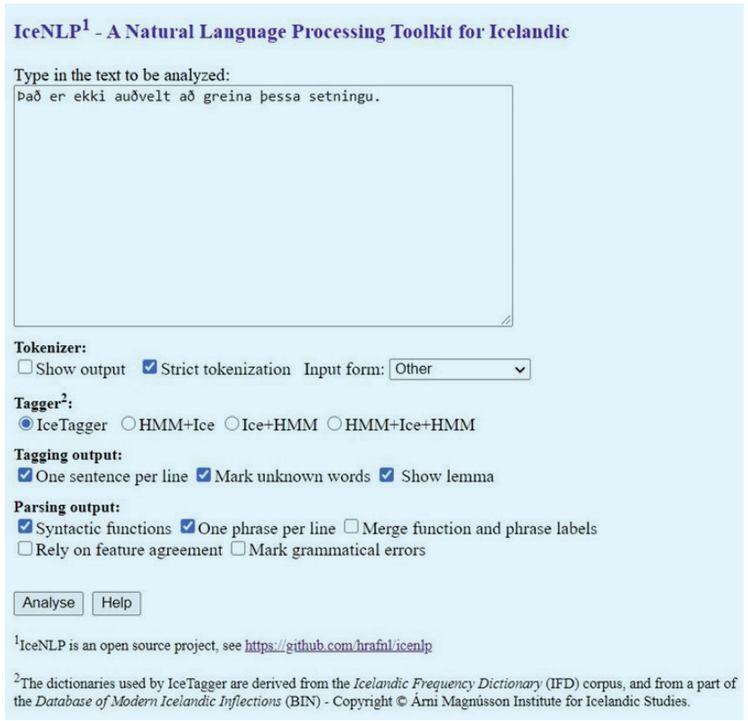


Figure 33: The user interface for the online version of IceNLP (also available in Icelandic).

References:

Brandt, M.D., Loftsson, H., Sigurþórsson, H., and Tyers, F.M. 2011. Apertium-IceNLP: a rule-based Icelandic to English machine translation system. In *EAMT 2011: Proceedings of the 15th Conference of the European Association for Machine Translation*, 217–224.

Einarsdóttir, J.T., Pétursdóttir, A.L., and Rúnarsdóttir, Í.D. 2019. Tíðni orða í tali barna. [Word frequency in children's speech.] Reykjavík: Háskólaútgáfan.

Helgadóttir, S., Svavarsdóttir, Á., Rögnvaldsson, E., Bjarnadóttir, K., and Loftsson, H. 2012. The Tagged Icelandic Corpus (MÍM). In *Proceedings of Language Technology for Normalization of Less-Resourced Languages*, workshop at LREC 2012, 67–72.

Ingason, A.K., Helgadóttir, S., Loftsson, H., and Rögnvaldsson, E. 2008. A Mixed Method Lemmatization Algorithm Using a Hierarchy of Linguistic Identities (HOLI). In *Advances in Natural Language Processing*, 205–216.

Loftsson, H., and Rögnvaldsson, E. 2007. IceParser: An Incremental Finite-State Parser for Icelandic. In *NODALIDA 2007 Conference Proceedings*, 128–135.

Loftsson, H. 2008. Tagging Icelandic text: A linguistic rule-based approach. *Nordic Journal of Linguistics* 31 (1): 47–72.

Rögnvaldsson, E., Ingason, A.K., Sigurðsson, E.F., and Wallenberg, J. 2012. The Icelandic Parsed Historical Corpus (IcePaHC). In *Proceedings of LREC 2012*, 1978–1984.

Volodina, E., Borin, L., Loftsson, H., Arnbjörnsdóttir, B., and Leifsson, G.Ö. 2012. Waste not, want not: Towards a system architecture for ICALL based on NLP component re-use. In *Proceedings of the SLTC 2012 workshop on NLP for CALL*, 47–58.

Resource | **The Database of Modern Icelandic Inflection**

Written by **Kristín Bjarnadóttir** and **Eiríkur Rögnvaldsson**

The Database of Modern Icelandic Inflection (DMII) contains inflectional paradigms and has a vocabulary of 300,000 lemmas with approximately 6.5 million inflectional forms. Uninflected words are also included.[21] Downloadable data for use in language technologies is available from the CLARIN-IS repository under the CC BY-SA 4.0 licence. The download package contains four versions of the data in the CSV format, ranging from a simple list of word forms to data linking lemmas and inflectional forms to grammatical tags and usage information. A detailed description of the data is available on the DMII website.

The DMII was originally created in the context of the initiative to provide Icelandic language technologies at the start of the millennium. The first version of the DMII was a set of XML files with 173,389 paradigms, made available on CDs for use in LT in 2004 (Bjarnadóttir 2012), and individual paradigms have been accessible on the DMII website since the same year. Extensively used by the Icelandic public as a reference on inflection, the website has been very popular from the start – more than 298,000 users viewed over 5.6 million pages in the year starting September 1, 2019.

The DMII Core is a subset of DMII, which contains the core vocabulary of current Icelandic, i.e., common non-domain specific words, and a selection of named Icelandic entities, i.e., personal names, common place names, and a few names of important institutions. The vocabulary contains approximately 58,000 words. Its sources are The Dictionary of Modern Icelandic (íslensk nútímamálsorðabók), containing approximately 50,000 headwords, with additions from the top 50,000 most frequent words (lemmas) of the Gigaword Corpus (Risamálheild). The DMII Core was created to be used for third party publications, and is accessible through a RESTful API that is open to everyone. The API allows users to send simple queries and receive full paradigms in JSON-format as a response.

[21] http://hdl.handle.net/20.500.12537/5

The DMII has been used in a number of different language technology projects. It has proven its usefulness in increasing the accuracy in PoS tagging (Loftsson et al. 2011, Steingrímsson et al. 2019); in the post-processing of OCR texts (Daðason et al. 2014); in linking lexicographic resources (Bjarnadóttir 2016); in developing a high-accuracy lemmatizer for Icelandic (Ingólfsdóttir et al. 2019); in developing Context-Free Grammar for Icelandic (Þorsteinsson et al. 2019).

Furthermore, the DMII is currently being used to develop The Database of Icelandic Morphology (DIM; Bjarnadóttir et al. 2019), which is a multipurpose linguistic resource. Whereas the DMII is descriptive, the DIM is partly prescriptive, i.e., the "correctness" of both words and inflectional forms is marked in accordance with accepted rules of usage. This greatly improves the scope of applications using the data, from the purely analytical possibilities of the DMII (used for examples in search engines, PoS tagging, named-entity recognition, etc.), to the productive possibilities of the DIM, such as correction and formulation of text. The analysis has recently been extended to include genre, style, domain, age, and various grammatical features. Work on error analysis of sub-standard forms is in progress, as is work on an analysis of word formation, including linkups of all constituents to lemmas in the DMII. More components of the DIM will be added to the CLARIN-IS repository as soon as they are finalized.

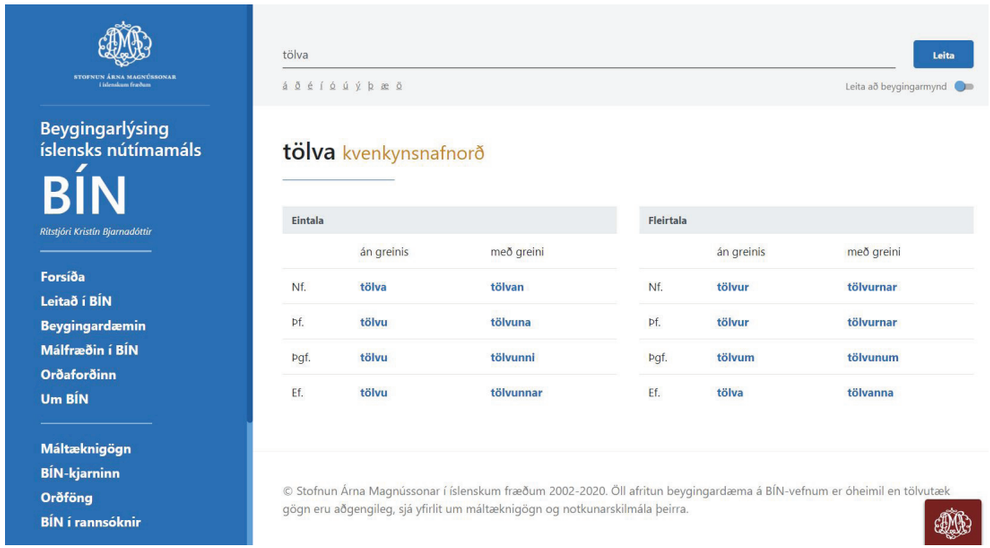


Figure 34: The online version of the DMII, showing the inflection of the word *tölva* ("computer").

References:

Bjarnadóttir, K. 2012. The Database of Modern Icelandic Inflection. In *Proceedings of Language Technology for Normalization of Less-Resourced Languages,* workshop at LREC 2012, 13–18.

Bjarnadóttir, K. 2016. The Case for Normalization: Linking Lexicographic Resources for Icelandic. In *Nordiske Studier i Leksikografi*, 79–88.

Bjarnadóttir, K., Hlynsdóttir, K.I., and Steingrímsson, S. 2019. DIM: The Database of Icelandic Morphology. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, 146–154.

Daðason, J.F., Bjarnadóttir, K., and Rúnarsson, K. 2014. The Journal *Fjölnir* for Everyone: The Post-Processing of Historical OCR Texts. In *Proceedings of Language Resources and Technologies for Processing and Linking Historical Documents and Archives – Deploying Linked Open Data in Cultural Heritage*, workshop at LREC 2014, 56–62.

Ingólfsdóttir, S.L., Loftsson, H., Daðason, J.F., and Bjarnadóttir, K. 2019. Nefnir: A high accuracy lemmatizer for Icelandic. *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, 310–315.

Loftsson, H., Helgadóttir, S., and Rögnvaldsson, E. 2011. Using a Morphological Database to Increase the Accuracy in POS Tagging. In *Proceedings of Recent Advances in Natural Language Processing*, 49–55.

Steingrímsson, S., Kárason, Ö., and Loftsson, H. 2019. Augmenting a BILSTM tagger with a Morphological Lexicon and a Lexical Category Identification Step. In *Proceedings of Recent Advances in Natural Language Processing*, 1161–1168.

Þorsteinsson, V., Óladóttir, H., and Loftsson, H. 2019. A Wide-Coverage Context-Free Grammar for Icelandic and an Accompanying Parsing System. In *Proceedings of Recent Advances in Natural Language Processing*, 1397–1404.

Event | **Launching the National Language Technology Programme**

Written by **Eiríkur Rögnvaldsson**

The Ministry of Education, Science and Culture decided to fund Icelandic membership of CLARIN ERIC because of a proposal made by an expert group who wrote the Icelandic National Language Technology Project Plan published in June 2017. The Project Plan has a special chapter on CLARIN where the importance of Iceland joining CLARIN ERIC is emphasized. In light of this, it was only natural that CLARIN-IS, then an observer in CLARIN ERIC, was asked to be one of the organizers of a conference ("Er íslenskan góður bissness") celebrating the launch of the Icelandic National Language Technology Programme on October 16, 2019.[22]

The aim of the conference was to publicize and promote the LT Programme and create enthusiasm for it among Icelanders, and to make people aware of the importance of language technology for small languages like Icelandic. The conference thus had a very wide target group – researchers from various fields and people from the IT sector, the finance sector, the health sector, media, and the general public. The conference was addressed by the President of Iceland, Mr. Guðni Th. Jóhannesson, the Minister of Education, Science and Culture, Ms. Lilja Dögg Alfreðsdóttir, and the Rector of the University of Iceland, Prof. Jón Atli Benediktsson.

Furthermore, the CLARIN National Coordinator, Prof. Eiríkur Rögnvaldsson, and the Manager of Almannarómur (the Consortium for Language Technology), Ms. Jóhanna Vigdís Guðmundsdóttir, gave an overview of the situation with regard to Icelandic language technology – past, present, and future. In his talk, Eiríkur emphasized the importance of building Icelandic language resources and making them available through CLARIN in order for the LT Programme to be successful. He also illustrated how Icelandic language technology would benefit from multinational cooperation – especially within CLARIN ERIC, but also other European initiatives such as the European Language Resource Coordination (ELRC) and the European Language Grid (ELG). These talks were followed by the launching of a special campaign collecting speech samples from the general public.

After that, a number of people representing the different target groups gave short talks on their experiences with language technology and their vision for the future of Icelandic language technology. The talks were divided into three sessions: Language and Computers, The Icelandic

[22] https://www.hi.is/vidburdir/er_islenskan_godur_bissness

Language and Mass Media, and Language Technology in the Financial Sector. Among the speakers were university professors in computer science, electrical engineering and translation studies, and representatives from the three largest banks in Iceland, the largest IT company in Iceland, the National Broadcasting Service, the National University Hospital, and three start-up companies.

Last but not least, the former CLARIN ERIC Vice Director Bente Maegaard, who is a member of the LT Programme's Expert Panel, together with Estonian National Coordinator Kadri Vider and CLARIN Senior Advisor Steven Krauwer, gave an introductory talk on CLARIN. In her talk, Bente gave an overview of the origins, structure, governance, and aims of CLARIN. She emphasized the benefits of CLARIN for business and industry, and briefly introduced the European Open Science Cloud.

Around 120 people attended the conference. The audience was very mixed, and represented the target group fairly well. The conference received good media coverage on national TV and in the largest newspapers, and was a great success. A video recording of the whole conference is available on YouTube. Our plan is to organize similar events annually for the duration of the LT Programme.

Figure 35: The audience at the "Er íslenskan góður bissness" conference.

Figure 36: Eiríkur Rögnvaldsson giving a talk on the importance of Icelandic language technologies and resources being made available through CLARIN.

Figure 37: Bente Maegaard giving a talk on CLARIN ERIC.

Interview | **Jóhannes Gísli Jónsson**

Jóhannes Gísli Jónsson is Professor of Linguistics at the University of Iceland. He has used the CLARIN-IS Gigaword Corpus and the Icelandic Parsed Historical Corpus for his research into theoretical syntax.

Could you please introduce yourself, and what are your main research interests?

<

I am Professor of Icelandic Linguistics at the University of Iceland. My main research interests are theoretical syntax and the syntax-semantics interface. I have mainly worked on Icelandic and Faroese, and also a little bit on Icelandic Sign Language.

>

What has inspired you to research Icelandic syntax by using corpora?

<

I like to use corpora to get information that goes beyond my native speaker intuitions and raises new questions. Of course, there is no alternative if you are working on older stages of Icelandic for which you do not have any reliable intuitions, but I have found that corpora are also very useful for the study of Modern Icelandic. Still, corpus data must be complemented with experimental data, such as judgement tasks on grammatical constructions that cannot be found in corpora, and you also need a good theoretical framework to make sense of all the data in linguistic corpora.

>

In your research, you have used the Icelandic Gigaword corpus.[23] Could you summarize some of the work you have done on the basis of this corpus, and which features made it crucial for your work?

<

In the past 16 months or so I have been looking at inversion in Icelandic, i.e., word order where the direct object precedes the indirect object in active clauses. An example of this would be *Ég gaf bókina Jóni* "I gave the book (to) John". The new Gigaword Corpus has provided me with a lot of information that I could not have gotten through pure introspection. For instance, it turns out, quite surprisingly, that inversion is much more common with some ditransitive verbs than others. It is for instance much more common with *afhenda* "deliver, hand over" than *gefa* "give". Another surprising finding is that the indirect object is heavier than the direct object in about 90% of the cases in the corpus when it follows the direct object, in the sense that it has more words or a stressed word as opposed to an unstressed pronoun. Thus, the Gigaword Corpus has opened up all kinds of questions that I did not even have at the beginning of this study.

The crucial feature of the Gigaword Corpus for this work is the possibility of searching for strings where the first word is a particular ditransitive verb, followed by some word bearing accusative case and then another word with dative case. Furthermore, the Gigaword Corpus is morphosyntactically tagged, using around 700 different tags. Since Icelandic is a highly inflectional language, a lot of syntactic information can be deduced from morphological tagging, which can compensate for the lack of syntactic parsing. Finally, since the Gigaword Corpus has both an Icelandic and an English user interface, it is also useful for researchers who are not fluent in Icelandic and are not familiar with Icelandic linguistic terminology (which is rather special, since we do not use the Latin-based terms most languages do).

>

[23] https://clarin.is/en/resources/gigaword/

You have also done research on the basis of the Icelandic Parsed Historical Corpus (IcePaHC) corpus. Could you summarize your findings from this work?[24] Which linguistic phenomena were you looking at, and what were you able to establish?

<

I have used the corpus in a paper on subjecthood in Old Icelandic, where I argued that word order is a reliable subject test in Old Icelandic, even if the word order is freer than in Modern Icelandic. Similarly, Brynhildur Stefánsdóttir and I have used the IcePaHC corpus to study the incorporation of prepositions, which refers to leftward syntactic movement across lexical verbs, participles, nouns, and adjectives, in the history of Icelandic. An example of this grammatical process, taken from a 15th century text in the corpus, can be seen in the clause *og hefir Oddur **af** virðing **málunum*** ("and Oddur gains respect from the affairs"), where the preposition *af* "from" is displaced from its complement *málunum* "the affairs" via leftward movement so that the noun *virðing* "respect" now intervenes between the preposition and its object complement. Such instances of incorporation have disappeared as a productive process in contemporary Icelandic, which is attested by the data in the IcePaHC corpus. Furthermore, we have argued on the basis of such examples with prepositional incorporation that Old Icelandic was uniformly a Verb-Object language in terms of word order.

>

Could you discuss the IcePaHC corpus itself? How did you use the corpus to extract the relevant syntactic data?

<

IcePaHC is a large diachronic corpus containing Icelandic texts from the 12th century until the early 21st century, so it is a crucial resource for researchers who are interested in historical changes that span many centuries. IcePaHC has excerpts from many different texts from each century, and these texts belong to many different genres – the medieval subpart of IcePaHC is mostly represented by narratives like Old Norse sagas, but biographic, religious, judicial and scientific texts are also included in the corpus. I do, however, wish that there were a syntactically parsed corpus for Old Icelandic with search capabilities similar to those of IcePaHC, which would comprehensively include all the relevant texts from that time period instead of just excerpts.

[24] https://clarin.is/en/resources/icepahc/

Iris Edda Nowenstein, who is currently a PhD student in Icelandic linguistics, performed the searches in IcePaHC for my study on word order and subjecthood in Old Icelandic. I was mostly interested in strings where the finite verb immediately precedes two determiner phrases, and finding such strings is easy to do in IcePaHC because it is syntactically parsed.

>

How does the Icelandic CLARIN research infrastructure benefit your research community?

<

Up to now, digital language resources for Icelandic have been scarce and scattered. It is very convenient to be able to use CLARIN-IS as a hub and access all existing resources from there – both online databases (where available) and downloadable resources in the CLARIN-IS repository. Since new resources are constantly being added to the repository, its value for researchers will increase greatly in the future.

For instance, a large machine-parsed corpus of Modern Icelandic (The Icelandic Contemporary Treebank) has recently been added.[25] I have not yet had the opportunity to use it in my research, but it is potentially highly valuable for syntacticians.

>

References:

Gísli Jónsson, J. 2018. Word order as a subject test in Old Icelandic. In *Non-Canonically Case-Marked Subjects: The Reykjavík-Eyjafjallajökull papers*, 135–154.

Gísli Jónsson, J., and Stefánsdóttir, B. 2014. P-incorporation in the history of Icelandic (abstract). In *16th Diachronic Generative Syntax Conference Research Institute for Linguistics*.

[25] http://hdl.handle.net/20.500.12537/21

FRANCE

Introduction

Written by **Nicolas Larousse** and **Christophe Parisse**

France has been an observer of CLARIN since 2017. CLARIN-FR is coordinated by Huma-Num and its national coordinator is Nicolas Larrousse.[26] It involves several national partners:

- Analyse et Traitement Informatique de la Langue Française (ATILF, “The Analysis and Computer Processing of the French Language”)
- Bases, Corpus, Langage (BCL, “Databases, Corpora, and Language”)
- Cognition, Langues, Langage, Ergonomie (CLLE, “Cognition, Languages, Language, Ergonomics”)
- Centre de Linguistique Inter-langues, de Lexicologie, de Linguistique Anglaise et de Corpus-Atelier de Recherche sur la Parole (CLILLAC-ARP, “Centre for Interlanguage Linguistics, Lexicology, English Linguistics and Corpus-Speech Research”)
- Centre de Recherche en EthnoMusicologie (CREM, “Ethno-Musicology Research Centre”)
- Dynamique du Langage (DDL, “Language Dynamics”)
- FOrmes et REprésentations en Linguistique, Littérature et dans les arts de l’Image et de la Scène (FORELLIS, “Forms and Representations in Linguistics, Literature and in the Fine and Performing Arts”)
- Histoire des Théories Linguistiques (HTL, “History of Linguistic Theories”)
- Interactions, Corpus, Apprentissages, Représentations (ICAR, “Interactions, Corpora, Learning, Representations”)
- Langage, Langues et Cultures d’Afrique (LLACAN, “The Language(s) and Cultures of Africa”)
- LAngues et CIvilisations à Tradition Orale (LACITO, “Languages and Civilisations with Oral Traditions”)
- Langues, Textes, Traitements Informatiques, Cognition (LATTICE, “Languages, Texts, Computer Processing, Cognition”)
- LInguistique et DIdactique des Langues Etrangères et Maternelles (LIDILEM, “Linguistics and Didactics of Foreign and Native Languages”)
- LInguistique, Langues, PArole (LILPA, “Linguistics, Language, and Communication”)
- Laboratoire d’Informatique pour la Mécanique et les Sciences de l’Ingénieur (LIMSI, “Computer Science Laboratory for Mechanics and Engineering Sciences”)
- Laboratoire de LINGuistique de Nantes (LLING, “Nantes Linguistics Laboratory”)
- Laboratoire de Linguistique Formelle (LLF, “Formal Linguistics Laboratory”)
- Laboratoire Ligérien de Linguistique (LLL, “Ligérien Linguistics Laboratory”)
- MOdèles, DYnamiques, COrpus (MODYCO)
- PRAXIs el LINGuistique (PRAXILING, “Praxis and Linguistics”)
- Structures Formelles du Langage (SFL, “Formal Structures of Language”)
- Savoirs, Textes, Langage (STL, “Knowledge, Texts, Language”)

The French consortium is mainly composed of French linguists involved through the CORLI expert group, which promotes the use of good practices for corpus creation, focusing on how to maximize corpus reuse and disseminate the corpus data. CORLI is also active in promoting the application of FAIR principles. Moreover, French NLP researchers are involved via the ATALA association and the group GDR TAL. Additionally, CLARIN-FR has recently started to establish contacts with the French cognitive sciences community, in particular with Institut Carnot pour la Cognition for its “Cognition & Langage” research project.

CLARIN-FR has so far established three C-centres:

- The COCOON Centre provides a data repository with access to oral resources (with a focus on dialectal texts) and an interactive web portal that offers a chain of navigational and analytical tools to the French digital research community. A unique feature of COCOON is that all the oral resources are tagged with precise geolocational metadata so they can be searched geographically on a state-of-the-art interactive map of the world offered on the web portal.

[26] https://www.huma-num.fr

- The ORTOLANG centre provides a general-purpose repository for secure long-term storage of language data mostly pertaining to the languages spoken in France, although resources from other origins are accepted as well, especially when the data come from countries where no public repositories like ORTOLANG are available. For example, COMERE is one of the most visible corpora of ORTOLANG and constitutes computer-mediated language resources such as tweets and text messages.
- The MMSH's Sound Archives Centre (Phonothèque) preserves, archives and disseminates the archived recordings of the sound heritage related to ethnology, languages, history, music and literature from the Mediterranean area. For example, the Phonothèque makes accessible, with ethical and legal rules, recordings of different dialects of Occitan or variants of colloquial Arabic (Syria, Lebanon, Sudan, Algeria, Yemen).

In 2020, CLARIN-FR established the French K-centre for Corpora, Languages and Interaction. The K-centre focuses on providing information, tools, and continuing education to help PhD students and professional linguists work on corpus linguistics. It is run by a panel of corpus linguists who provide their expertise to the community. CLARIN-FR has also successfully added two end points to the Federated Content Search from the ORTOLANG and the COCOON C-centres.

Following the Work Plan for the 2019 renewal of the status as a CLARIN observer and after establishing the K-centre, CLARIN-FR now aims to obtain the Core Trust Seal certification for the ORTOLANG repository so that it can become a CLARIN B-centre. Although the observer period ends in 2021, CLARIN-FR is continuing discussions with the French Ministry of Research, the French National Centre for Scientific Research CNRS and national communities about the opportunity for France to become a full CLARIN member.

Figure 38: Nicolas Larrousse, coordinator of CLARIN-FR.

Tool | **The COCOON Factory**

Written by **Nicolas Larouse** and **Christophe Parisse**

The COCOON platform is aimed at individual researchers and research teams in the Digital Humanities and Social Sciences for the management of their digital oral resources.[27] It combines the functionalities of a data repository, archive and discovery portal.

COCOON contains a large collection of 13,000 recordings and 4,600 transcriptions, which amount to 5,600 hours of speech in 248 different languages from all over the world, as well as historical French language data, which can all be browsed and downloaded. Prominent collections include corpora that were prepared on the basis of dialectological surveys carried out in France and elsewhere, such as the Linguistic and Ethnographic Atlas of Gascony, the Corpus Of Parisian Spoken French From 2000 Onwards, and the Oral Corpus of Afro-Asian Languages. The COCOON corpora can be found through search engines such as CLARIN's Virtual Language Observatory. The transcriptions can also be searched in CLARIN's Federated Content Search.

COCOON also offers a chain of services for navigating and archiving the oral resources. A prominent navigation tool is the geographic search function, which provides a scalable map that shows the distribution of the oral recordings both across the world (Figure 39) and, when zoomed in, within each country (Figure 40). Each recording, which can be directly listened to from the map itself, is also equipped with metadata showing the year the recording took place, information about the collection in which it is contained, as well as biographic information about the speaker (Figure 41). Due to such detailed metadata, COCOON is particularly valuable for typologists and sociolinguists, and as such represents a key tool for the linguistics community in France.

[27] http://cocoon.huma-num.fr/

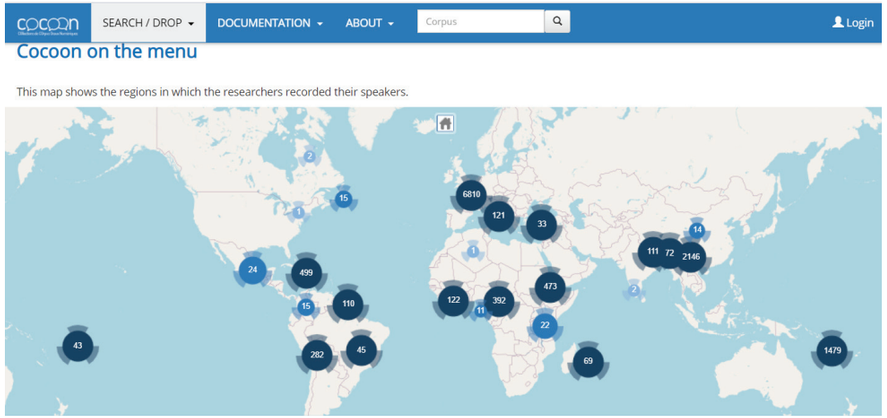


Figure 39: The distribution of the COCOON oral resources across the world. For instance, there are 6,810 oral recordings in Europe.

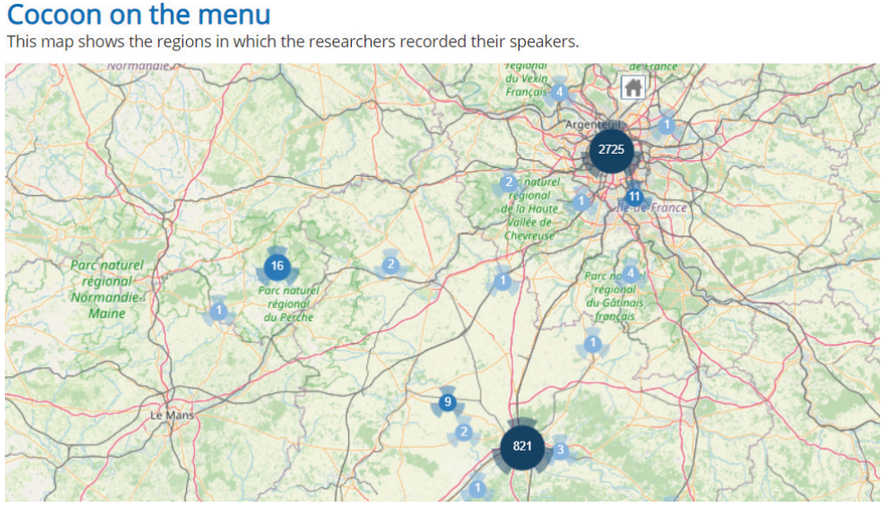


Figure 40: The distribution of the resources in Paris, Orleans, and surrounding areas.

As an archiving tool, COCOON ensures that the researchers' data (mainly audio or video recordings) is automatically normalized in formats supported by the system archiving operator CINES (National Computer Centre for Higher Education). A file in a broadcast format is also produced in a degraded quality to facilitate its use in web mode.

COCOON has been successfully re-used in several research projects. For instance, researchers in the Epic Nepal project have taken existing COCOON resources as a basis for building a corpus of *huḍkelī*, which are sung performances of oral epics in the shamanic tradition of Western Nepal. These recordings can now be accessed online with the Karaoke Tool in COCOON, a user-friendly online interface for listening to the recordings, which are presented along with sentence-aligned transcriptions of the original sung Nepalese and their translations, as well as annotated for the musical instruments. Similarly, the Pangloss collection, which is an online archive of the endangered and under-documented languages of the world, has also been built from COCOON resources and can also be listened to with the Karaoke Tool.

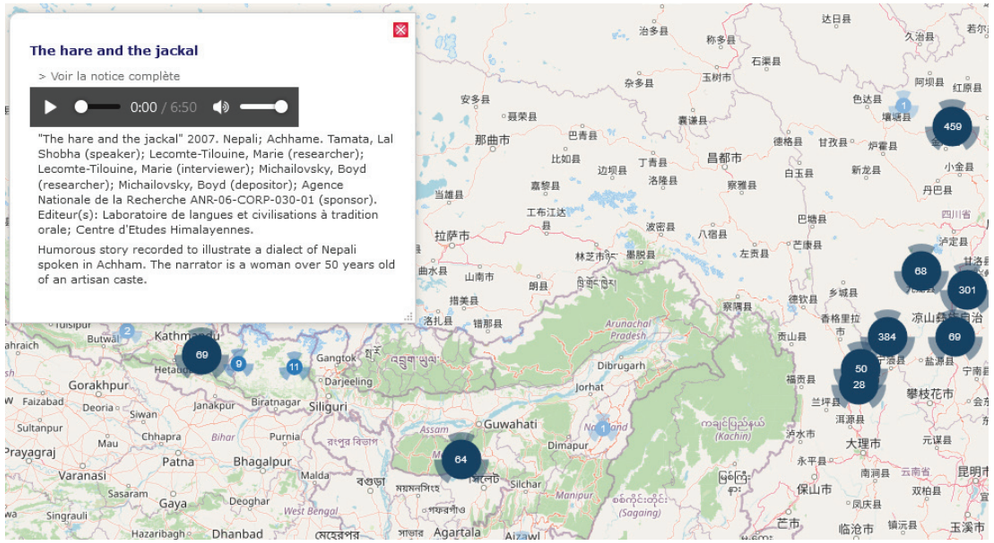


Figure 41: The metadata of a Nepalese recording in COCOON.

Resource | **The CoLaJE Corpus**

Written by **Nicolas Larousse** and **Christophe Parisse**

One of the most prominent CLARIN-FR resources is the CoLaJE corpus,[28] which was developed as part of the Project ANR CoLaJE. The corpus pieces together the emergence and development of communication and language in young children, using an interdisciplinary and multimodal approach. The project provides a crossroads for researchers and artists from all over the world (France, Belgium, Italy, England, United States, Canada, Brazil) and from a variety of fields (linguists, psychologists, speech therapists, filmmakers, musicians, composers) whose paths lead them to child language.

The corpus builds upon a shared database made up of the first-ever longitudinal recordings of children's spontaneous productions from age 1 to age 7 and consists of data from children learning French as well as French Sign Language. The corpus is available for download from the ORTOLANG repository and can be accessed online through a companion website. It contains more than 250 hours of video recorded spontaneous child-parent interaction, and the available transcriptions in the CHAT format (CHILDES) contain more than one million words. All the data is available for research and teaching purposes, and it has already been used for several PhD and master's theses.

The corpus has been used to analyse many features of language development. The simultaneous study of phonology, prosody, gesture, and dialogue has allowed the project team to enrich their perspective on the linguistic development of children. For instance, one working group that was part of the original CoLaJE project team successfully used the corpus to examine the interaction between prosody and gesture in deaf-signing children. Another group focused on the emergence of the discursive, pragmatic, and intersubjective competence in children's communication, while a third group studied the development of grammatical markers, such as the acquisition of the French verbal system, through a comparative study of all children included in the corpus

[28] https://hdl.handle.net/11403/colaje/v2.3

The data in the corpus has also been used by external researchers in various methodologies, emphasizing the multimodality of language development. A large number of publications have been using the CoLaJE corpus, and a special issue of the *Journal of French Language Studies* (volume 22, issue 1) presents some of the main developments in the research produced on the basis of this corpus. Amongst others, Leroy-Collombel and Morgenstern (2012) have used the corpus to prepare a longitudinal study showing how a French-speaking child acquired possessive markers from when she was one and a half years old to three years old, while Sekali (2012) has used the corpus to study the acquisition of French adverbial clauses. A more detailed description of the corpus can be found in Morgenstern and Parisse (2012).

Figure 42: Members of the CoLaJE project.

References:

Leroy-Collombel, M., and Morgenstern, A. 2012. Rising grammatical awareness in a French-speaking child from 18 to 36 months: uses and misuses of possession markers. *Journal of French Language Studies* 22 (1): 57–75.

Morgenstern, A., and Parisse, C. 2012. The Paris Corpus. *Journal of French Language Studies* 22 (1): 7–12.

Sekali, M. 2012. The emergence of complex sentences in a French child's language from 0; 10 to 4; 01: causal adverbial clauses and the concertina effect. *Journal of French Language Studies* 22 (1): 115–141.

Event | **The CoReFo 2018 Study Day**

Written by **Damien Chabanal, Simon Ensor, Anne-Laure Foucher, John Fynn, Véronique Quanquin, Christine Rodrigues Blanchard, Ciara Wigham, Winnie Boingotlo Kaome, Oneil Nathaniel Madden, Albulfatah Al Kishik, Jose Vasques Lopes, Frédéric Mourier, Agnès Pétillat, Aïda Ter-Ghazaryan** and **Siglinde Pape**

Since joining CLARIN as an observer, a number of activities have been organized to help French researchers understand what CLARIN can do for them. On 16 November 2018, the CA2LI research team, which is a member of CLARIN-FR, hosted a study day entitled Corpora at the interface between research and vocational training at the Maison des Sciences de l'Homme, Univerity of Clermont Auvergne.[29] The main goal was to bring together researchers, trainers, and practitioners in the fields of language sciences, education, and healthcare to discuss ways in which multimodal research corpora can be exploited from a pedagogical perspective for vocational training.

The idea for a study day started from the initial observation that although corpus-based and corpus-driven research has a long tradition in France, the use of corpora for the training of practitioners and professionals in the fields of education, language sciences, and healthcare is still not well documented. Indeed, too often the link between research and vocational training is reduced to the simple idea of knowledge transmission and is not considered in terms of appropriation of these research corpora by professionals.

The study day was attended by around 50 participants. Half of them were researchers from the fields of language sciences, education, and healthcare, while the other half constituted students from different master's programmes.

The meeting was structured around a certain number of questions regarding the possible links between research and vocational training and the technicality of the tools. Speakers at the meeting talked about which training objects can be built from the corpora, and how such corpora and linguistic tools available through CLARIN-FR nodes, such as ORTOLANG, can be accessed and used by both trainers and learners. The exchanges during the day have shown, on the one hand, the relevance and transversal nature of these questions and, on the other hand, the effectiveness of the use of corpora in certain areas of vocational training, such as speech therapy and foreign language teaching. However, in certain fields, the needs for training corpora are not always clearly identified. Better knowledge of these needs, stemming from the end-users, would make it possible to guide the collection, processing, and transformation of research corpora into training objects accompanied by guided pedagogical scenarios.

The interdisciplinary nature of the study day made it possible to envisage future collaborations between researchers for the development of training corpora guided by best practices seen and discussed during the presentations. At the same time, master's students who participated became aware of the interactions between research and training, and of the possible contributions of interdisciplinary approaches while acquiring knowledge to complement their current training.

During this event, a presentation of CLARIN infrastructure was given by Michel Jacobson from Huma-Num, the French coordinating institution for CLARIN. This presentation was intended to present the CLARIN infrastructure to participants, but also the organization of the French consortium by Huma-Num and CORLI. An emphasis was placed on the tools and resources made available by CLARIN. For instance, Michel showed how the Virtual Language Observatory can be used to identify resources pertaining to French Sign Language. He showed how the Language Resource Switchboard can be used to explore and analyse the language data in the French corpora in the VLO, which seemed to be a good entry point for the French research community and wholly in line with the theme of the conference. Demonstrations were made using Voyant Tools, UDPipe, and Weblicht on the French VLO resources.

Figure 43: The CoReFo study day.

[29] https://corefo2018.sciencesconf.org/

Interview | **Amalia Todirascu**

Amalia Todirascu is a computational linguist who specializes in NLP. She is a member of the steering committee of CORLI, a group of experts in linguistics, and has successfully used CLARIN language technologies in teaching and research.

Please introduce yourself. Could you describe your academic background and your current academic position?

>

I have been trained in computer science with a specialization in NLP. My thesis was about building ontologies from texts. My work is at the crossroads of NLP and linguistics. For instance, I was part of the ALECTOR project, which aimed to measure the complexity of a text and how to simplify it to facilitate access by children with dyslexia. I am also involved in the adaptation of NLP resources for education, specifically for language learners.

I am now specialized in the construction and preparation of linguistic resources for automatic translation, semantic and discourse annotation, etc. For instance, I was part of the team that built the DEMOCRAT corpus which provides a large manually annotated corpus of coreference, but also develops specialized tools for automatic coreference detection (cofr), for manual coreference annotation (SACR) and for corpus exploration (TXM extensions for handling coreference annotated corpora). In 2004 I joined the LiLPa laboratory, which focuses on linguistics, phonetics, language learning, sociolinguistics and computational linguistics.

>

How did you get to know CLARIN?

>

I discovered CLARIN a long time ago in the context of a collaborative project for collocation detection in German, Romanian and French. My German (Ulrich Heid) and Romanian colleagues (Dan Tufis) were very active in CLARIN at that time (2008/2009), and I was invited to attend a workshop in Berlin organized by CLARIN-D. I was also able to participate in different TEI workshops organized by the German CLARIN consortium.

At that time, France was not officially involved in CLARIN activities, but some people, like Jean-Marie Pierrel (ORTOLANG founder) and Laurent Romary (former director of DARIAH), were already very active in disseminating information about CLARIN among French researchers.

>

What is your role in the French consortium?

>

I am a member of the steering committee of CORLI, a group of experts in linguistics funded by Huma-Num, which is the cornerstone of the French CLARIN-FR consortium.

As a committee member, I was able to be actively involved in the establishment of the recently established CORLI Knowledge Centre. I think the centre is important at the national level because it will provide help and training activities related to the use of French linguistic tools and resources, such as those that are offered by the COCOON and ORTOLANG Centres to different research communities in France. The K-centre also aims to give more visibility to French resources and tools, especially by organizing in working groups aiming to create and moderate research networks that target tools and practices in French linguistics.

As a teacher, I realize that it is necessary to sensitize doctoral students and even master's students to the tools and practices around digital technologies as early as possible. For this purpose, the K-centre is an opportunity to promote digital practices and the general philosophy of CLARIN.

>

What CLARIN tools and resources do you use in your own work?

>

I mostly use CLARIN resources in master's courses in linguistics and language technologies. CLARIN is particularly important in my introduction to my corpus linguistics class, where I use resources (Language Resource Inventory, Virtual Language Observatory) to find corpora, but also tools like WebLicht to show students how they can search for information in a corpus, as well as how to build a simple corpus from scratch. I also use the CLARIN services to illustrate the usefulness of annotations and how to encode corpora in TEI.

WebLicht in particular is a very user-friendly tool for showing what can be done with language tools to students and young researchers who lack experience with NLP or computational linguistics. To give a simple example, the fact that you can annotate simple formats, like DOCX files produced by Word, with most tools that are offered on the CLARIN Switchboard, has proven itself to be crucial from the perspective of accessibility.

Generally, the tools and resources provided by CLARIN are very interesting not only for linguists, but also for other researchers in the Social Sciences. For example, a historical corpus (The Chronicles of Jean Froissart about the first part of the Hundred Years' War) was annotated with person, organization and place named entities. This information was used to retrieve all the occurrences of an entity in the corpus and to study the relations to the organizations, places and persons represented in these texts.

However, it must be taken into account that there is a lack of NLP tools for French in the CLARIN infrastructure; for instance, currently, only UDPipe, so only one out of eight tools on the Switchboard, offers parsing for French. Going forward, I believe it necessary that CLARIN-FR identifies the missing tools, such as keyword extractors, terminology extractors, coreference annotators or NER tools, dedicated to French and integrate them into existing toolchains like WebLicht in order to facilitate their dissemination and use.

>

What do you think needs to be developed to enrich CLARIN and make it better known within the French communities?

>

First, we need to develop more training materials that showcase the use of CLARIN services, especially the language tools and how they can be applied to the resources in repositories like ORTOLANG in the case of France. For this purpose, it would be a good idea to invite external collaborators from the CLARIN network that have already prepared and used such materials.

Additionally, a good idea would be for Huma-Num and the K-centre to organize workshops or webinars about the use of tools for French, but with a focus on practical examples that pertain to applied fields in the Digital Humanities and Social Sciences. For instance, how to choose the right tools for named-entity recognition for a French text in geography. SSH communities would also benefit from CLARIN-FR, as it promotes the adoption of good practices, like the use of standards. For young researchers (but not only them), the French consortium should make better use of mobility grants proposed by CLARIN to allow them to discover other research done at the other consortia, thus strengthening cross-border collaboration.

An important point, which is perhaps unique to the situation in France, is to facilitate exchange among different research communities dealing with linguistics, which are still quite uncoordinated.

>

SOUTH AFRICA

Introduction

Written by **Liané van den Bergh** and **Juan Steyn**

The South African Centre for Digital Language Resources (SADiLaR) forms part of the South African Research Infrastructure Roadmap (SARIR) programme of the South African government's Department of Science and Innovation (DSI).[30] SADiLaR joined CLARIN as an observer in 2018 and is the only digital language research centre outside Europe which forms part of the European CLARIN research network as a C-centre (a metadata providing centre). Current efforts are directed towards SADiLaR becoming a B-centre (a service providing centre) in the near future.

SADiLaR's mandate is to support researchers through facilitating the creation and access to digital data through its Digitization Programme and building overall research capacity through training and dissemination activities as part of its Digital Humanities programme.

[30] https://www.sadilar.org/index.php/en/

SADiLaR is a multi-partner entity located at the North-West University (NWU), which functions as a host and hub of a network of linked nodes, comprising:

- the University of Pretoria (Department of African Languages)
- the University of South Africa (Department of African Languages)
- the Council of Scientific and Industrial Research (HLT Research Group)
- the North West University (Centre for Text Technology)
- the Inter-Institutional Centre for Language Development and Assessment (ICELDA)

SADiLaR runs two programmes to support the South African research community:
Our **digitization programme** entails the systematic creation of relevant digital text, speech, and multi-modal technologies and resources primarily related to the 11 official languages of South Africa. Currently, SADiLaR facilitates access to 378 resources of which 238 are downloadable through a dedicated online repository, which is harvested by the Virtual Language Observatory (VLO). Prominent tools and technologies include the NCHLT text processing web services and the Autshumato machine translation web services, as well as other downloadable applications and software packages for use within the HLT and NLP domains.

Our **Digital Humanities programme** facilitates the building of research capacity by promoting and supporting the use of digital data and innovative (computational) methodological approaches within the Humanities and Social Sciences. This has been done primarily through the commissioning and support of more than 60 workshops, conferences and events within Digital Humanities and Social Sciences since 2017. SADiLaR Workshops have covered a wide variety of topics such as the use of domain applicable computational tools and approaches as well as the general awareness related to what the broad domain of Digital Humanities entails.

These programmes make an impact in three domains:

Language technology domain: As part of the South African Constitution all our official languages are guaranteed parity of esteem and must be treated equitably. However, the reality is that almost all of our languages are under-resourced. For this reason, a key part of what the centre does is creating new high-level resources and natural language processing tools for all South African languages. This is needed to ensure that our language communities and researchers can have access to technologies and datasets that support research and societal equity of access, where language can be a barrier. Practical technologies that are generated and refined using SADiLaR language resources include machine translation engines for local languages, automatic speech recognition systems, text-to-speech systems, speech-to-speech translation systems, interactive

communication systems, as well as a variety of text-related applications such as grammar and spelling checkers, and online electronic dictionaries.

Humanities and Social Sciences domain: Having resources and technologies available is only part of the solution, as it is also necessary for scholars to be able to access and effectively use them. Therefore, SADiLaR actively works toward establishing communities of practice via initiatives for building centralized research capacity. Capacity building ranges from raising awareness about what is available and how researchers can get involved with SADiLaR activities, as well as practical training pertaining to the use of digital data, innovative research methods, and software tools. Through this SADiLaR hopes to enable South African scholars to ask and pursue previously unanswerable questions within their respective disciplines.

Socio-economic domain: Reusable digital language resources are important building blocks that can be used not just for researcher activities, but also by commercial entities to build end-user applications that have a direct impact for language communities. A practical example of this is a recent application called AwezaMed COVID-19, which was developed by a SADiLaR node and aims to remove the communication barriers between health providers and patients. This application was initially developed for maternal healthcare and obstetrics, but was adapted as part of the South African response to the COVID-19 pandemic. The application features speech recognition, machine translation, and text-to-speech developed by the Council for Scientific and Industrial Research in partnership with Aweza. The full report is available online. There is also a YouTube video which shows how the system functions. It is also a good example of how the research outputs are having a translational impact through collaboration with private sector companies and the healthcare sector.

Figure 44: Langa Khumalo, the director of SADiLaR.

Figure 45: SADiLaR members.

Tool | **NCHLT Web Services and CTexTools**

Written by **Martin Puttkammer**

Over the past two decades, research and development projects in South African language technologies (mostly funded by the South African government through their National Centre for Human Language Technologies initiative) have generated several natural language processing (NLP) resources in the form of data, core technologies, applications, and systems, which are immensely valuable for the future development of the official South African languages. Although these tools and resources can be obtained in a timely fashion from the Resource Management Agency of the South African Centre for Digital Language Resources (SADiLaR), their accessibility can still be considered limited, in the sense that technically proficient people or organizations are required to utilize these technologies. This makes it difficult for other potential users, such as Digital Humanities and Social Science researchers, to benefit from them in their work.

To increase the visibility, access, and impact of these technologies, 61 of the existing text-based core technologies, developed by the Centre for Text Technologies (CTexT) over a ten-year period, were ported to Java-based technologies. These were in turn made available to developers via the RESTful API and to end-users through an intuitive web interface (NCHLT Web Services) as well as in a stand-alone interface (CTexTools). The technologies include optical character recognition (OCR) engines, tokenizers, sentence boundary detectors, part-of-speech (PoS) taggers, named-entity recognizers, and phrase chunkers as well as a language identifier for ten of the official South African languages.

NCHLT Web Services

Apart from the RESTful API aimed at developers, SADiLaR has also developed an automated system to assist end-users. This was done through the development of a web-based, user-friendly graphical user interface (see Figure 46) which provides predefined chains of the web services available via the API. For example, if a user needs to perform part-of-speech (PoS) tagging on a document, he or she can upload the document and select PoS tagging and the relevant language. The system will automatically perform tokenization and sentence boundary detection before using the PoS tagging service to tag the user's document.

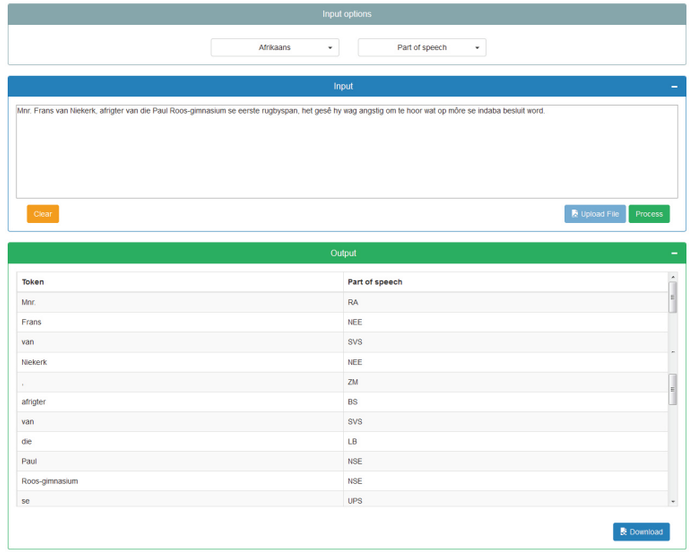


Figure 46: Web interface.

CTexTools

The technologies have also been integrated in a downloadable package of corpus analysis tools (CTexTools; see Figure 47) for users with limited access to the internet. In addition to the above-mentioned technologies, CTexTools is able to compile corpora from a collection of files, to normalize certain inconsistencies, extract frequency and word lists and perform basic collocation searches and keyword comparisons.

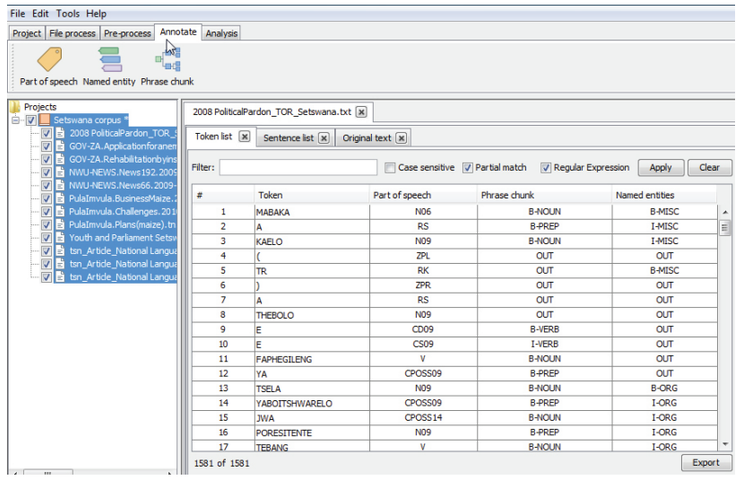


Figure 47: Main window of CTexTools.

The web services and API can be accessed at a dedicated website hosted by SADiLaR,[31] while CTexTools can be downloaded from the SADiLaR repository.[32] More detailed descriptions of the technologies and web services are available in Eiselen and Puttkammer (2014) and Puttkammer et al. (2018).

References:

Eiselen, R., and Puttkammer, M. 2014. Developing Text Resources for Ten South African Languages. In *Proceedings of LREC2014*, 3698–3703.

Puttkammer, M., Eiselen, R., Hocking, J., and Koen, F. 2018. NLP Web Services for Resource-Scarce Languages. In *Proceedings of LREC2018*, 43–49.

[31] https://hlt.nwu.ac.za/

[32] https://hdl.handle.net/20.500.12185/480

Resource | **Setswana Test Suite and Treebank**

Written by **Liané van den Bergh** and **Ansu Berg**

The Setswana Test Suite and Treebank resource was developed as part of the PhD study of Ansu Berg,[33] who performed a rule-based computational syntactic analysis of Setswana with a specific focus on simple sentences in the Setswana language. In order to develop a parser for Setswana, Ansu employed Lexical Functional Grammar (LFG) to frame the description of Setswana grammar. LFG is a non-derivational, constraint-based theory of grammar that distinguishes between two levels of syntactic analysis of a natural language utterance; namely, a constituent and a functional structure. Figure 48 shows the LFG constituent structure for the Setswana simple sentence *mosadi ga a ketla a reka ditlhako* ("The woman will not buy shoes"), while Figure 49 shows its functional structure. The Setswana grammar was implemented in the XLE parser software, which is an environment for parsing and generating grammars corresponding to the LFG formalism with a rich graphical user interface for writing and debugging such grammars. Setswana is the first Bantu language to be parsed in the XLE parser.

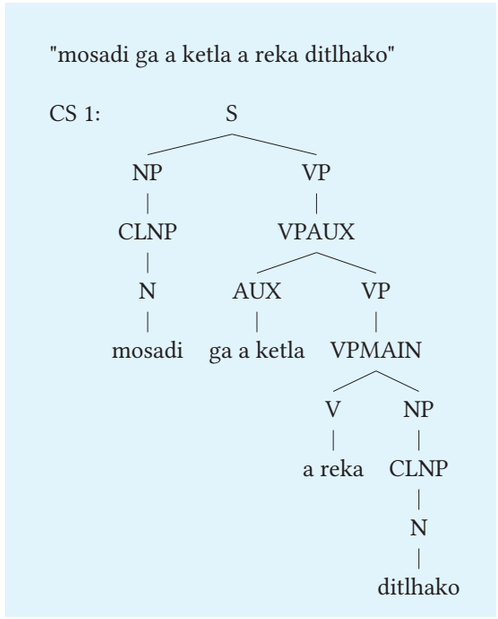


Figure 48: The constituent structure of the simple Setswana sentence *mosadi ga a ketla a reka ditlhako*, where the NP *mosadi* ("the woman") is the syntactic sister of the entire VP node and therefore the grammatical subject, while *ditlhako* ("shoes"), which is embedded deeper in the structure as the complement of the lexical verb, is the grammatical object.

[33] https://hdl.handle.net/20.500.12185/478

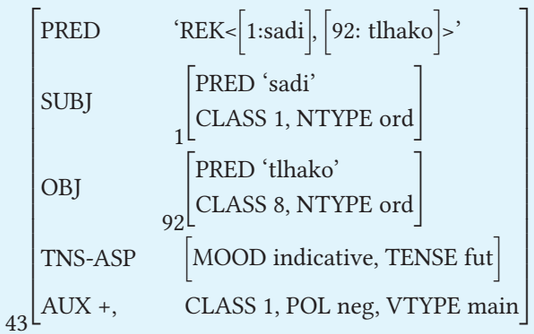


Figure 49: The corresponding functional structure, indicating for instance that the auxiliary system (the syntactic head of the VPAUX node) in Figure 48 expresses a negated future tense in the indicative mood.

A corpus of simple Setswana sentences was not available, so Ansu developed the first test suite for Setswana. The test suite contains a set of constructed linguistic examples, both grammatical and ungrammatical variants, corresponding to the main grammatical categories of Setswana. Ansu used the test suite to develop the first computational grammar for Setswana, and on the basis of this used the functionality provided by the XLE user interface to create the treebank. The resulting treebank is annotated for deep syntactic information corresponding to dependency relations between sentential constituents. For instance, in Figure 48, the simple NP *ditlhako* "shoes" is embedded within the verbal phrase and is the object complement of the main verb *a reka* "buy".

As one of the first syntactically parsed resources for Bantu languages, the Setswana treebank is important for both the computational and non-computational linguistic research community in South Africa. As a resource for computational purposes, the treebank serves as a gold standard for future Setswana grammar testing and evaluation. This contribution to Setswana can also enable similar projects for South African languages that share their syntactic structures with Setswana. For general linguistic research, it is a crucial resource not only for local South African linguists, who can use the syntactically parsed sentences in the treebank for lexicological and general grammatical purposes, but also for syntacticians who can avail themselves of the LFG dependencies in the treebank to research Setswana in the context of Universal Grammar. As an example of the successful use of the treebank in the research domain, Ansu and her colleagues have used the treebank to perform an LFG-style analysis of those auxiliary verbs in Setswana that indicate tense (Pretorius and Berg 2019). Furthermore, they have used the argument structure level of the LFG formalism to determine subcategorization frames (such as the omission of the logical subject during passivization) in the lexical verbal system of Setswana (Berg et al. 2020).

References:

Berg, A., Pretorius, L., and Pretorius R. 2020. Using LFG a-structure to determine the subcategorization frames of Setswana verbs. *Nordic Journal of African Studies* 29 (2): 1–31.

Pretorius, R., and Berg, A. 2019. An LFG Analysis of Setswana Auxiliary Verb Phrases Indicating Tense. In *Proceedings of the LFG'19 Conference*, 233–250.

Event | **SADiLaR and the International Year of Indigenous Languages 2019**

Written by **Liané van den Bergh**

UNESCO declared 2019 the International Year of Indigenous Languages (IYIL2019),[34] which offered an international platform for promoting all indigenous languages, based on five key elements:

1. Increasing understanding, reconciliation and international cooperation;
2. Creation of favourable conditions for knowledge-sharing and dissemination of good practices with regard to indigenous languages;
3. Integration of indigenous languages into a standard setting;
4. Empowerment through capacity building; and
5. Growth and development through elaboration of new knowledge.

The above-mentioned elements fall right within SADiLaR's mandate to serve South Africa as a research infrastructure to ensure that our indigenous languages are not left behind. At the beginning of 2019 SADiLaR launched its IYIL2019 programme, with a main focus on the 11 official languages of South Africa. The centre dedicated each month of the year to a specific language and celebrated it during the entire month through different channels, such as our social media platforms. However, what had a major impact on our user involvement were the language celebrations that were organized for each language at various tertiary institutions all across the country.

The organization started with a general call for collaboration that was sent to various universities, language bodies as well as the National Lexicography Units of South Africa to join the centre as a collaborator and to assist in the organization of the events. Each of SADiLaR's 11 language researchers were project managers for the events and were responsible to host celebrations o at a tertiary institution which supports and offers their language as a field of study. The main idea of the events was to create a platform for academics, researchers, students, and the public to get together and celebrate their mother tongue languages. The following celebrations took place in 2019:

- isiZulu Celebration with 56 participants at the University of KwaZulu-Natal on March 1;
- Sesotho Celebration with 81 participants at the University of the Free State on March 19;
- Afrikaans Celebration with 50 participants at the North-West University (Potchefstroom Campus) on April 17;
- Setswana Celebration with 91 participants at the North-West University (Mafikeng Campus) on July 31
- Tshivenda, Xitsonga and Sesotho sa Leboa Celebration with 350 participants at the University of Limpopo on July 31;
- isiXhosa Celebration with 80 participants at Rhodes University on September 20;
- Siswati Celebration with 149 participants at the University of Mpumalanga on October 24;
- isiNdebele Celebration with 100 participants at the University of Mpumalanga on November 1;
- South African English Celebration, which took place online due to COVID-19.

As Digital Humanities and Natural Language Processing are almost unknown within the South African scholars and academics, the main idea of the events were to illustrate how the South African languages have developed and the new avenues that can be explored within the fields of Digital Humanities and natural language processing. The events also stressed the importance of the latter to ensure that our languages will undergo further development and are thus preserved for future generations.

Each language researcher gave a presentation on SADiLaR as a research infrastructure, with a demonstration of the language resources and tools that are available from the SADiLaR repository and how resources can be downloaded for research purposes. The researchers also demonstrated some of the end-user services/technologies that are available from the SADiLaR website, such as the CTexT NCHLT Web Services and Autshumato MT Web Services. Lastly, the researchers invited the audience to get in contact with SADiLAR via an open call if they have any research proposals or projects that need funding, or if there is a need for a training workshop in some of the tools or resources.

[34] https://en.iyil2019.org/

The programme for each language was unique, with various speakers from different research backgrounds. There was also entertainment in the form of folk dances, poetry readings, singers and much more. For instance, the Setswana celebration, which took place at the North-West University Mafikeng Campus, focused on Setswana in the 21st century. Speakers at the event included the Secretary-General of the South African Mission to UNESCO, Mr. Carlton Mukwevho, who spoke on behalf of UNESCO. He focused on the role the organization is playing towards promoting indigenous languages on an international level. Professor Daniel Matjila, who is the Setswana Language Head at the Department of African Languages at the University of South Africa, gave a presentation on new ways of teaching Setswana literature. Dr. Motheo Koitsiwe, who is Acting Director at the Indigenous Knowledge Systems Centre of the North-West University, then spoke about Indigenous Knowledge Systems and the role they play in teaching and learning. Dr. Baile Mareme, National Lexicography Unit leader for Setswana, followed this with a presentation on the development of dictionaries, with a specific focus on the Setswana language.

All of the events were well received at all the institutions, and they created great visibility and awareness of SADiLaR. The language celebrations created a footprint for the Centre throughout South Africa, and provided a basis for new research opportunities to develop South African languages for future generations.

Figure 50: Speakers at the Setswana language celebration event.

Figure 51: Muzi Matfunjwa (SADiLaR Siswati Researcher at the Siswati Event at the University of Mpumalanga).

Figure 52: Attendees at the isiXhosa language celebration event.

Figure 53: Mmasibidi Setaka (SADiLaR Sesotho Researcher at the Sesotho Event at the University of the Free State).

Interview | **Menno van Zaanen**

Menno van Zaanen is Professor of Digital Humanities at the South African Centre for Digital Language Resources (SADiLaR).

Could you please introduce yourself, your academic background and current work?

<

I am Menno van Zaanen, research manager and Professor in Digital Humanities at the South African Centre for Digital Language Resources (SADiLaR). I received a master's degree in Computer Science (focusing on "low-level" computer science, such as operating systems and computer networks) at the Vrije Universiteit, and a master's degree in Computational Linguistics at the University of Amsterdam (both in Amsterdam, the Netherlands). In these educational programmes, I noticed that there is an overlap of techniques used in both fields. I was wondering if this could be extended, so I focused on the interaction between computational linguistics and computer science. My graduation project for computer science dealt with applying a robust parsing technique from computational linguistics in the area of error correction in the compilers of computer programmes.

For my PhD (University of Leeds, UK), I initially wanted to develop a grammar checking tool by applying error correction techniques from compilers in natural language processing. However, I realized that for that to work, a complete grammar of the natural language would be required. So instead, I started working on language learning tools. I developed a grammatical inference system that aims to learn syntactic grammars for natural language. After that, I realized that such systems can also be used for other sequential information, such as music.

Following that, I have always tried to reuse techniques, such as grammatical inference and machine learning, in different fields. This has led to interesting research and wonderful collaborations in a range of research areas. Currently, I am working together with several of my colleagues to see how far Digital Humanities techniques that are successful, say for English, can also be applied to South African languages. For these languages, limited amounts of resources (datasets and tools) are available, and the quality of the tools is not always very high (this seems to be mostly to do with the domain on which the tools have been trained). We aim to identify approaches that are robust, in the sense that errors in earlier steps in the process do not have a major impact on the final result.

>

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

<

When I was still working in the Netherlands (Tilburg University, Tilburg), I regularly received information on the events and activities organized by CLARIN. Several of my colleagues were active in CLARIN (and later CLARIAH) funded projects, and I was also involved in the OpenSoNaR project. This aimed to provide a user-friendly interface for the Dutch SoNaR corpus, which is a 500 million-word corpus of Dutch and Flemish texts. Having access to such a corpus is wonderful; however, given the size of the corpus, handling such large amounts of data is non-trivial, especially if you also take the annotations that are present in the dataset into account. A resource like OpenSoNaR is therefore essential.

Here in South Africa, I am involved on a different level. Working at SADiLaR means that I now see more of what is needed to be a CLARIN centre. This includes providing detailed documentation of the services and facilitating various aspects of standardization for the resources.

>

How has CLARIN influenced your way of working? How does your research benefit from the CLARIN infrastructure? Which CLARIN resources, tools and services would you recommend to your colleagues?

<

The fact that CLARIN exists has made me realize more that we as researchers do not work in isolation, but are part of a larger network. As such, I especially enjoy the events organized by CLARIN. For instance, the Twin Talks events in which people present how they work together with researchers from different fields provide wonderful ideas on how to tackle the interdisciplinary communication problems.

Inge van de Ven (Tilburg University, Tilburg) and I also presented how we collaborate as researchers with different backgrounds (Culture Studies and Computational Linguistics/Digital Humanities) and for the preparation of this presentation we had to reflect on what problems we ran into and how we solved these. Hopefully, such presentations also help other researchers with their collaborations.

Additionally, events such as the CLARIN Bazaar are nice places to meet other researchers in the field. During the informal discussions at the Bazaar, I have learned about new resources that I did not know existed and I have met people working in the same field that I did not know yet.

Another CLARIN service that I like is the Virtual Language Observatory. This is a tool that allows for searching in a wide range of repositories. Not many South African language resources are available, so this tool is very useful to identify the resources that are out there. Having a central point, like the Virtual Language Observatory, where different resources can be found, makes it easier for me as well as interested researchers to start working with, for instance, different South African languages. As such, I also find it very important that SADiLaR's resources are findable in these services.

At the moment, many Digital Humanities researchers in South Africa are interested in using more computationally oriented tools, but do not know where to start. CLARIN provides training and resources for such researchers, and SADiLaR specifically provides training (currently in the form of workshops, but we are also developing online course material as well) to boost the computational skills of researchers from a South African language perspective. This includes information on digitization, but also on the practical use of the computational linguistic tools and general computational skills.

>

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

<

Recently, I started using OCR systems and named-entity recognizers developed for the South African languages, which are available through the SADiLaR repository. As I do not speak any of the indigenous South African languages personally, I rely on my colleagues who do speak them to interpret the results and evaluate the quality of the output. The discussions based on the output of the OCR systems and named-entity recognizers has made the collaboration between us much more concrete.

For instance, we have experimented with identifying social networks in fiction, mainly in novels and plays, in different South African languages. We first apply the OCR system to scans of the physical books and then use named-entity recognition to identify the characters in the books. The next step is to try to identify the relationships between the characters (based on co-occurrence). These relationships are then visualized in graph form. This approach allows us to get a sense of the robustness of the approach. For instance, we know the named-entity recognizers are not perfect (they miss characters and also tag words that are not names of characters), but we hope that the errors made by the named-entity recognizers do not have a major impact on the identification of the relationships. This can be evaluated by looking at the resulting social networks.

>

Why is the tool and corpus you described important for your research? Which specifics does it possess?

<

In general, there are not many language technologies available for the South African languages. This holds for corpora as well as tools. If you want to do research on one (or more) of the South African languages, you quickly end up using tools that are in SADiLaR's repository. Without the

available tools, much of the research in the field of Digital Humanities for the South African languages simply cannot be done or will require huge amounts of manual annotation. If, for instance, the Afrikaans, Xitsonga, and Tshivenda OCR systems and named-entity recognizers were not available, then the identification of social networks that we have done would be impossible.

>

What are the methodological and technical challenges that you face in your particular field?

<

Unfortunately, for many of the South African languages the tools that are available do not always provide high quality output. I think there are several reasons for this, but the main complicating factor is the limited availability of (annotated) corpora. These linguistic datasets are essential in the development and training of the computational linguistic tools. For some languages and tasks it is seriously difficult to find suitable datasets, limiting the training and evaluation of these tools. With the research that I described earlier pertaining to identifying social networks in South African fiction, I am trying to come up with methods that still yield useful results even if the quality of the output is not always perfect, such as by performing minimal manual corrections. Another approach that several of my colleagues take is to use language independent tools. For instance, they take the text of the constitution of South Africa, which is available in all the 11 official languages, and align the different language versions. Based on the aligned texts, they try to identify the terminology that is used in the English version of the constitution and analyse how the terminology is translated in the other languages to see if similar translation strategies have been used. No language specific tools are needed for this kind of analysis, although linguistic knowledge of the different languages is essential.

>

What would you recommend to students who are interested in the Digital Humanities?

<

I think the field of Digital Humanities in South Africa is still rather fresh. This also means that there is a wide range of opportunities. When more people become active in the field, the field itself will grow in turn, leading again to new opportunities. This is thus the right time to start with research in the field of Digital Humanities, and a perfect time for a research infrastructure such as SADiLAR to be fully embedded in the discipline from the very beginning!

Practically, I think what is needed for students interested in the field is to acquire the right set of skills. For students from the field of humanities, this most likely means learning some computational skills: how to handle and convert files in different formats, how to execute (computational linguistic) tools, and so on. For students who already have a computational background, this means learning more of the open problems, methodologies, terminology, and theories used in the field of humanities. SADiLaR is organizing workshops (currently mostly aimed towards humanities researchers) to help boost their computational skill set. During these workshops, participants learn to use language-specific and language-independent tools. Typically, South African language datasets are used during these workshops, making it easier for participants to understand the applicability. This training program will be extended in the near future with the aim to involve all universities in the country. Unfortunately, during the lockdown no face-to-face training events have taken place. To still allow for training, SADiLaR is currently making course material available online. These training events are excellent starting points for researchers interested in Digital Humanities.

>

What is your vision for CLARIN and the Digital Humanities 10 years from now?

<

I hope that in the next 10 years, the field of Digital Humanities will have grown considerably in South Africa, as well as and in the rest of Africa. There are still so many open questions, for instance related to African culture, African languages, and so on, that need further attention. This is not something researchers can do by themselves. This requires a solid research field with knowledgeable researchers, which is exactly what SADiLaR as the South African CLARIN consortium aims to foster. Once the field is more active, I expect that more datasets will become available, which again will lead to the development of more computational tools, which again allows for more research.

I think that CLARIN can have an important function in boosting Digital Humanities research. I hope that SADiLaR will be able to make the South African resources that are out there more findable and accessible, which is especially crucial given the fact that the Bantu languages are generally under-resourced and under-researched from a Digital Humanities perspective. Training will also help in getting more researchers active in interdisciplinary digital research, and I think that through collaboration, as well as really making use of the worldwide network provided by CLARIN at the ERIC level, this will improve the sharing of information and more importantly experiences.

Knowledge Centres featured in this volume:

IMPACT-CKC Knowledge Centre

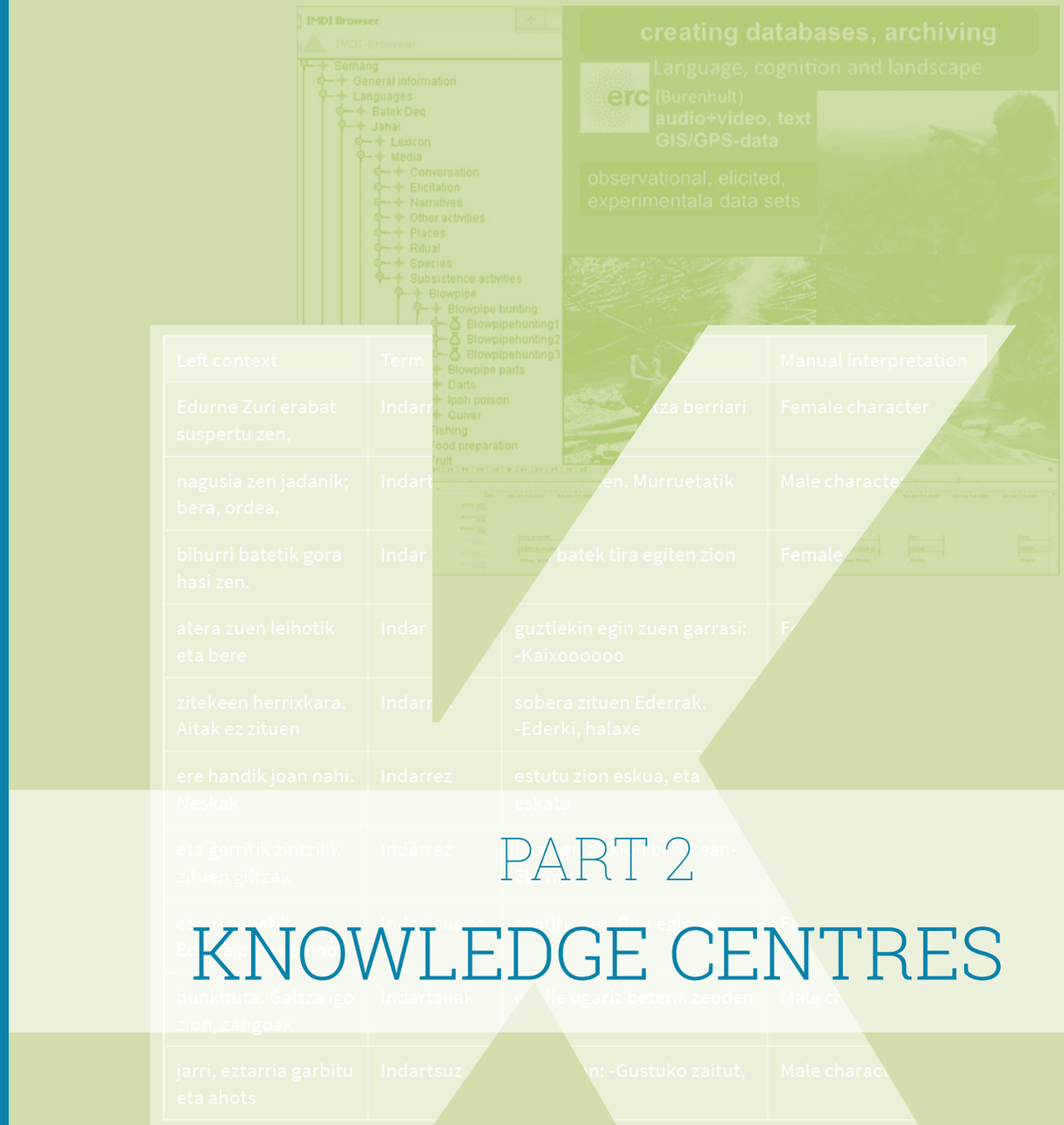
The Knowledge Centre
for Polish Language Technology

The Phonogrammarchiv of the Austrian Academy
of Sciences Knowledge Centre

The Knowledge Centre
for Atypical Communication Expertise

The LUND University Humanities
Lab Knowledge Centre

The Spanish CLARIN
Knowledge Centre



PART 2
KNOWLEDGE CENTRES

IMPACT-CKC Knowledge Centre

Introduction

Written by **Isabel Martínez-Sempere**

The IMPACT Centre of Competence in Digitisation,[35] founded in 2012, is a non-profit organization that aims at making digitization of historical texts better, faster and cheaper. Depending on the language, the period covered by historical texts is different as language change is not homogeneous. With regard to languages, IMPACT is mainly focused on the European ones, but always open to widen the scope worldwide as our members' expertise also includes non-western languages.

IMPACT is based in Spain and hosted by Fundación Biblioteca Virtual Miguel de Cervantes. From an organizational point of view, IMPACT is governed by an Executive Board composed by representatives of its main member institutions. IMPACT is managed by a General Director, Francis Ballesteros (Spain), a Scientific and Technological Director, Tomasz Parkola (Poland), the Executive Board Chair, Frieda Steurs (the Netherlands), and a Manager, Isabel Martínez (Spain).

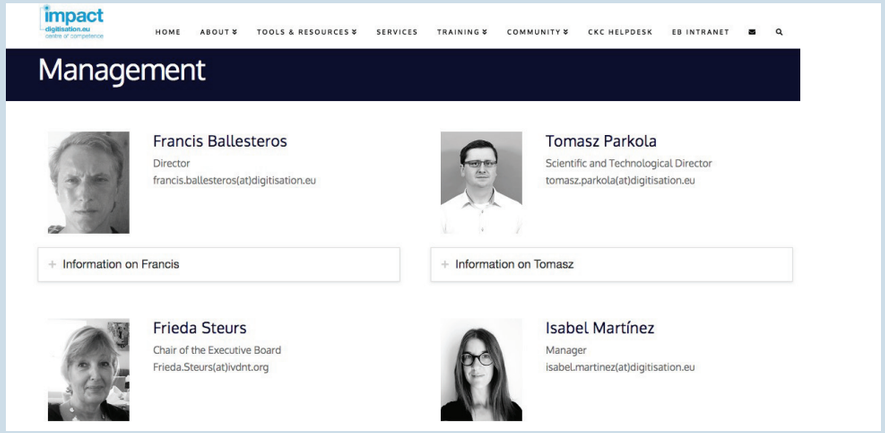


Figure 54: Members of the IMPACT-CKC K-Centre.

[35] https://www.digitisation.eu/

In 2018, IMPACT was recognized as a CLARIN Knowledge Centre with the name IMPACT CLARIN K-Centre in Digitisation (IMPACT-CKC). This recognition is aligned with the centre's objectives, i.e., supporting humanities researchers, cultural heritage professionals and computer scientists in their daily activities. In this context IMPACT offers the following:

IMPACT-CKC Helpdesk

Through its helpdesk, IMPACT provides first-line assistance to researchers on digitization techniques, tools, materials, etc. Researchers are welcome to seek advice on digitization and related fields.

DATeCH International Conference

DATeCH is a unique biennial conference at the intersection between Digital Humanities, cultural heritage and computer science. The first edition of DATeCH was organized in 2014 in Madrid, Spain; the second edition in 2017 in Göttingen, Germany; and the third in 2019 in Brussels, Belgium. During the three editions, DATeCH has received over 140 worldwide submissions, published 72 papers in ACM International Conference Proceedings Series and was attended by over 400 people. Through the organization of the DATeCH International Conference, IMPACT provides researchers with a meeting forum to showcase and discuss their latest research.

Figure 55: The DATeCH international conference.

Online training

IMPACT offers periodic webinars on digitization-related topics of interest for researchers and practitioners such as IIIF, quality control in mass digitization, business models, and so on. Webinars are open to IMPACT members and remaining places are offered to the general public. Similarly, webinar recordings are available to members.

Project proposals corner

The project proposals corner is a space where we publish funding opportunities related to digitization. There is also information on other ways of funding, such as the public-private partnerships. The corner also provides links to EU relevant portals such as CORDIS, Ideal-ist, the international ICT National Contact Point network, and the EC Partner Search service.

IMPACT Language resources

The IMPACT Language resources comprises downloadable historical lexica, corpora and search services for 10 European languages: Bulgarian, Czech, Dutch, English, French, German, Polish, Slovene, Spanish and Latin.

IMPACT Dataset

The IMPACT Dataset is a unique collection of about 500,000 high resolution images, 50,000 of which are accompanied by the corresponding ground truth information provided by direct observation in PAGE XML format, which includes information on layout. This collection is available under different licences, and most resources are free for research purposes. Currently, a new version of the dataset browser is being implemented in order to improve its usability and accessibility.

Training materials

IMPACT provides training materials for specific digitization tools covering the whole digitization workflow, digitization techniques with best practices and briefing papers, and recommendations for digitization projects on licensing and formats and standards. Many researchers produce their own digital materials; consequently, these resources provide a good base to researchers and practitioners interested in heading a digitization project. The tools' training materials are also included in a wiki, DigitWiki, to enable the community to update the content.

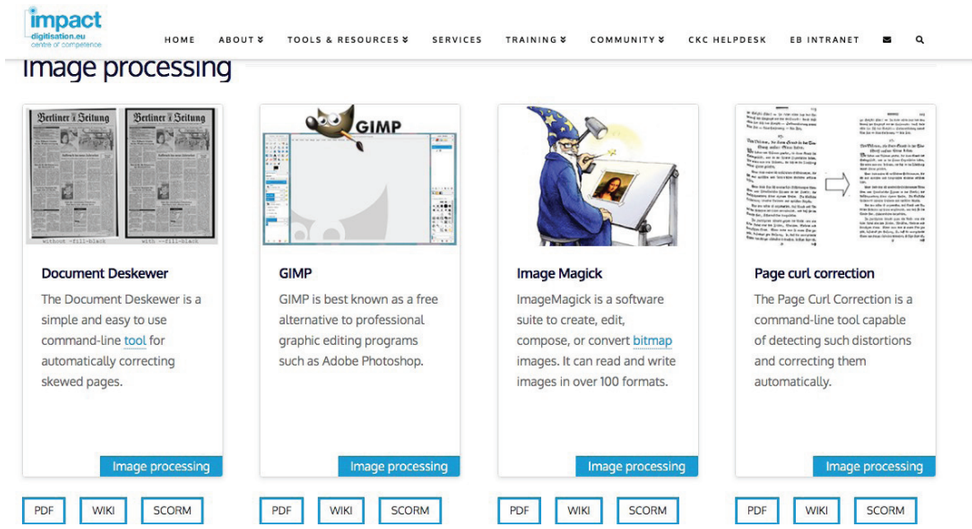


Figure 56: IMPACT training materials.

A map of the digitization landscape

The map of the digitization landscape provides an overview of initiatives and projects in digitization and related fields. The content can be filtered according to the type of initiative, the scope of its activities, and the country. The map contains over 350 initiatives and users can easily add new initiatives by registering on the website.

A tool browser

Similar to the map of the digitization landscape, the IMPACT digitization tools browser provides information on over 250 tools for digitization and related techniques. These tools can be filtered by group and type. Users can also add new tools once logged in by clicking on "add tool" and filling in a simple form (Figure 57).

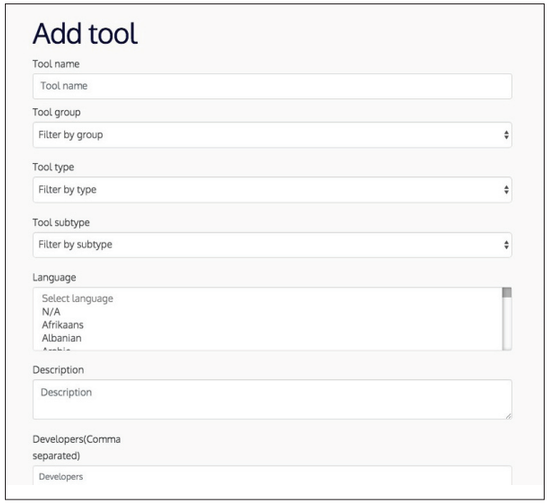


Figure 57: IMPACT tool browser.

IMPACT Demonstrator platform

The IMPACT Demonstrator platform provides researchers with the opportunity to test about 40 digitization related tools in an online environment without the need to install any software.

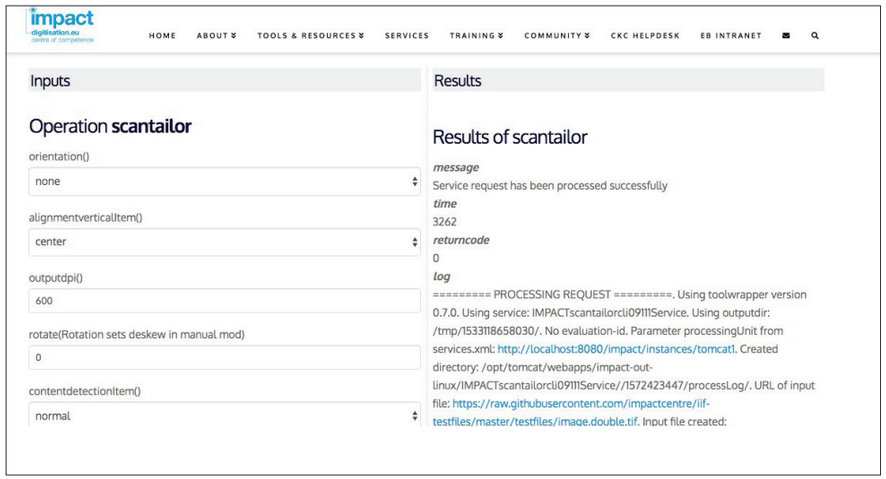
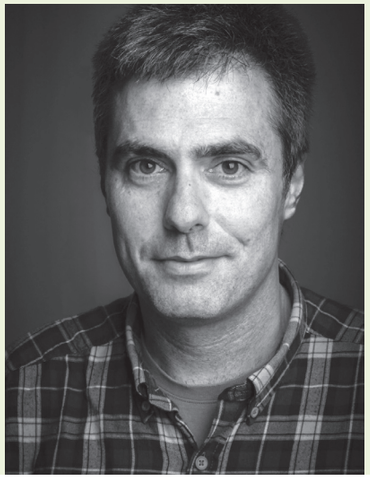


Figure 58: IMPACT demonstrator platform.

Apart from the activities and resources mentioned above, the strength of IMPACT-CKC lies in our wide network of experts covering different scientific fields such as Digital Humanities (Instituut voor de Nederlandse Taal, Ghent Centre for Digital Humanities, Georg August Universität, KU Leuven), Computer Science (Poznan Supercomputing and Networking Centre, Universidad de Alicante) and Cultural Heritage (Koninklijke Bibliotheek, British Library, Bibliothèque nationale de France, Biblioteca virtual Miguel de Cervantes, Berlin State Library).

Interview | **Mikel Iruskieta**



Mikel Iruskieta is a computational linguist who is part of the Ixa Research Group and the Didactics of Language and Literature Department at the University of the Basque country. He has collaborated with the CLARIN IMPACT-CKC Knowledge Centre, which helped him and his colleagues digitize Basque texts.

Could you briefly describe your academic and research background?

<

My current research focuses on the didactics and analysis of Basque, mostly regarding discourse parsing and evaluation of discourse structure. For the last five years, I have mainly worked on adapting language technologies for teaching and learning purposes. With that goal, I have created and now co-lead a postgraduate programme in Basque (University Specialist in ICT and Digital Competences in Education, Continuing Education and Language Teaching), a research group working in Digital Humanities and Education. Our aim is to build a research community that will conduct research and teach in Basque by adopting a critical approach and using language technologies in a pedagogical context. In this postgraduate programme, my colleagues and I are developing a new framework of the socio-tech pedagogy for Basque that will cover the following topics:

- The Basics of Technology and Pedagogy;
- Formal Education and Technology;
- Continuing Education and Technology;
- Language Teaching and Technology Development;
- Society and Education, Opportunities and Risk of Technology;
- E-learning: Approaches and resources; and
- Digital Research: Methods and resources.

>

Does the fact that Basque is a language isolate have any bearing on the development of language tools tailored to it?

<

The history and current situation of the Basque language are both complex and interesting. Basque has a relatively small community of speakers (751,700 active and 1,185,500 passive speakers) which lives in contact with three powerful language communities, namely Spanish and French (as official languages in the Basque Country) and English (as a foreign language). It is also not supported enough by official language policies. As a result, Basque is still considered an under-resourced language. In this context, the work of the Ixa Group for NLP is highly valuable. They have developed basic resources for Basque (as well as for other languages) which are used by the research community, for example IXApipes (a modular set of NLP tools which provide easy access to NLP technology for several languages that can be used or exploit its modularity to pick and change different components) and ANALHITZA (a web service to analyse Basque, Spanish and English texts without needing any technical experience). Many more basic and advanced tools and resources for Basque can be found on the website of the HiTZ: Basque Center for Language Technology.

>

How did you get involved with the IMPACT-CKC K-Centre and how did they help you with your research?

<

I learned about the IMPACT-CKC K-Centre when they joined CLARIN. Because I was working on several different digitization projects for Basque and for Spanish, I immediately got in touch with them and asked for their help. Isabel Martínez Sempere, the manager of IMPACT, helped me solve a digitization issue that I encountered when I was analysing the most frequently occurring words in *Pulgarcito*, which is a Cuban children's magazine written in Spanish from 1919 to 1920. This magazine consists of very diverse materials, such as drawings and handwritten texts, which are

normally very difficult to digitize. I first tried a commercial OCR tool, but the results were very poor. I then got in touch with IMPACT, telling them that I needed good quality OCR results presented in a machine-readable format like XML. IMPACT promptly responded to my request and managed to digitize the entire journal within a week, with significantly fewer errors than when I had used the commercial OCR tool.
In another project, which was led by the Ixa Group but also involved the Basque Ikastola Schools (which are a type of primary and secondary school in which pupils are taught either entirely or predominantly in the Basque language) and Faculty of Informatics, we had three corpora that contained texts for four- to six-year old children. The first corpus is a Basque collection of stories that is used in education.[36] The second is a corpus of old European fairy tales, such as *Rapunzel, The Beauty and the Beast, Sleeping Beauty,* and *Snow White,* which are translated and adapted into Basque. The third corpus is a modern version of the European fairy tales which have been adapted for co-educational purposes, meaning they are suitable for mixed-gender classrooms. However, the co-educational modern version was not machine-readable, so we asked IMPACT if they could give us the OCR version of this collection. IMPACT were again happy to do so, and their experts extracted all the pages from the corpus and performed OCR with ABBYY FineReader (version SDK 11) on the Basque texts.

>

Can you share any interesting results?

<

As soon as IMPACT digitized the fairy tale corpora, my colleagues and I used ANALHITZA, which is a tool for extracting linguistic information from large corpora, to determine whether the texts in the corpora contained gender-inclusive language from the perspective of the characters' roles in the narratives. To this end, we performed an analysis of several expressions, such as *eder* ("beautiful"), *polit* ("beautiful"), *gaizto* ("evil"), and *indarra* ("power"), which we extracted with the Voyant Tools for stylometry from the OCRed corpora.

[36] http://ixa2.si.ehu.es/clarink/corpusak/ipuinak/

In the traditional fairy tale corpus, it turned out that expressions associated with concepts such as beauty and fear (e.g., *eder* "beautiful") were almost exclusively used in reference to female characters, while expressions related to concepts such as power (e.g., *indarra* "strength") were used to refer to male characters. Such a sharp linguistic division between the two genders in learning materials for very young children reinforces problematic gender dichotomies, like the idea that male characters inherently play an "active" and adventurous role in the story, whereas female characters are "passive", dependent characters associated with concepts such as home but not power.

Let's give concrete examples from the two corpora. In the traditional fairy tales corpus (Figure 59), the noun *indar* ("power") and its inflectional variants refer to male characters four out of five times. By contrast, in the modern co-educational corpus (Figure 60), *indar* is now used six out of ten times in reference to female characters, so the usage is almost evenly split between female and male characters, which is desirable if one wants to ensure that the language is used a gender inclusive manner.

Left context	Term	Right context	Manual interpretation
bat ikusi zuen eta, azkeneko	indarrak	ateraz, haraino joan zen. Hondartza	Male character
eta orduan, braust!, Gretelek bere	indar	guztiarekin bultzatu zuen sorgina labe	Female character
eta orduan, braust!, Gretelek bere	indar	guztiarekin bultzatu zuen sorgina labe	Male character
asko nekatu zen. Ez zuen	indarrik	igerian jarritzeko eta itoko zela	Male character
txiki-txiki batzuk ziren. Gulliver	indarka	hasi zen bere burua askatzeko	Male character

Figure 59: Usage of the expressions *indar* ("power") in the traditional fairy tale corpus, where it is associated with male characters in four out of five cases. For instance, the first KWIC line – *bat ikusi zuen eta, azkeneko indarrak ateraz, haraino joan zen* – is roughly translated into English as "He saw one other person, and, drawing his last strength, he went on", describes an action of a male character. By contrast, the second KWIC line – *eta orduan, braust!, Gretelek bere indar guztiarekin bultzatu zuen sorgina labe* – is roughly translated into "And then, Gretel with all of her strength pushed the witch", which this time around describes the action of a female character.

Left context	Term	Right context	Manual interpretation
Edurne Zuri erabat suspertu zen,	Indarrez	bete zen eta bizitza berriari	Female character
nagusia zen jadanik; bera, ordea,	Indartsua	eta arina zen. Murruetatik gora	Male character
bihurri batetik gora hasi zen.	Indar	bitxi batek tira egiten zion	Female character
atera zuen leihotik eta bere	Indar	guztiekin egin zuen garrasi: -Kaixoooooo	Female character
zitekeen herrixkara. Aitak ez zituen	Indarrak	sobera zituen Ederrak. -Ederki, halaxe	Male character
ere handik joan nahi. Neskak	Indarrez	estutu zion eskua, eta eskatu	Female character
eta gerritik zintzilik zituen giltzak	Indarrez	kentzen zizkion bitartean- Eta niri	Female character
etorri zenetik, Ederra piztia baino	Indartsuago	sentitu zen. Bira egin, eta	Female character
hunkituta. Galtza igo zion, zangoak	Indartsuak	eta ile ugariz beterik zeuden	Male character
jarri, eztarria garbitu eta ahots	Indartsuz	esan zuen: -Gustuko zaitut, Monty	Male character

Figure 60: The use of the expression *indar* (“power”) in the modern, co-educational corpus, where it is used six times with female characters and four with male ones. For instance, the first KWIC line *Edurne Zuri erabat suspertu zen, indarrez bete zen eta bizitza berriari* is roughly translated as “Sleeping Beauty was completely revived, full of strength (*indarrez*) and new life”, where the concept of power is associated with Sleeping Beauty, a female character, while the second line – *nagusia zen jadanik; bera, ordea, indartsua eta arina zen. Murruetatik gora* – is roughly translated into “he was already the boss; but he was strong (*indartsua*) and agile”, where strength is associated with a male character.

What is your vision for the future of the IMPACT-CKC K-Centre?

<

Generally speaking, Artificial Intelligence and the work of the IMPACT-CKC K-Centre are crucial to exploring our past. There are many documents which are still not available in a machine-readable form.

Making these information sources accessible, analysing the language in them, linking objects, data and documents, enriching texts with metadata, and making accessible or referenceable the virtual corpus thus created through decentralized collections will enable a new way to interact with our past, understand the present and plan for the future.

Researchers will be able manage a large amount of data and finish their research in less time, allowing them to use more of their time to focus on the truly interesting, ground-breaking research questions rather than on non-innovative technical tasks.

As for my more concrete wishes for the future development of the IMPACT infrastructure, an important topic that would be highly valuable to tackle next is handwritten texts. For us, this would be particularly valuable because we have a large handwritten learner corpus of Basque annotated with errors, as marked by professional language testers who have passed a rigorous ALTE audit. Digitizing handwritten text, however, is notoriously difficult in comparison to printed text. Nevertheless, I think the process can be streamlined with the development of new machine-learning techniques. Luckily, IMPACT already provides an enormous amount of OCRed data which could be used to train new models for digitizing handwritten texts.

The Knowledge Centre for Polish Language Technology

Introduction

Written by **Jan Wieczorek**

The Knowledge Centre for Polish Language Technology (PolLinguaTec) is a CLARIN K-centre that is part of Language Technology Centre CLARIN-PL (LTC) at Wrocław University of Science and Technology.[37] The main aim of PolLinguaTec is to provide knowledge on the application of tools and systems for natural language analysis, especially Polish, within the Digital Humanities and Social Sciences. PolLinguaTec provides extensive documentation, such as instructions, guidelines and tutorials, as well as experienced experts able to solve problems related to the use of language processing tools, such Morpho, Tagger, WSD (for the disambiguation of lexical meanings), Chunker, Parser and Spejd (for shallow processing and morphological disambiguation), and resources, such as the valency lexicon Walenty and plWordNet, that are developed at LTC. PolLinguaTec also helps researchers with the use of language services tailored to interdisciplinary research in the Digital Humanities and Social Sciences, such as the Literary Exploitation System LEM and the WebSty text similarity analysis system (and its multilingual counterpart WebstyML), as well as various tools for extracting information from text, such as TermoPL, and tools for sentiment analysis and topic modelling. Since May 2019, PolLinguaTec has helped plan the implementation of the CLARIN infrastructure in about 10 research projects. The authors of six of the projects have also asked for help in preparing their grant applications.

[37] http://kcentre.clarin-pl.eu/index.php

PolLinguaTec disseminates knowledge about the use of NLP in Digital Humanities and Social Sciences research at conferences and other scientific events. To this end, PoLinguaTec members have organized a number of User Involvement events, such as workshops focusing on the NLP research infrastructure (e.g., ten editions of “CLARIN-PL in Research Practice” workshops), as well as seminars for smaller research teams, such as the seminar CLARIN-PL Tools in Scientific Research in Psychology Seminar).

Figure 61: Participants at one of the “CLARIN-PL in research practice” workshops.

Ever since PolLinguaTec was founded in 2017, it has been involved in the implementation of language services in research projects, many of which have resulted in publications with significant results for the Digital Humanities and Social Sciences. For instance, Ewa Geller collaborated with PolLinguaTec in conducting a diachronic linguistics study of the lexical and semantic effects that arise from long-term language contact by looking at Polish lexical borrowings in Yiddish. The research team needed a tool to create a digital lexical database, which in its conception resembled WordNet. PolLinguaTec developed a specially adapted application, which was a modified version of WordNet Loom – a program used to edit plWordNet. Later it became clear that the new version of WordNet Loom was very useful to the African Wordnet development team, which was looking for a tool to speed up and improve the editing of their database. The results of the cooperation between PolLinguaTec and the African Wordnet team are presented in Griesel, Bosch, and Mojapelo (2019).

In relation to corpus building, Jerzy Malinowski collaborated with PolLinguaTec to construct the Corpus of Henryk Siemiradzki Paintings. In addition, Mariusz Zięba collaborated with PolLinguaTec in his work on "The search for meaning and sense of life and personal growth in the consequence of trauma: prospective studies", which resulted in the paper by Zięba, Wiecheć, Biegańska-Banaś and Mieleszczenko-Kowszewicz (2019). Finally, the PolLinguaTec team has trained researchers to conduct a study of people with post-traumatic stress disorder using NLP methods, where the training and cooperation concerned such topics as the creation of annotated lists of words, sentiment analysis, and the application of stylometry.

References:

Griesel, M., Bosch, S., and Mojapelo, M.L. 2019. Thinking globally, acting locally – progress in the African Wordnet Project. In *Proceedings of the Tenth Global Wordnet Conference*, 191–196.

Zięba, M., Wiecheć, K., Biegańska-Banaś, J., and Mieleszczenko-Kowszewicz, W. 2019. Coexistence of Post-traumatic Growth and Post-traumatic Depreciation in the Aftermath of Trauma: Qualitative and Quantitative Narrative Analysis. *Frontiers in Psychology* 10.

Figure 62: Members of PolLinguaTec and LTC.

Interview | **Dominika Hadro**

Dominika Hadro is Assistant Professor at Wrocław University of Economics and Business. She has fruitfully collaborated with the PolLinguaTec CLARIN Knowledge Centre whose experts have helped her apply topic modelling in her recent research on corporate finance and accounting.

Please describe your academic background and current position.

<

I received my PhD in Economics in 2010 from the Wrocław University of Economics and Business and my work now mainly deals with corporate finance and accounting. In my research, I generally take a qualitative as well as quantitative approach to the analysis of financial reporting. I also study general disclosure and communication practices aimed at stakeholders of publicly listed companies and universities. I am currently an Assistant Professor at the Wrocław University of Economics and Business, but collaborate with a lot with other universities as well. I was a research fellow at Bocconi University and the University of Manchester and visiting researcher at the University of Bologna, University of Navarra and Pablo de Olavide University.

>

How did you get involved with the Polish PolLinguaTec CLARIN Knowledge Centre?

<

Since 2015, I have been the coordinator of the research project Transparency of Listed Companies. At the beginning of the project, I was interested in finding language tools with which my teammates and I could perform a more automatic textual analysis of financial reports in Polish. Because I received my master's degree at the Wrocław University of Technology, I knew that there was a research group working on machine learning. I was checking the profiles of the researchers involved in the group, and I found Maciej Piasecki and the Polish PolLinguaTec CLARIN Knowledge Centre. I contacted him and we immediately started collaborating and using their tools in this and related research projects.

Could you briefly describe the goals and results of this research project?

<

This project is divided into two research streams. In the first, my colleagues and I study impression management (discursive strategies used by company owners to influence stakeholders' impressions), and we apply content analysis to identify which impression management techniques are used in letters to shareholders. The letters are written in Polish by the 60 largest companies listed on the Warsaw Stock Exchange in 2008 and 2013. We identify the patterns in these techniques with the use of k-means clustering, a statistical method that helps us to group the letters into four different categories based on the types of impression management techniques that prevail.

One of the main findings, which are presented in Hadro, Marek Klimczak, and Pauka (2017), is that the more concentrated the ownership of a company, the shorter the letters are, which indicates that the management puts less effort into communicating with investors at these firms. This is particularly visible in companies held by insiders (that is, directors who own more than 10% of a company's voting shares), who tend to produce short, formal letters, devoid of impression management techniques. In contrast, companies controlled by foreign shareholders prepare letters that are longer and are more likely to present defensive arguments, while institutional non-controlling shareholders favour extensive disclosure. The results show that the largest cluster includes letters that praise the management, while the rest include defensive arguments, discussions of negative outcomes and short, formal letters.

The second stream focuses on the use of tone. We examine letters to shareholders as an example of mandatory textual disclosure in which managers have the freedom of choosing the tone used. Our goal is to develop a model that ties their choices to situational incentives (e.g., how attractive a given company is for specific investors), and subsequently test the model with empirical data. The results of our analysis show that managers are on average sincere in their use of tone and that tone is correlated with company performance.

>

How do PolLinguaTec experts help you with regard to the content analysis of the letters and related materials? Which tools, resources or services offered by PolLinguaTec do you use or plan to use with regard to the impression-management analysis? And what are the advantages of the collaboration?

<

Before we started collaborating with PolLinguaTec, we had to manually annotate the textual features required for our content analysis of the impression management techniques. This manual approach required that each text be hand-coded by two individuals. Because we wanted to have as few mistakes as possible, the manual annotation process was extremely time-consuming, since we were annotating for many variables. Now, thanks to PolLinguaTec, we are able to use their automatic stylometric and topic modelling tools such as WebSty, Topic, and LEM, which have not only significantly simplified and streamlined our annotation process, but have also managed to significantly increase the number of texts we can analyse, jumping from around 200 pages of text, which was our limit when we were annotating by hand, to basically an infinite number.

The PolLinguaTec tools are great because they allow us to detect many latent textual features which would go unnoticed in manual annotation, so their tools have proven to be crucial for increasing the validity of our research from a qualitative perspective.

The PolLinguaTec tools can also be easily customized to a very specific task at hand. For instance, LEM allows us to add a readability index to our texts, which helps us ascertain the complexity of the financial statements and degree of voluntary disclosure. Similarly, the tool Topic allows us to perform topic analysis on our corpus almost instantaneously, whereas in our previous manual approach we had to carry out such analysis in several time-consuming steps.

>

Has collaboration with PolLinguaTec helped you advance the state-of-the-art in your discipline?

<

Together with PolLinguaTec, we are currently developing a tool for analysing the tone/sentiment of financial texts in Polish, so in this sense our current collaboration is highly relevant for our work in the project on impression management techniques. This sentiment analysis tool is based on a massively overhauled version of the Loughran-McDonald wordlist, which is an English sentiment

wordlist that is widely recognized as the best resource to measure tone in finance and accounting. We have identified three major areas for improvement of the Loughran-McDonald method.

First, the method is weakly supported from a theoretical perspective – although it has proven itself to be empirically effective, it was first and foremost developed with the pragmatic goal in mind to obtain the best results in empirical research, so theoretical considerations were largely disregarded. Researchers have very often applied these standard empirical tools, but they cannot tackle key questions about the nature of the textual characteristics that they want to measure. Second, the lack of a theory limits the Loughran-McDonald method to one language. With our tool, we want to remedy this linguistic limitation by combining the Loughran-McDonald list with the Princeton WordNet, which provides data that is significantly more conducive to multilingual applicability. Third, the Loughran-McDonald wordlist relies on the bag-of-words approach, which has proven itself to be severely limited in the case of sentiment analysis in finance. This is because wordlists inherently provide noisy measures, so we have to disregard a lot of data if we want to achieve good results in our field.

Furthermore, it is now recognized that sentiment analysis should be more driven by dynamic machine-learning methods and approaches rooted in artificial intelligence rather than on the basis of sentiment lexica, which are in essence static wordlists. This sort of AI-driven approach is exactly what we aim to achieve with the tool that we are developing together with PolLinguaTec – one of the main advantages in this sense will be that the tool is going to take the context and discursive features of the text into consideration in order to improve accuracy, which is in line with the current research trends in computational linguistics.

>

Are there any other on-going projects that also involve collaboration with PolLinguaTec that you would like to highlight?

<

Yes, my colleagues and I are also collaborating with the K-centre on three other projects. In the first one, we are performing textual analysis of the non-financial information of publicly listed companies from an ethical perspective. We are using PolLinguaTec's topic modelling tools to analyse the companies' annual reports to determine if and to what extent the companies engage in the Corporate Social Responsibility (CSR) model. With topic modelling, we try to unveil textual patterns that link the company's CSR strategies to broader ethical issues and trends.

In the second project we also use PolLinguaTec's topic modelling tools, but this time to perform textual analysis of sustainability reporting in public universities. We are trying to determine the most common topics that university admins present in their sustainability reports, as well as what influences the structure of the reports.

The third and most recent project, which is currently in its initial stages, has to do with textual analysis of the differences between annual financial reports before and after the change from cash to accrual accounting at public universities. Specifically, we want to determine how the universities' communication with stakeholders has changed as a result of this switch to accrual accounting. To this end, we plan to use PolLinguaTec's tools to look for textual characteristics like the differences in readability and types of topics between the Italian universities. Later on, we want to extend our scope beyond Italy and look at the situation in other countries as well, and possibly also look at institutions other than universities.

>

What is in store for your future collaboration with PolLinguaTec?

<

After the development of the sentiment analysis tool is completed, we would like to extend its functionality so that it could be used to identify other discursive strategies in finance. Together with Walter Aerts from the Antwerp Management School, my team and I are mainly interested in having a tool that can be used to automatically analyse attributional framing, which refers to the framing techniques with which managers explain the performance of a company in a financial report. Research on this topic generally shows that good performance is explained as the result of good management, while bad performance is often attributed to external factors. However, attributional framing can vary depending on company culture and the individual incentives of a particular manager, such as the manager's salary and whether he or she is also the owner of the company. To the best of my knowledge, computational tools that can code attributional framing do not exist yet, so I hope that we can fill in this technological gap together with PolLinguaTec.

Reference:

Hadro, D., Marek Klimczak, K., and Pauka, M. 2017. Impression Management in Letters to Shareholders: Evidence from Poland. *Accounting in Europe* 14 (3): 305–330.

The Phonogrammarchiv of the Austrian Academy of Sciences Knowledge Centre

Introduction

Written by **Kerstin Klenke, Christiane Fennesz-Juhasz, Katharina Thenius-Wilscher, Christian Liebl, Nadja Wallaszkovits** and **Johannes Spitzbart**

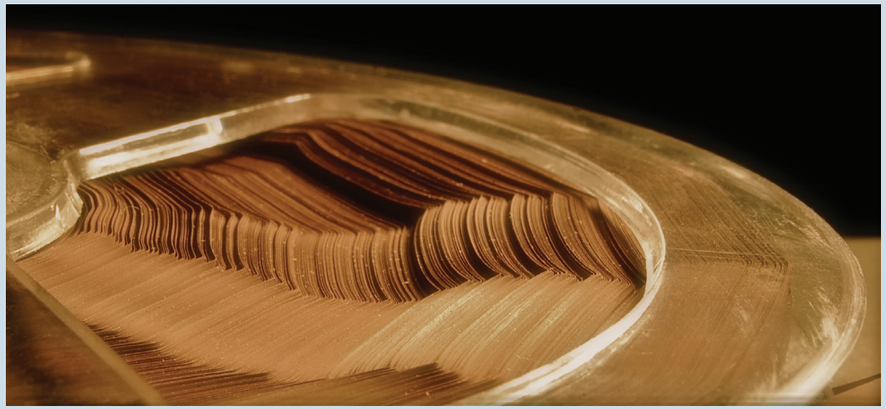


Figure 63: Magnetic tape before regaining playability [photo: Bernhard Graf].

The Phonogrammarchiv of the Austrian Academy of Sciences is a research archive which houses large collections of sound and video recordings from all across the world.[38] Founded in 1899 at the then Imperial Academy of Sciences, the Phonogrammarchiv was set up as a multi-disciplinary institution from its very beginning. The Phonogrammarchiv became a CLARIN Knowledge Centre in 2015, and is part of the Austrian CLARIN group within the CLARIAH-AT Consortium. It has built up – and continues to enlarge – its collections through:

- Technical and methodological support of researchers whose recordings are subsequently archived and catalogued;
- Research projects conducted by the archive's own staff;
- The acquisition of already existing research collections (e.g., scholarly estates).

Research by Austrian scholars and by scholars working in Austria is at the focus of the Phonogrammarchiv's collections. The majority of the recordings originate from studies in social and cultural anthropology, ethnomusicology and linguistics, but disciplines such as medicine and zoology, among others, are also represented. The activities of the Phonogrammarchiv centre upon the following goals:

- Long-term preservation of its collections (audio and video recordings plus accompanying materials), their contextualization and annotation as well as their accessibility (online/onsite and through publications) for scholars, communities of origin/communities of practice as well as the general public;
- Advancement of audio and audio-visual research through developments in technology (digitization/re-recording, restoration, long-term storage strategies) as well as methodology (the Social Sciences and Humanities);
- Critical reflection of the premises and predicaments of audio and video archiving – past and present.

As a CLARIN Knowledge Centre, the Phonogrammarchiv offers various services. It disseminates knowledge by providing access to data and metadata resources, through individual advice, group trainings and workshops, internships, practical assistance and institutional cooperation.

[38] https://www.oeaw.ac.at/phonogrammarchiv/

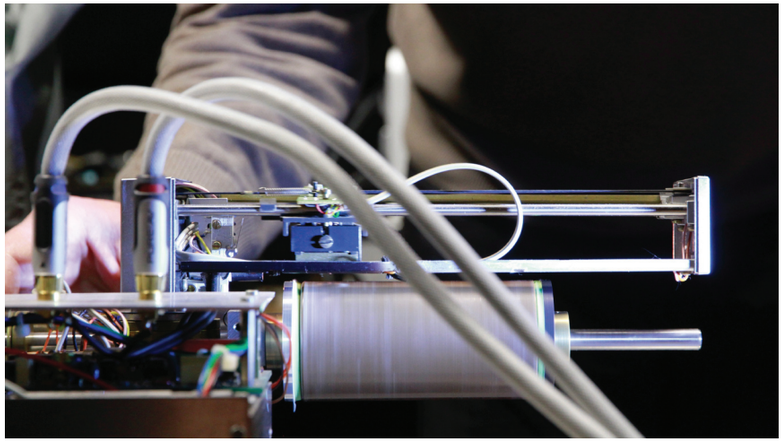


Figure 64: Digitizing historic sound and video carriers is one of the Phonogrammarchiv's areas of expertise [photo: Clemens Gütl].

The Phonogrammarchiv offers the following access points for its services:

- Online: A continually expanding online catalogue offers two modes of access (full text and expert search) to about 85% of the collections, which encompass 71,000 items or 12,500 hours of audio recordings and 5,900 items or 1,800 hours of video recordings. Catalogue entries include information about the year and place of recording, the person who made the recording, archive number, content, duration, language of the recording, ethnic/national/cultural group that the recorded people belong to, collection name, musical instruments, etc. The respective sound files (complete or in part) can be accessed through about 1,000 catalogue entries.
- Onsite and remote access: Due to legal and ethical restrictions, only an excerpt of the Phonogrammarchiv's comprehensive database can be displayed in the form of the online catalogue; for the same reasons, access to the actual recordings and accompanying materials is provided onsite on the Phonogrammarchiv's premises, or on request. Remote access to full-length recordings and/or documentation and accompanying materials requires usage authorization by the archive.
- Usage Authorization: The Phonogrammarchiv provides users with copies of recordings and other materials on the basis of written agreements – if legal and ethical aspects as well as the envisaged usage allow. Usage authorization is mostly granted to scholars and communities of origin/communities of practice, but also to museums and media, or for educational and artistic purposes.

The Phonogrammarchiv's staff advise scholars – individually or in group settings – on the technical aspects of sound and audio-visual recording as well as in research methodology in the Social Sciences and Humanities, including broad knowledge with regard to tailoring recording techniques to specific research settings. Scholars are supported with audio and audio-visual equipment for research purposes. Research advice is also given with regard to using the collections for academic or other purposes. Due to their wide expertise in all matters of audio and video archiving, the staff of Phonogrammarchiv offer support to individuals and institutions on a broad range of technical topics, starting from physical restoration, digitization and re-recording, through format obsolescence and migration to long-term preservation and storage. The Phonogrammarchiv engages in cooperation with archives abroad in the field of digitization, metadata structures and cataloguing and provides support concerning a broad variety of archival topics.

Figure 65: Monitoring the video digitization process [photo: Bernhard Graf]

As a CLARIN Knowledge Centre, the Phonogrammarchiv can build on its decade-long experience in the field of training and disseminates its expertise in various formats. In 2018, for example, the Phonogrammarchiv participated in the international training event "10 years of preservation of our memory for the future" at the National Phonoteque in Mexico City, which was attended by around 200 participants from Mexico, other Latin American countries, as well as experts from Europe and Australia. At this event, the Phonogrammarchiv was represented with lectures and hands-on workshops, as well as a training program especially tailored to the demands of the National Phonoteque. For instance, Nadja Wallaszkovits, who is a musicologist and audio engineer at the Phonogrammarchiv, offered workshop units on the topics of Sound Heritage Restoration and the Preservation of Sound Heritage.

Another training format offered by the Phonogrammarchiv is internships. In 2018, the Phonogrammarchiv offered a three-month internship to a student of HTW Berlin – University of Applied Sciences with training in the following fields: recording and documenting the condition of records for conservation and long-term preservation; cleaning, parameter optimization and digitization of historical sound carriers; preparation of the digitized material for long-term archiving and use, including technical documentation and database entries.

Figure 66: Participants from Taiwan at an international training course at the Phonogrammarchiv [photo: Nadja Wallaszovits].

The Phonogrammarchiv also cooperates successfully and internationally with a great variety of institutions, among them the following:

- Ongoing since 2015: AfricAvenir International in Douala, Cameroon: digitization of audio tapes and mastering of CDs.
- 2015–2017: The Institute of Ethnomusicology – Centre for Music and Dance Studies in the Faculty of Social Sciences and Humanities at the NOVA University Lisbon in Portugal: digitization of magnetic tapes and videos.
- Ongoing since 2017: the Salzburg State Institute for Folklore, Austria: digitization of magnetic tapes and audio cassettes.
- Planned: the Uzbek Academy of Sciences in the fields of restoration, digitization and cataloguing.

In February 2020, the Phonogrammarchiv hosted Associate Professor Dr. Silvia Calamai (University of Siena, Italy) as a CLARIN Mobility Grant Scholar.

Interview | **Beate Eder-Jordan**

Beate Eder Jordan is a literary scholar focusing on minority and Roma literature at the University of Innsbruck in Austria. She has successfully collaborated with the Phonogrammarchiv Knowledge Centre in the RomArchive (2015–2019) project.
[Photo by Monika Raič.]

Please describe your academic background and current position.

I am University Assistant at the Department of Comparative Literature at the University of Innsbruck in Austria and have been involved in research on topics related to minority literatures and cultures for many years.

>

In your research, you focus on Roma and minority literature, especially from an intercultural and transcultural perspective. Could you briefly present some of your current work on this topic?

<

When I analyse literature and art in my research and teaching, I always focus on its political dimension. I am interested in questions of production and reception, identity-building, in the process of minoritization and in cultural memory. I also very much appreciate interdisciplinary work. I am a member of the DFG (German Research Foundation) network Aesthetics of Roma: Literature, Comic and Film by Roma in Areas that Speak a Romance Language, where I collaborate with researchers from other disciplines focusing on different aesthetics of Romani literature and art. In 2018, I organized an interdisciplinary lecture series at the University of Innsbruck called Cultural Encounters and Conflicts: Minoritization. Representation and Alliances, and this year I organized the lecture series Meeting of Knowledges and the Sociopolitical Relevance of Research and Cultural Work.

In order to increase the visibility of Romani authors, it is necessary that their literary works are printed and translated. For this reason, in collaboration with innsbruck university press, I helped republish the German-language edition of József Holdosi's novel *Kányák*, an important Hungarian generational novel influenced by magic realism. Moreover, together with Erika Thurner and Elisabeth Hussl, I edited the volume *Roma und Travellers. Identitäten im Wandel* ("Roma and Travellers: Changing Identities"), which was also published by innsbruck university press in 2015. In the same year, I organized the Writer in Residence program at the University of Innsbruck, where we hosted Jovan Nikolić, a prolific Yugoslavian-born author from the Romani community, who now lives in Germany.

>

Can you present the aims of the project RomArchive (2015–2019)? What kind of materials do you collect there? How is the project connected to the Phonogrammarchiv?

<

RomArchive's primary aim is to increase the visibility of the arts and cultures of Roma, and thereby emphasize their contribution to European cultural history and at the same time counter persistent stereotypes.[39] This international project addresses Roma as Europe's largest minority as well as their relationship to the majority societies. It has resulted in an openly accessible and curated digital archive, which brings together Romani art of all kinds, as well as scholarly texts and historical documents.
The project is characterized by Romani leadership: members of Romani communities have been involved as curators, artists, scholars, and members of the advisory board.

RomArchive consists of ten sections: dance, film, literature, music, theatre and drama, visual arts, flamenco, material on the politics of photography, first-person testimonies related to the persecution of Roma under the Nazi regime, and scholarly material on the civil rights movement. RomArchive was awarded the prestigious Grand Prix of the European Heritage Awards / Europa Nostra Awards 2019.

[39] https://www.romarchive.eu/en/

I curated one of the ten sections of RomArchive, namely literature, which presents a wide range of oral and written literary works of Roma, Yenish and Travellers in a variety of languages, Romani dialects and genres. I collaborated with an international team of 30 scholars from Europe, Russia and the United States, who are experts on Romani literature in their countries. The archive mainly contains examples from European countries as well as some examples from overseas.
The linguist Petra Cech was the curator responsible for the orality subsection. Together with the team of experts, we collected various texts, such as poems, extracts of prose, examples of children's literature, scholarly articles, and photos, as well as sound recordings of Romani oral literature, interviews with authors, lectures, and readings.

This extensive project was only possible thanks to the cooperation with the Phonogrammarchiv of the Austrian Academy of Sciences, which is a CLARIN Knowledge Centre and hosts the largest collection of field recordings of Romani language, oral tradition and music, including the Heinschink Collection, as well as the Collections Hübschmannová and Davidová. Covering the time span from the mid-1950s to date and consisting of well over 1,000 hours' worth of materials, these holdings include thousands of (predominantly audio) recordings from various European countries and Turkey. For the literature section of RomArchive, the Phonogrammarchiv's curator responsible for collections of Romani culture, together with Petra, selected, edited and annotated 18 sound recordings from the Phonogrammarchiv's field collection that feature Romani oral literature from various countries recorded between 1965 and 2012. In addition to this, the Phonogrammarchiv edited literary readings from its holdings as well as readings and interviews with famous authors, mainly recorded by one of our team members in the course of the project.

>

How does the Phonogrammarchiv help with the preservation and annotation of the audio-visual and written materials in this project? What are the main obstacles related to working with audio-visual materials? And how does the Phonogrammarchiv overcome them?

<

The collaboration of the Phonogrammarchiv, as a K-centre, with researchers like us on how to handle materials on the cultural heritage of minority communities is an excellent example of putting the idea of CLARIN's research infrastructure into practice.

At the outset, the RomArchive project wanted to achieve a high standard in the digitization, editing and cataloguing of its materials. To this end, the literature section teamed up with the Phonogrammarchiv, which fulfilled a whole range of tasks for our section. These included technical support and advice in the editing of the sound recordings of the interviews, the digitization of printed materials and analogue recordings, and the production of digital copies for online presentations. In addition, the metadata capture of all the materials collected by the literature section was carried out together with a specialist in information studies, who had gained experience with audio-visual materials through an internship at the Phonogrammarchiv while she was a student. The Phonogrammarchiv also provided a trusted repository for the digital transfer of all texts, sound files and other materials during the review process by board members of RomArchive. Finally, the Phonogrammarchiv successfully conducted rights clearance with authors and speakers (or their heirs) and other rights holders for the online publication of their works and performances in the RomArchive.

As stated above, I collaborated with many international colleagues the RomArchive project. However, the editing, capturing and digital archiving of the materials go beyond the know-how of many scholars in literary studies. It was therefore essential for me to have gained the Phonogrammarchiv as a partner in the project, as its expertise and technical support were necessary in all steps of the digitization process. Aside from RomArchive, my colleagues and I have also benefitted from the Phonogrammarchiv's support and advanced technologies in other research and teaching endeavours, for instance when the archive's staff performed the digital transfer from obsolete audio-visual formats that were used some decades ago.

>

One of the aims of the Phonogrammarchiv is to help the communities that provide the audio-visual recordings and the accompanying written materials. How do the Roma community, as well as immigrant and minority communities in general, benefit from such help and preservation on the part of the Phonogrammarchiv? What future developments would you like to see in relation to this?

<

The archived field recordings of tales, stories and songs are part of the cultural heritage of the people recorded, and their oral histories or other individual accounts are first-hand sources for (future) historical narratives of these communities. Their preservation and documentation in an institution such as the Phonogrammarchiv is important because it guarantees the steady availability of such resources for the communities themselves.

Aside from providing a permanent archive, the Phonogrammarchiv has been cooperating with Romani and other minority communities for a long time. Since the early 1990s, the Phonogrammarchiv collaborated with the Romani Project, which is led by sociolinguists at the University of Graz who have been working on the codification of several Romani dialects, including highly endangered varieties. Here, the Phonogrammarchiv has provided numerous recordings for analysis and publication. As a significant outcome of this collaboration, five bilingual volumes with Romani tales and other narratives were published between 2000 and 2012. The tales and narratives included in the volumes are largely based on recordings from the Phonogrammarchiv's collections. Three of these anthologies were published together with CD editions containing the original audio recordings. Since such publications are bilingual, in the sense that they contain the original Romani texts and their German translations, they also address the wider public in addition to the Romani communities. The same is true for the educational website RomBase – Didactically Edited Information on Roma, for which the Phonogrammarchiv contributed sound samples illustrating various topics on Romani history and culture. In 2006–2008, the Phonogrammarchiv cooperated with representatives of a Romani organization in the Austrian province of Burgenland during their field research in the oral history project Mri Historija.

One recent project involving the Romani community in Austria was the exhibition Romane Thana ("Places of Roma"), which was shown at the Wien Museum in 2015 as well as at other Austrian museums. It was organized by the NGO's Romano Centro (Vienna), Initiative Minderheiten and the Wien Museum. Many members of Romani communities were directly involved and made their own contributions at this extensive exhibition, which was visited by nearly 23,000 people in Vienna. The Phonogrammarchiv contributed a video (documenting the work of a Romani coppersmith)

and audio samples of songs, tales and interview excerpts from the Heinschink Collection, so visitors could learn about the variety of Romani dialects. The website of Romane Thana also contains some of these items in its virtual exhibition, as well as educational material for students. The Phonogrammarchiv has provided relevant audio materials about Romani culture for school teachers, general educational purposes and for radio programmes focusing on minority issues.

In short, such cooperation is crucial both for the Phonogrammarchiv and the focal communities themselves. Staff members of the Phonogrammarchiv have gained an outstanding knowledge of the Romani language and cultures, and the archive's multifaceted expertise is, in my experience, very much appreciated by the minority communities. Transcultural cooperation is made possible, such as in the field of education, where there are now more expert lectures on topics about Romani and other minority communities at schools and universities.

>

How do the Phonogrammarchiv's experts advise scholars on implementing audio-visual technologies in Humanities and Social Science research? Could you give an example from your research on Roma and minority literature that shows how you have benefited from Phonogrammarchiv's advice?

<

Generally speaking, the Phonogrammarchiv supports Austrian researchers by providing state-of-the-art equipment and training on the use of audio-visual technologies and methodologies in field recording, as well as knowledge on the documentation and processing of the recorded materials. It is also important that the field recordings are archived at the Phonogrammarchiv, since they thereby become accessible for future research and users such as myself. The Phonogrammarchiv as a K-centre serves the CLARIN community with knowledge transfer and individual and group training on various aspects of audio-visual recording and long-term preservation.

In my case, the Phonogrammarchiv's advice and assistance were especially crucial for creating many parts of the RomArchive literature section. In addition to chapter 4 on Oral Literature, the Phonogrammarchiv's experts edited all audio material presented in two other chapters in this section – in chapter 5, which includes interviews, and in chapter 7, which includes literary readings and recitations by well-known Romani poets and writers. Several members of my team, as well as other colleagues at the RomArchive, have also given very positive feedback on the intensive assistance and support that they have obtained from the Phonogrammarchiv.

>

What is your vision for the Phonogrammarchiv 10 years from now?

<

I would like to see that the Phonogrammarchiv continues to maintain its extraordinarily high standards in supporting Romani and minority communities, as well as scholars and partner institutions at an international level. I hope that new research communities will benefit from the Phonogrammarchiv's expertise on audio-visual research and from their thorough knowledge, which aside from minority studies extends to other relevant disciplines such as ethnomusicology, linguistics and social anthropology. I also hope that new generations of minority communities and scholars will have the opportunity to cooperate with the Phonogrammarchiv in the same way as I have been able to for many years.

>

The Knowledge Centre for Atypical Communication Expertise

Introduction

Written by **Henk van den Heuvel**

The K-Centre for Atypical Communication Expertise (ACE for short) is run by the Centre for Language and Speech Technology (CLST) at Radboud University.[40] The mission of ACE is to support researchers engaged in investigating what can be characterized as atypical communication, which is an umbrella term used here to denote language use by second language learners, people with language disorders or those suffering from language disabilities, but also to languages that pose particularly difficult issues for analysis, such as sign languages and languages spoken in a multilingual context. It involves multiple modalities, such as text, speech, sign, gesture, and encompasses different developmental stages. The target audience for ACE includes linguists, psychologists, neuroscientists, computer scientists, speech and language therapists and education specialists.

ACE offers the following services through its website:

- Information and guidelines about:
  - consent (forms);
  - hosting corpora and datasets containing atypical communication;
  - where to find corpora and datasets containing atypical communication;
- Helpdesk/consultancy for questions on the above topics;
- Technical assistance with designing, creating, annotating, formatting and adding metadata to resources of atypical communication;
- Outreach: publications, workshops contributions, etc.

[40] https://ace.ruhosting.nl/

Data originating in the context of atypical communication is particularly sensitive as regards privacy and ethical issues. While collecting, storing, processing and using such data, researchers are bound by strict rules and procedural requirements imposed by ethical committees and the GDPR. For hosting data and corpora for atypical communication and making these accessible in a FAIR- and GDPR-compliant manner, CLST has established close collaboration with The Language Archive (TLA) at the Max Planck Institute for Psycholinguistics (MPI) in Nijmegen. TLA, which is a CLARIN B-centre, offers storage of sensitive data (audio, video and transcripts) in a CMDI-supported repository that provides strong authentication procedures, layered access to data, and persistent identification. They focus on collecting spoken and signed language materials in audio and video form along with transcriptions, (speech) analyses, (linguistic) annotations and other types of relevant material, such as photos and accompanying notes.

For corpora of speech from people with language disorders, ACE works closely together with the DELAD initiative, whose goal is to facilitate the sharing of disordered speech corpora among researchers in a GDPR-compliant manner. Especially for these types of resources there is close collaboration with Carnegie Mellon University TalkBank / Clinical banks. Our collaboration makes it possible for our corpora and datasets to be registered at TalkBank and obtain its metadata and landing page at the TalkBank website. By contrast, the storage and authentication of access to the raw data (commonly audio and video data) is handled at TLA.

At all stages, appropriate measures must be in place so as to prevent unwanted disclosure. In some cases, this requires that the original data remains stored in a dark archive so that it is not accessible to users and cannot be copied or distributed in any form. To this end, ACE provides through its website a helpdesk to advise resource owners and users on how they can preserve sensitive data in a safe manner, from the point at which the raw data come into existence up to the moment where the data and information obtained from it are shared with others. Assistance in designing and collecting corpora containing atypical communication with consent forms that are GDPR-proof is considered of great value, as are references to available guidelines and tools for annotating such resources. How to make the resources accessible and share them with other researchers is another issue for which special expertise is requested.

Atypical communication data is also special when it comes to the methods and tools for processing and using it, if only because specific requirements apply in the light of the GDPR. Often guidelines and tools that have been developed for standard data cannot be used for atypical communication data or require adaptations or special settings; in some other cases dedicated tools are available. For example automatic speech recognizers which are used to generate transcriptions from audio files have a far lower performance when used for the speech

of language learners or of people suffering from dysarthria. ACE is thereby well positioned to inform researchers who want to work with language development data, data from adults and children with speech disorders, or users of sign language, on the availability of such tools and guidelines.

Figure 67: The main areas of research of K-ACE.

Reference:

Van den Heuvel, H., Oostdijk, N., Rowland, C., and Trilsbeek, P. 2020. The CLARIN Knowledge Centre for Atypical Communication Expertise. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC2020)*, 3312–3316.

Interview | **Katarzyna Klessa** and **Anita Lorenc**

Katarzyna Klessa and Anita Lorenc are experimental phoneticians who have prepared a Polish corpus of disordered speech, which has been released in a GDPR-compliant manner in collaboration with the ACE Knowledge Centre.

Could you describe your academic background, research interests, and current position? What inspired you to focus your research on speech prosody and the creation of corpora for speech technologies?

Katarzyna Klessa (KK): My academic background is primarily in experimental phonetics and applied linguistics. I work as a university professor at the Department of Multimodal Communication, Institute of Applied Linguistics at Adam Mickiewicz University in Poznań, Poland. My research interests originate in my fascination with two areas: human communication and technology. My formal education was technical only in secondary school, where I graduated from a technical automation class. Afterwards, I continued with a more humanities-oriented university education at the Faculty of Modern Languages and Literatures here in Poznań.

In my research, I am interested in the amazing phenomena that occur during human communication when people produce speech, hear and perceive sound, and interact with one another. I am excited about technology because it is something that gets us closer to inspecting the nature of these phenomena, by allowing us to precisely measure the properties of the sounds that we generate and decode. Importantly, communication also goes beyond the sounds themselves since we communicate with the way we shape the form of words and use melody. We also use certain special sounds such as hesitation markers as well as gestures or mimicry.

I have been involved in a number of different projects that focused on designing, creating, annotating and exploring speech corpora. At some point, my colleagues and I also decided to develop our own tool called Annotation Pro to support such tasks. Because speech corpora are used to verify assumptions about spoken language and are also fundamental for the development of speech synthesizers or various speech or speaker recognition applications, they are highly appreciated in disordered speech analysis, speech therapy and training. Corpus-based technologies may be helpful in both research-educational and clinical practice. These disciplines have been the common ground for Anita and me since the times of our PhD studies, when we first met at the Department of Psycholinguistics in Poznań.

Anita Lorenc (AL): Yes, we met in Poznań, where I wrote my PhD thesis under the supervision of Professor Piotra Łobacz. The thesis was about the acoustic analysis of the speech of hearing-impaired children (e.g., VOT parameter). Currently, I work at Warsaw University, at the Institute of Applied Polish Studies as a phonetician and as a Polish philologist and speech, language and hearing pathologist. I am the head of the Maria Przybysz Piwko Laboratory of Applied Phonetics at the University.

As the head of a research project funded by the Polish National Science Centre, I initiated a study of contemporary Polish pronunciation based on electromagnetic articulography, which is the first wide-scale research of this type in Poland. As an important outcome of the project, we established an interdisciplinary research-development team that consists of Łukasz Mik and Daniel Król from the University of Applied Sciences in Tarnów, in addition to Katarzyna Klessa and myself. Katarzyna, Łukasz and I have also collaborated on a smaller project on an articulography study of infant- and adult-directed speech in the Maria Przybysz Piwko Laboratory.

My habilitation thesis was dedicated to the normative pronunciation of Polish vowels and lateral consonant /l/ using electromagnetic articulography and a newly developed microphone array. In recent years, I have been involved in several other projects where corpora of speech disorders were created.

Corpora of speech disorders are difficult and expensive to obtain and collect, and can be tricky to distribute because of privacy issues, which is why both I and Katarzyna have been great supporters of the DELAD group since 2017, which is an initiative to establish a digital archive of disordered speech in a GDPR-compliant way and at secure repositories in the CLARIN infrastructure.

>

How did you get involved with the ACE CLARIN Knowledge Centre?

<

KK: The involvement with ACE started from the DELAD group and Henk van den Heuvel's proposal to share with ACE some of the spoken language data collected by individual DELAD members. We discussed data-sharing issues during our meetings in Cork and Utrecht.

AL: Yes, and it was the DELAD group who inspired me to share my PhD dissertation data related to the speech of hearing-impaired children with the help of Henk, Katarzyna and the ACE centre who took care of the technical aspects of preparing the data for publication.

>

Could you describe this corpus? Which kinds of speech disorders does your corpus contain? What was ACE's involvement in the creation process?

<

AL: The corpus is based on read and elicited speech recordings that I collected as part of my doctoral dissertation around 15 years ago.[41] The utterances were produced by hearing-impaired Polish children. Their pronunciation included various kinds of disorders specific for that group of speakers, for example voicing disorders. I described a number of these disorders in publications that are listed in the corpus public profile. Aside from the speech recordings, the original version of the data includes some basic metadata such as information about the speakers' age and gender. The elicited recordings come from a picture naming test. The orthographic transcription that is accessible along with the recordings is based on the prompts presented to the children during the recording sessions.

[41] https://phonbank.talkbank.org/access/Clinical/PCSC.html

KK: We received practical guidelines on how to prepare Anita's collection and how to properly describe its metadata for distribution. ACE helped us by quickly answering all kinds of questions and also with technical support before we could actually share the data. For example, the original material in Anita's collection was saved as separate files. The utterance transcriptions (prompts) were available from audio file names. I extracted the transcriptions into text file format. Then ACE helped us combine the multiple isolated files into a single larger one which is easier to handle in many cases.

>

What kind of research can be carried out with such a corpus?

<

KK: The corpus can be used for the analysis of certain paralinguistic components of the utterances, such as the acoustic and perceptual correlates of disordered speech, maybe even the speaker's vocal effort or voice quality. It could also be useful in contrastive studies of speech intelligibility in the contexts of hearing vs. hearing-impaired speakers. I believe the material might also be a great sample for teaching and demonstration purposes for students of phonetics, phonology and, primarily, speech pathology and speech therapy.

AL: The children whose disordered speech is presented in the corpus were educated using the Cued Speech method, so it would be interesting to compare the phonetic features of their pronunciation with the features of a similar group of hearing-impaired children educated with a different method. In addition to voicing disorders, which I described in my publications, there are noticeable disorders of speech fluency and the rhythm structure of speech, expressed by chanting and syllabifying utterances. Perhaps they are caused by the need to make gestures illustrating syllables and their particles in the Cued Speech method? A subjective, perceptual assessment of this group of children also suggests a non-standard realization of certain speech sounds – for instance, whereas a healthy speaker would pronounce certain sounds as just one consonant or vowel, a hearing-impaired child seems to pronounce them as a diphthong or even more complex sound.

>

What kind of speech technologies can be developed on the basis of this corpus?

<

KK: The corpus is not very large, so considering the fact that contemporary solutions often require large amounts of data, it might not be sufficient by itself as a basis for speech technology development. But I still believe that it can become a valuable resource for this purpose. I imagine it could become a subset of a larger resource dedicated to testing or training tools for speech or speaker recognition. Another example might be educational applications where various speech samples are contrasted to demonstrate certain phonetic features of utterances. In addition, the recordings of isolated words in the corpus could possibly serve as perception test stimuli. And perception tests are often a practical tool used to evaluate various speech applications.

>

Which GDPR issues did you encounter when developing the corpus, and how did ACE assist you on this matter?

<

KK: The DELAD team and ACE are experts on legal or administrative issues. We discussed GDPR in much detail during the 2019 DELAD/CLARIN workshop in Utrecht and while preparing the DELAD group publications. The GDPR issues that emerge when sharing disordered speech corpora are often even more problematic than those observed for corpora including the speech of healthy persons. The main reason is that for some studies it is necessary to include medical information and highly sensitive personal data. What is also important to note is that the data for our corpus were collected 15 years ago, which is long before the GDPR and when the conditions for licensing and participant consent were quite different and less strict than they are today. So, for the current corpus the data had to be anonymized and limited. The metadata in the corpus now includes speakers' gender and age, but without names or other personal information.

>

Aside from ACE, which helped you with the curation process, the corpus is also associated with another CLARIN Knowledge Centre, TalkBank, which will release the corpus. Why did you choose TalkBank as the depositing service, and why was it important for you to involve several K-Centres in your work?

KK: TalkBank is one of the more popular online services dedicated specifically to disordered speech,[42] so it offers a highly visible platform for reaching out to a very important group of possible users of the corpus. My participation in DELAD and discussions with researchers dealing with corpus-based studies of disordered speech have only reinforced my conviction of the importance of releasing the corpus through TalkBank. Additionally, the idea behind DELAD as well as CLARIN in general is to share as much data and disseminate it as broadly as possible for the sake of progress in research and technology. As a member of the programme board for the Polish CLARIN consortium, I have observed very intensive efforts towards not only developing the technological infrastructure, which of course is very important, but also towards reaching out to as many potentially interested parties as possible. Importantly, such efforts necessarily involve several institutions, each with its own specialization, which is also why it was so important to involve several K-centres, as ACE helped us with the creation and documentation of the corpus, while TalkBank helped with its dissemination.

[42] https://www.clarin.eu/blog/talkbank-clarin-knowledge-centre

Do you plan to continue collaborating with CLARIN K-centres in the future? When would you advise your fellow researchers to seek help from K-centres?

KK: Yes, definitely!

I see the collaboration with CLARIN K-centres as a very successful and instructive experience. First of all, I am happy that the data have received the chance for a "second life" with the help of ACE and other centres. Secondly, I learned a lot about the procedures and data flow, which is very interesting to me as a person involved in the creation and maintenance of several digital databases and repositories.

I am therefore grateful for the opportunity to participate in the data curation process and hope for more opportunities like this. We will probably continue with some other Polish datasets of disordered speech or other types of recordings, either archival or maybe new ones. As for advice, I would say that if you need help or expertise related to language tools or data, CLARIN K-centres are a great choice. The experience with ACE has shown me that it's both easy to start working with a K-centre and that you can expect a lot of support and encouragement.

The LUND University Humanities Lab Knowledge Centre

Introduction

Written by **Johan Frid**

Lund University Humanities Lab is a department for research infrastructure, interdisciplinary research and training. Since 2017, the Lab has been a certified CLARIN Knowledge Centre with a special focus on multimodal and sensor-based methods.[43] As of 2020 we are also a CLARIN C-centre, meaning that our datasets are integrated with CLARIN's Virtual Language Observatory (VLO). The Lab is a member of the Swedish national consortium for language resources and technology, Swe-Clarin.

We provide access to sensor-based technologies, methodological know-how, data management, and archiving expertise. Our mission is to facilitate and help diversify research around the issues of cognition, communication, and culture – traditional domains for the Humanities. That said, many projects undertaken at the Lab are interdisciplinary and conducted in collaboration with the Social Sciences, Medicine, the Natural Sciences, Engineering, and e-Science. The Lab enables researchers to combine traditional and novel methods, and to interact with other disciplines.

We have a wide range of facilities for measurement and recording: articulography, electrophysiology, EEG, eye-tracking, professional audio and video recording, motion capture and virtual reality. The Lab also offers support and consultancy on statistics, machine learning-related research on language data, and keystroke logging for the study of the writing process.

As a node in the Swedish national infrastructure Swe-Clarin, the Humanities Lab provides speech and language technological support to a wide range of projects and contributes to the development of resources for Swedish language technology. For instance, the Swe-Clarin consortium has formed a thematic working group to develop a resource for benchmarking Swedish Named-Entity Recognition and Classification (NERC) systems. The NERC group links Swe-Clarin nodes at Lund, Gothenburg and Linköping. The aim is to develop a tool for finding and replacing Swedish names in written materials in order to anonymize or pseudonymize them. All the resources developed in this working group will be made available from the Language Bank Of Sweden.

The Lab also provides tools and expertise related to language archiving, corpus and (meta)data management, with a continued emphasis on multimodal corpora, many of which contain Swedish resources, but also other (often endangered) languages, multilingual or learner corpora. A primary service is the Lund University Humanities Lab corpus server, containing a varied set of multimodal language corpora with standardized metadata and linked layers of annotations and other resources.

The corpus server hosts two sets of corpora, the Lund Corpora, and the Repository and Workspace for Austroasiatic Intangible Heritage (RWAAI) corpora. The facility contains a wide variety of data types including audio, video, text, images, and eye-movement data. The Lund Corpora offer data from major world languages and lesser-described minority languages, including longitudinal child language studies, adult language acquisition data, dialect surveys, and corpora with linked eye-tracking data. The RWAAI corpora constitute a unique digital resource preserving multidisciplinary research collections documenting the languages and cultures of communities from the Austroasiatic language family of Mainland Southeast Asia and India. The collections span over half a century of research in fields such as linguistics, anthropology, botany, ethnomusicology, and human ecology. More than 50 predominantly endangered minority languages are currently represented in the collection. Metadata is provided in CMDI format, and harvested by CLARIN's VLO.

[43] https://www.humlab.lu.se/

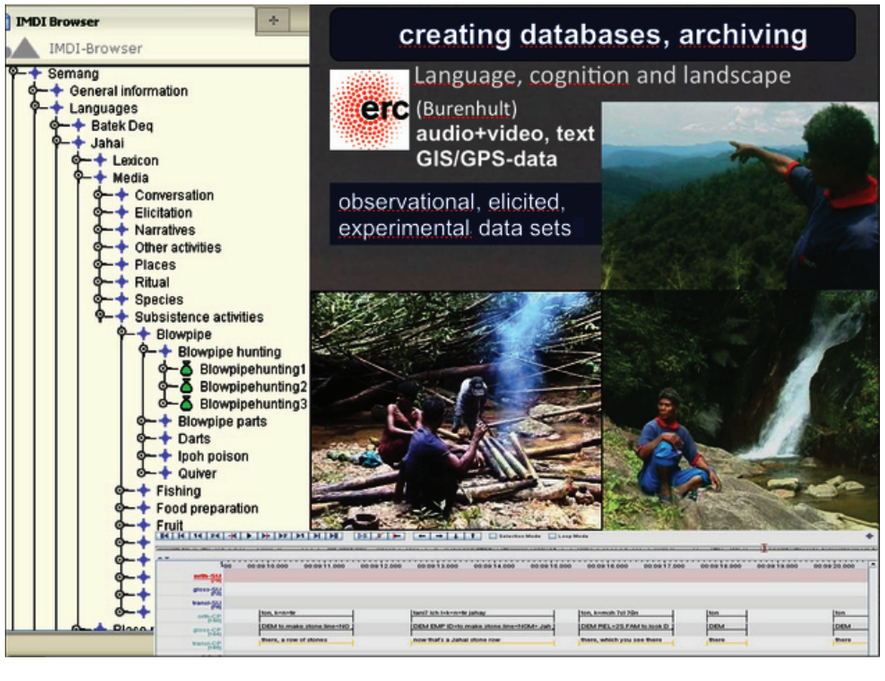


Figure 68: The Lund corpus server.

Interview | **Gerd Carling**

Gerd Carling is Associate Professor at the Department of Linguistics at Lund University who specializes in linguistic phylogenetics. She is the main editor of the Diachronic Atlas of Comparative Linguistics, which is hosted by the LUND CLARIN Knowledge Centre.
[Photo: Idun Blomé.]

What sparked your interest in historical linguistics? What motivated you to take a digital approach in this field?

<

My interests within linguistics have always been in the direction of typology and historical linguistics. Even though I began my PhD and postdoctoral studies within a more traditional, philological orientation, I was always interested in the digital aspects of humanities, including both corpus and phylogenetic linguistics. Languages, like biological populations, inherit traits from their ancestors, diverge into different lineages, go extinct, and engage in horizontal transfer. Therefore, linguistic phylogenetics and similar computational historical methods draw extensively on techniques pioneered in biology.

>

You are the main editor of the Diachronic Atlas of Comparative Linguistics (DiACL) database.[44] How did this project come about?

<

The compilation of a typological and lexical database started as a project involving teachers and students in courses at the Department of Linguistics in Lund. We were all dealing with the typological aspects of language, both of grammar and of lexicon, and our field sites and expert areas were distributed around the world. Some researchers and students were more interested in grammar, others inclined more towards lexicon, some worked on Indo-European modern and

[44] https://diacl.ht.lu.se/

ancient languages, and some worked with Austronesian languages, while others worked with Amazonian ones.

To bring all of these approaches together, we decided to compile and pool data in a more consistent way, so that we could more easily collaborate and compare our findings. I developed two projects that allowed us to do this in a more systematic fashion, where we were able to hire students to compile data and started planning the database infrastructure. At the beginning (2013–2015), the knowhow for building this type of database in Lund was relatively restricted, and we discussed several possible solutions for designing the resource, mainly with the Lund University GIS Centre and the IT Department at the Faculty of Humanities and Theology, which provided support for licences and the database infrastructure.

>

What is the role of the Lund University Humanities Lab CLARIN Knowledge Centre in DiACL?

<

Very soon, in the early stages of the project, we met a number of difficult obstacles, not the least digital ones. The first pilot infrastructure for DiACL that we built with a programmer from the IT department was not able to meet the demands of the linguistic database that we wanted to build. Therefore, I had to recruit a programmer from Leiden University with education in historical linguistics, who took over the programming and building of the infrastructure, with the support of programmers from the IT department. The hosting of the database was taken over by the Lund University CLARIN K-Centre together with the Swedish CLARIN consortium. In particular, the building of the lexical database turned out to be very complex and tricky. Finally yet importantly, the migration of the data (2016–2017), which was compiled in CSV files, into the database infrastructure generated an enormous amount of errors and mistakes, which required many recodings and screenings of the database.

The Lund Humanities Lab and the CLARIN consortium played an important role in overcoming these problems, mainly in the later phase of the project (2016–2019), when we worked towards getting the data ready for publication. Researchers from the Humanities Lab, in particular Johan Frid, who is the coordinator of the Lund University CLARIN K-Centre, were very important in designing the code for extracting and analysing the data.

>

Could you briefly describe the make-up of this database (e.g., overall size, the languages and time periods covered by the database)?

<

The database currently contains 569 languages, including 21,542 grammatical data points (i.e., generalizations of grammatical structure in languages) and 72,033 lexemes (words in languages), which are connected by 43,095 cognacy links (connections of words to a joint ancestor in a tree-like structure) and 71,730 links to concepts (connections between a word in a language to a prototype meaning, such as MOTHER). The data spans 25 language families and covers a period of 3,500 years, making it one of the largest diachronic databases, as well as one of the most well-annotated ones.

>

Why is DiACL important for the historical linguistics community? What are the main features of the database that are especially useful for conducting diachronic research in lexicography, phonology or morphosyntax?

<

The DiACL database is special since it is a joint grammatical and lexical database. In addition, the database has a strictly comparative *and* diachronic focus. Many other databases of the same type, such as the World Atlas of Language Structures (WALS) or South American Indigenous Language Structures (SAILS), are synchronic and focus on either typology, morphology, or lexicon. Additionally, the DiACL database is now an integrated part of two larger attempts to bring together all similar databases globally – CLICS for lexicon and Grammaticon for grammar.

>

Have you used the DiACL database in any of your own recent research or publications? Could you briefly discuss some of the results?

<

DiACL has been used in several recent publications, first and foremost the monograph *Mouton Atlas of Languages and Cultures* (Carling 2019). This monograph compiles and publishes the part of the data that covers Eurasia and discusses both the motivations for this work and findings from the database in a non-technical manner that is accessible to researchers, students, and interested non-academics. A second volume to come will deal with South America in the same fashion.

My colleagues and I have also published several papers on the phylogenetic trends of typology on the basis of our database. For instance, in one paper (Cathcart, Carling et al. 2018), we present a case study using the Indo-European data in DiACL to show that linguistic complexity is dependent on the notion of "areality", which means that genealogically unrelated languages come to share linguistic features, often because they are spoken in the same geographic area. In another paper (Carling, Larsson et al. 2018), we examine the Eurasian subset of DiACL, which includes Indo-European, adjacent languages from different families, and earlier states of contemporary languages, dead branches, as well as later stages of migrated languages, from the earliest sources up to the modern period, to show how languages developmentally cluster by areality on the one hand and genealogy on the other.

In a forthcoming study, we will demonstrate that universal hierarchies of grammar, connected to the frequency and economy of language (e.g., the singular is more common than plural), affect the general evolution of grammar. In another study, we will show that grammatical gender correlates with environmental aspects and spreads by migration, due to its stability within the language family (i.e., gender systems do not change or disappear). In lexicology we have demonstrated that borrowing depends on cultural factors and is highly affected by sociolinguistic factors, including language size and power (Carling, Cronhamn et al. 2019). Other studies to come will investigate various causes for gender assignment, based on large amounts of lexical data.

>

Do your graduate students used DiACL in their own work, and would you like to highlight any of their recent findings or publications?

<

The database and the research involved in the database are highly integrated with the work by graduate students. Often, students work in parallel on a project involving the big data of the database, and their own, more limited datasets, which focus on an individual language or a branch of a family. One MA student, Anne Goergens, investigated alignment in Arawakan languages. Another student, Sandra Cronhamn, looked at lexical borrowing in Tupí, using DiACL data. Filip Larsson wrote his MA thesis on areal structure in typology. All based their work on DiACL data.

>

Aside from DiACL, have you collaborated with the Lund University Humanities Lab in any other research project? And could you please briefly describe what you have done in this regard?

<

My work with the Lund University Humanities CLARIN Centre is mostly connected to the type of research that relates to work with the DiACL database. However, I have involved the computational knowhow of the Lund University Humanities CLARIN Centre in other projects where we have used data from other international databases, including a recent project with the NorthEuralex lab, where we find that the stability of sounds over time correlates with preference in first language acquisition.

>

References:

Carling, G. (ed.). 2019. *The Mouton Atlas of Languages and Cultures*. De Gruyter Mouton.

Carling, G., Cronhamn, S., Farren, R., Aliyev, E., and Frid, J. 2019. The causality of borrowing: Lexical loans in Eurasian languages. *PLOS ONE*.

Carling, G., Larsson, F., Catchcart, C., Johansson, N., Holmer, A., Round, E., and Verhoeven, R. 2018 Diachronic Atlas of Comparative Linguistics (DiACL)–A database for ancient language typology. *PLOS ONE*.

Cathcart, C., Carling, G., Larsson, F., Johansson, N., and Round, E. 2018. Areal pressure in grammatical evolution An Indo-European case study. *Diachronica* 35 (1): 1–34.

The Spanish CLARIN Knowledge Centre

Introduction

Written by **Ainara Estarrona, Mikel Iruskieta, German Rigau** and **Núria Bel**

The Spanish CLARIN K-Centre was the first CLARIN K-centre to be established.[45] In 2015, CLARIN granted the K-centre recognition to three research groups that decided to work together as a distributed centre with the aim of supporting research in all the languages of Spain. Initially the Spanish-K-Centre covered Spanish, Basque and Catalan, but soon after a Galician specialist research group joined the Spanish-K-Centre.

Thus, the Spanish CLARIN K-centre brings together the specialities of the following four nodes or groups with complementary skills and experience:

- IIULA-UPF-CCC (UPF Barcelona) specializes in Text Analytics and Language Technologies and Resources. This centre offers services to researchers working with textual data, especially in Spanish and Catalan. This group has led Spain's participation in CLARIN as a natural language processing web service provider. It has developed several web applications, such as ContaWords. It also provides web service access to tools for enriching texts with syntactic dependencies, part-of-speech tagging and named-entity recognition using the open source language analysis tool suite FreeLing.
- ILINDH (UNED Madrid) is a research centre on Digital Humanities that works as a hub for innovation, consultancy and training to support researchers and projects in Spain and other Spanish-speaking countries.
- IThe IXA group (UPV/EHU Donostia-San Sebastián) is a multidisciplinary team that includes members of different research areas at the University of the Basque Country, such as linguists, computer scientists, and educators. It is specialized in Text Analytics and Language Technologies and Resources and offers services to researchers working with Basque, Spanish and English. It has developed more than 20 NLP tools, some of which have been reused and redesigned for Digital Humanities and Social Sciences researchers who have requested technical support through the CLARIN K-centre. An example of such a tool is ANALHITZA, which is a user-friendly web service for the general linguistic analysis of Basque, Spanish, and English.
- ITALG (UVigo, Vigo) specializes in the development of digital language resources for facilitating the use of Galician, its translation, study, and learning in the contemporary digital framework. TALG has developed the automatic analyser DContado, which helps researchers extract linguistic information from texts in Spanish, Galician, and English.

The Spanish CLARIN K-Centre supports an average of five requests per year from postdoctoral researchers and senior researchers in a variety of topics, such as poetry analysis, identification of linguistic features for detecting fraud, analysis of bilingual child corpora, and representing discourse relations in political debates. In addition to creating user-friendly tools like ANALHITZA and DContado, the K-centre has also participated in several summer schools and the development of teaching materials for promoting the benefits of the CLARIN infrastructure.

The Spanish CLARIN K-Centre has supported the creation of a strategic research network, INTELE, in order to promote the participation of Spain in CLARIN ERIC as well as other European infrastructures like DARIAH. The INTELE network aims to reduce the digital divide in Spain by promoting new lines of multidisciplinary research in the Humanities and Social Sciences and by participating in their digital transformation with the help of language technologies. The seven research groups and universities involved in INTELE are the following:

- IULATERM-TRL-UPF Group at Pompeu Fabra University (Barcelona)
- Ixa Group at the University of the Basque Country (Donostia-San Sebastian)
- LINDH at the National Distance Education University (Madrid)
- La otra Edad de Plata: Proyección Cultural y Legado Digital Group (LOEP-UCM) at the Complutense University of Madrid (Madrid)
- Fundación Biblioteca Virtual Miguel de Cervantes at the University of Alicante (Alicante)
- Galician Language Tecnologies and Applications Group at the University of Vigo (Vigo)
- Grupo de Sistemas Inteligentes de Acceso a la Información (SINAI) at the Jaen University (Jaen)

[45] http://clarin-es.org/en/

The specific objectives of the INTELE network are to coordinate and organize the activities of the seven research groups in order to support and foster the interest and use of the CLARIN and DARIAH infrastructures for Digital Humanities and Social Sciences, encourage meetings between researchers working in Digital Humanities and Social Sciences, and create a catalogue of tools and use cases made by researchers working with Spanish and the co-official languages Basque, Catalan-Valencian and Galician. INTELE also aims to promote the collaboration and the internationalization of the seven Spanish research groups, so that Spain can become a member of the European infrastructures CLARIN and DARIAH.

Figure 69: The kick-off meeting of INTELE.

Interview | **Jose Pérez-Navarro**

Jose Pérez-Navarro is a PhD candidate at the University of the Basque Country who works in developmental cognitive neuroscience.

Please introduce yourself. Could you briefly describe your research background and current position?

<

I am a developmental cognitive neuroscientist, interested in the cognitive and neural underpinnings of language acquisition. Currently, I work as a predoctoral researcher at the Basque Centre on Cognition, Brain and Language (BCBL), and I am a PhD candidate at the University of the Basque Country (UPV) at San Sebastian, Spain.

>

What is your involvement with the Spanish CLARIN K-Centre? Could you briefly describe the collaboration?

<

My collaboration with the Spanish CLARIN K-Centre is through Mikel Iruskieta, who is the coordinator of the centre. I reached him when I read the article by Otegi et al. (2017) about ANALHITZA,[46] which is a tool for performing linguistic annotation of Basque, English and Spanish. I thought the tool could be used to enrich the corpus that me and my PhD supervisor Marie Lallier are working on, since to my knowledge it is the only tool for performing language annotation equally in Basque and Spanish. Although it is not specifically designed for spoken language (it is rather primarily aimed at written registers), it does a great job in annotating our corpus, which includes speech productions of Basque-Spanish bilingual children, and consequently a considerable proportion of Basque-Spanish code-switching.

>

[46] http://ixa2.si.ehu.es/clarink/sarrera.php?lang=es

How have you used the tools ANALHITZA? What made it beneficial for your research purposes?

<

Thus far I have benefited from the tool's lemmatization component. We have relied on the expertise of Mikel Iruskieta and the IXA group who helped us batch-analyse the corpus, and specifically the subsets that correspond to the two first stages of bilingual language learning, which takes place in children who are between four and five years old. ANALHITZA is also crucial for me because it takes into account both Spanish and Basque, and provides outputs with comparable indexes in each language. That is to say, the tool has allowed us to determine, in an equally robust way for both languages, whether linguistic aspects like the lexical diversity or the morphosyntactic complexity of bilingual children are similarly dependent on factors of the bilingual environment, such as the amount and quality of the exposure to each of the languages, rather than the greater linguistic differences between the two languages.

>

Could you describe some of your on-going work in child language acquisition and bilingualism?

<

My PhD work focuses on the amount of exposure to each language within bilingual contexts, and how it shapes language acquisition at a cognitive and neural level. To this end, my supervisor and I are acquiring data on children who have not yet learnt to read, and analysing how different naturalistic measures of language acquisition, such as spontaneous speech productions or the brain's ability to synchronize speech in either Spanish or Basque, are influenced by the relative amount of exposure to each of these languages. Some preliminary results of this longitudinal project, in which we also use ANALHITZA, show that the amount of exposure and the children's age are crucial factors for the acquisition of phonology, vocabulary and morphosyntactic proficiency. In other words, we show that, at least during early language learning, when the vast majority of linguistic input is oral, proficiency in a given language is highly dependent on the amount of exposure to it.

>

Why is the Spanish CLARIN K-Centre especially important for researchers working with Basque?

<

I believe that the K-centre is especially important because it provides knowledge on two languages, namely Basque and Spanish, that share only a small amount of linguistic complexity and features. Looking at both of these two languages contrastively can lead to relevant discoveries about what aspects of the influence of the amount of exposure to a language on language learning are more universally shared (e.g., phonological and speech comprehension abilities that allow us to understand speech as it unfolds over time) and which features could be more specific to either Spanish or Basque (e.g., acquiring a proficient degree of morphosyntactic knowledge).

The IXA group, which is leading the Spanish Knowledge Centre, is supporting us with their expertise on corpus annotation and language tools like the aforementioned ANALHITZA. We are thus able to more efficiently analyze how Spanish and Basque develop, and obtain a precise snapshot of how they are used in everyday settings.

>

Are there any tools and resources that you would like to see developed in the future that either your field or the Basque language community could benefit from?

<

In the field of bilingual language learning there are already several useful tools that the IXA group has developed; apart from ANALHITZA, there is also Ixati,[47] which provides morphosyntactic analysis and phrase chunking for Basque. I do not know if it is already in the catalogue of the Knowledge Centre, but a tool that accounts for code-switching between Basque and Spanish in a single utterance and disentangles whether it is a Basque utterance with certain Spanish nouns or the other way around would be wonderful for working with corpora of natural language. Another example of a tool that would be valuable for our current research would be a syntactic parser trained on Basque and Spanish spoken language, which to my knowledge does not yet exist.

>

[47] http://ixa.si.ehu.es/node/4455

What is in store for your future collaboration with the Spanish CLARIN Knowledge Centre?

<

Since we work on longitudinal projects, there are several linguistic aspects of our emergent corpora that we cannot predict very precisely, especially given that they have not been previously assessed in the Basque-Spanish bilingual combination. For instance, it is still unclear to which extent the morphosyntactic development of children remains independent in each language and at which age children start to generalize the knowledge in one language to build increasingly complex syntactic structures in both languages. By finding the most suitable methods that can account for all the linguistic variability that takes place in speech production during language learning, we can capture the emergence of structures of increasing complexity in both languages. Therefore, we plan to reassess whether the existing tools can be improved by the Knowledge Centre to account for such variability at an even more efficient rate – I am also happy to say that thus far the centre has always accommodated our requests.

>

Reference:

Otegi, A., Imaz, O., de Ilarraza, D., Iruskieta, M., and Uria, L. 2017. ANALHITZA: a tool to extract linguistic information from large corpora in Humanities research. *Procesamiento del Lenguaje Natural* 58: 77–84.

COLOPHON

Coordinated by
Darja Fišer and **Jakob Lenardič**

Edited by
Darja Fišer and **Jakob Lenardič**

Proofread by
Paul Steed

Designed by
Tanja Radež

Cover image:
National and University Library of Iceland (row 1, image 2)
Dominique Boutet, Jean-François Jégo, Vincent Meyrueis (row 1, image 3)
Anonymous, collected by Nadine Vakhnovsky (row 2, image 1)

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

Publication number
CLARIN-CE-2020-1781
November 2020

ISBN
9789082990928

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 2020