# Tour de CLARIN Volume Four

Darja Fišer, Jakob Lenardič, Francesca Frontini, Erik Axelson, Tomaž Erjavec, Maria Gavriilidou, Krzysztof Hwaszcz, Taja Kuzman, Krister Lindén, Therese Lindström Tiedemann, et al.

CLARIN

Common Language Resources and Technology Infrastructure

# Tour de CLARIN

VOLUME **FOUR**

Edited by **Jakob Lenardič, Francesca Frontini,** and **Darja Fišer**


Common Language Resources and Technology Infrastructure

# Tour de CLARIN

VOLUME **FOUR**

Edited by **Jakob Lenardič, Francesca Frontini,** and **Darja Fišer**

CONSORTIA





K-CENTRES



B-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. Originally only focussing on CLARIN consortia, this initiative was expanded twice, first in 2019 to also feature the work of CLARIN Knowledge Centres (or K-centres), which offer knowledge and expertise in specific areas to researchers, educators, and developers alike, and second in 2021 to feature Service Providing Centres (or B-centres), which serve as the technical backbone of the CLARIN infrastructure. For almost five years, Tour de CLARIN has been one of the flagship outreach initiatives, thus far released in the form of three printed volumes.

The fourth volume is organized in two parts. In Part 1, we present CLARIN Portugal in five chapters: an introduction to the consortium, its 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 six K-centres and two B-centres that have been visited since the publication of the second volume in November 2020: the K-centre for morphologically rich languages SAFMORIL, the French CORLI K-centre, the K-centre for South Slavic languages CLASSLA, the NLP:EL K-centre for Greek, the Austrian B-centre ARCHE, and the CLARIN-PL B-centre. Each centre is presented in two chapters: a presentation of what the centre offers to researchers and an interview with a renowned researcher who has benefited from the collaboration with the centre.

The volume would not have been possible without the contributions and dedication of the CLARIN national coordinators and user involvement coordinators, and centre representatives: Antonio Branco, João Silva, Erik Axelson, Eva Soroli, Nikola Ljubešić, Maria Gavriilidou, Martina Trognitz, and Jan Wieczorek.

We would also like to thank all the researchers who have kindly agreed to be interviewed for their time and invaluable insights: Pilar Barbosa, Jack Rueter, Thomas Gaillat, Zrinka Kolaković, Titika Dimitroulia, Peter Andorfer, Stephan Kurz, Martin Anton Müller, and Olga Czeranowska.

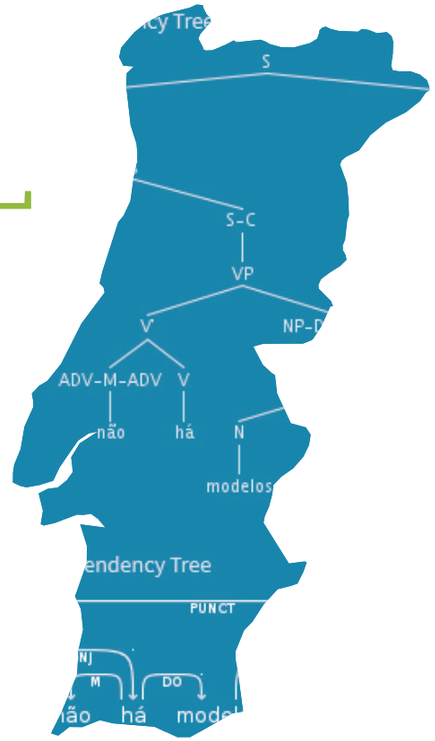
**Jakob Lenardič, Francesca Frontini,** and **Darja Fišer**
November 2021

Consortium featured in this volume:

Portugal

# PART 1
# CONSORTIA

# PORTUGAL




## Introduction

Written by **João Silva**

The PORTULAN CLARIN Research Infrastructure for the Science and Technology of Language is the CLARIN national consortium for Portugal,[1] a country that has been involved in CLARIN since the European EU-funded preparatory project for CLARIN began in 2008, and was invited in 2010 to be a founding member of what was to become CLARIN ERIC. The PORTULAN CLARIN consortium is currently formed by over 20 partners from various Portuguese institutions and fields related to language. Also included in the consortium are Camões I.P., the official national organization responsible with the promotion of the Portuguese language, and four partners from Brazil. The Director General as well as the National Coordinator of PORTULAN CLARIN is António Branco.

Scientific knowledge is grounded on falsifiable predictions and thus its credibility and raison d'être rely on the possibility of repeating experiments and getting similar results as originally obtained and reported. Scientific knowledge is also cumulative, with more recent advancements originating from developments obtained from previous breakthroughs. Crucial for the scientific endeavour, and for science-based activities, is the availability of data and of companion analytical devices. Also crucial is the moral, and many times physical, courage to challenge the status quo, together with the altruism to share the research results. Seeking to foster this scientific ethos, this research infrastructure is named *portulan*, which is based on that of portolan charts, designating the maps where discoveries by courageous sailors were documented in such a way that these discoveries and associated data could subsequently be confirmed or corrected, the original journey could be repeated with increased efficiency, and new discoveries and routes could be reached beyond those already known.

**Figure 1:** Portolan chart by Jorge de Aguiar (1492), the oldest known signed chart of Portuguese origin (Beinecke Rare Book and Manuscript Library, Yale University, New Haven, USA).

The mission of PORTULAN CLARIN is to support researchers, innovators, citizen scientists, students, language professionals and users in general whose activities rely on research results from the domain of science and language technologies. PORTULAN CLARIN supports them with the distribution of scientific resources, the supplying of technological support, the provision of consultancy, and the fostering of scientific dissemination. It supports activities in all scientific and cultural domains with special relevance to those that are more directly concerned with language – whether as their immediate subject, or as an instrumental means to address their topics – including among others, the areas of Humanities, Arts and Social Sciences, Artificial Intelligence, Computation and Cognitive Sciences, Healthcare, Language Teaching and Promotion, Cultural Creativity, Cultural Heritage, etc. It serves all those whose activity requires the handling and exploration of language resources, including language data and services, in all sorts of modalities, in all types of representations, and in all types of functions.

[1] https://portulanclarin.net/

The infrastructure pursues its mission by seeking to bring scientific, professional or personal advantages to its users by means of its operation being primarily centred on its users.

Users were involved with the infrastructure well before it entered into operation, right from the start of its planning, and contributed to its design and implementation through the Network of Implementation Partners. Users have a Helpdesk to contact when they need support in their utilization of the infrastructure, benefit from initiatives to enhance their engagement with the infrastructure. Users can also address the infrastructure at any moment and are encouraged to provide their feedback through the Scientific Advisory Forum.

Users have free access to the infrastructure, with no registration required and without any “members only” constraint. Through its repository, PORTULAN CLARIN allows users to access resources for language research. These resources include language processing tools and applications, as well as corpora, lexicons and various other kinds of data sets, such as word embeddings and records of brain potentials during reading. In addition, PORTULAN provides an online Workbench, through which users can run a variety of language processing tools. These include, among others, a corpus concordancer (CINTIL Concordancer); tools for language processing, such as dependency parsing (LX-DepParser), named entity recognition (LX-NER), and sub-syntactic annotation (LX-Suite); as well as tools for additional textual enrichment, such as the analysis of temporal relations and events in texts (LX-TimeAnalyzer).

Users distributing resources through the infrastructure are free to choose the distribution licences for their resources and grant PORTULAN CLARIN only the non-exclusive right to distribute those resources: users keep all rights, including the right to distribute their resources through other means, and to withdraw their resources from the infrastructure. No user data is retained related to their usage of the infrastructure, be it their scientific or personal data.

PORTULAN CLARIN ensures the preservation and fostering of the scientific heritage regarding the Portuguese language, thereby supporting the preservation, promotion, distribution, sharing and reuse of language resources for this language, which include text collections, lexicons, processing tools, etc. It represents an asset of utmost importance for the technological development of the Portuguese language and to its preparation for the digital age.



# Tool | **LX-DepParser**

Written by **João Silva**

LX-DepParser is a syntactic dependency parser for Portuguese.[2] In syntactic dependency parsing, a word (the head) is connected to one or more words (the dependents) by directed arcs, forming a directed graph. The link between a head and its dependent indicates that the occurrence of the dependent in its specific position in the sentence is made possible by the occurrence of the head. These arcs are typically labelled with the name of the grammatical function (e.g. subject, object, specifier, etc.) that mediates the relation between the head and dependent. The main predicate of the sentence, typically a verb, has no head and is marked by an arc labelled as a root.

LX-DepParser is based on the MaltParser machine-learning parsing engine, trained on roughly 22,000 sentences from CINTIL-DependencyBank, a corpus also presented in this Tour de CLARIN. Under a 10-fold cross-validation evaluation scheme, the parser achieves state-of-the-art scores for Portuguese, with 94.42% UAS (the unlabelled attachment score, or the amount of correct dependency arcs, ignoring the label) and 91.23% LAS (the labelled attachment score, or the amount of correct dependency arcs, also taking the label into account).

**Similar to many of the other tools in the PORTULAN CLARIN Workbench, LX-DepParser can be used in three different ways:**

- directly in the browser, which is convenient for processing small snippets of text or for getting a feel for the output given by the tool;
- by submitting files to be processed, which is useful for batch annotation of large amounts of data;
- and from code, through a web service API, which provides the greatest flexibility to users in integrating the tool into their own processing pipeline, but requires some coding knowledge.

When used in the browser, the user can input a small snippet of text and see the parsing result in the same browser window. This output can be shown in a user-friendly format, which provides a graphical representation of the dependency relations, meant for human readability. The other two in-browser output formats present the output in a machine-readable tabular format that matches the actual output format of the tool. The difference between these two machine-readable formats is in the tag set and grammatical dependency principles that are used: CINTIL

[2] https://portulanclarin.net/workbench/lx-depparser/

follows the tag set and principles defined in Branco et al. (2015), while the universal dependencies option converts the CINTIL dependencies into Universal Dependencies (de Marneffe et al. 2014).

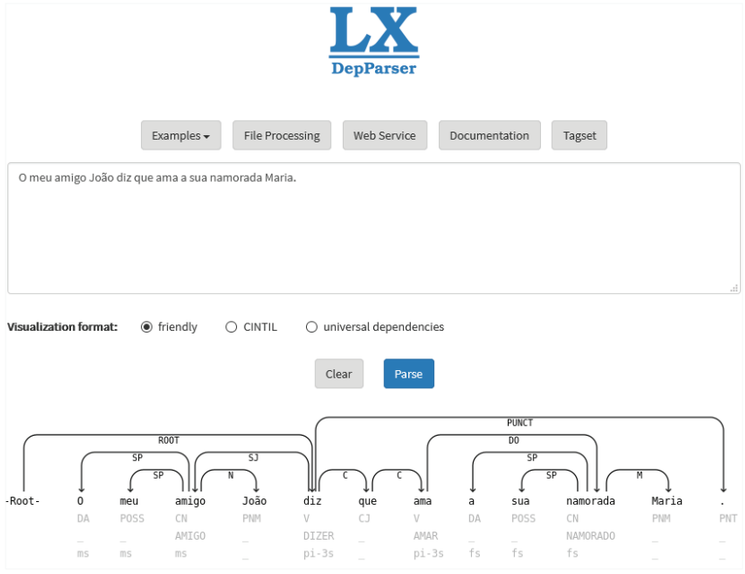



**Figure 2:** Output of the dependency parser, in-browser, in the "friendly" format.

Figure 2 shows the user-friendly output format of the tool for the example sentence *O meu amigo João diz que ama a sua namorada Maria* ("My friend João says he loves his girlfriend Maria"). The ROOT relation pointing to *diz* ("says") identifies it as the main predicate of the entire sentence, whereas the SJ (subject) relation pointing to *amigo* ("friend") defines the latter – or rather the entire phrase that it heads, *O meu amigo João* ("My friend João") – as the syntactic subject of the main clause.

LX-DepParser is run as the last step of a language processing pipeline. The preceding steps handle tokenization, part-of-speech tagging and morphological analysis. The annotation created by these steps is shown in grey beneath each word. For instance, *amigo* is annotated as a common noun (CN) with the lemma *AMIGO* as its base form and masculine singular (ms) as inflection features; and *diz* is annotated as a verb (V), with the lemma *AMAR* ("to love") as its base form and pi-3s (present indicative, third person singular) as its inflection features.



| #id | form | lemma | cpos | pos | feat | head | deprel | phead | pdeprel |
|---|---|---|---|---|---|---|---|---|---|
| 1 | O | - | DA | DA | ms | 3 | SP | 3 | SP |
| 2 | meu | - | POSS | POSS | ms | 3 | SP-ARG1 | 3 | SP-ARG1 |
| 3 | amigo | AMIGO | CN | CN | ms | 5 | SJ-ARG1 | 5 | SJ-ARG1 |
| 4 | João | - | PNM | PNM | - | 3 | N | 3 | N |
| 5 | diz | DIZER | V | V | pi-3s | 0 | ROOT | 0 | ROOT |
| 6 | que | - | CJ | CJ | - | 5 | C-ARG2 | 5 | C-ARG2 |
| 7 | ama | AMAR | V | V | pi-3s | 6 | C | 6 | C |
| 8 | a | - | DA | DA | fs | 10 | SP | 10 | SP |
| 9 | sua | - | POSS | POSS | fs | 10 | SP-ARG1 | 10 | SP-ARG1 |
| 10 | namorada | NAMORADO | CN | CN | fs | 7 | DO-ARG2 | 7 | DO-ARG2 |
| 11 | Maria | - | PNM | PNM | - | 10 | M-PRED | 10 | M-PRED |
| 12 | . | - | PNT | PNT | - | 5 | PUNCT | 5 | PUNCT |

**Figure 3**: Output of the dependency parser, in-browser, in the CINTIL format.

Figure 3 shows the output, for the same sentence, in the CINTIL tabular format. This format is akin to the commonly used CoNLL format and is amenable to being read by a computer. This is also the format produced by the other modes of accessing the tool: the file processing service and the web service.

For natural language processing, LX-DepParser has mostly been used as a component in larger processing pipelines, such as LX-SRLabeler (for semantic role labelling) or LX-Suite (a suite of shallow processing tools). It has also been used, for instance, to provide features for a machine-learning classifier in a work where the dependencies produced by the parser were used as features for a classifier that assigns deep lexical types for handling out-of-vocabulary words in a deep processing grammar (Silva 2014). These grammars make use of lexica with extremely fine-grained syntactic categorization, but cannot proceed when a word is not found in their lexicon, and relying only on the coarse annotation of a normal part-of-speech tagger leaves too much ambiguity unresolved for a useful and efficient analysis. The classifier was able to use the features provided by the parser to assign fine-grained tags.

For the study of language, the parser has been used to quickly provide a tentatively annotated corpus that was then manually corrected, leading to the creation of CINTIL-DependencyBank PREMIUM. Moreover, from personal communications we are aware that the dependency parser has been used in a classroom setting to show undergraduate students of Linguistics its grammatical analyses of input sentences. Note that the parser will sometimes have errors in its analysis, but these possible errors are then integrated into the discussion of the results with the students. If the teacher wishes to show only correct analyses, they can use the manually validated CINTIL-DependencyBank, though in that case they will be restricted to those sentences that are already in the corpus.

**References:**

Branco, A., J. Silva, A. Querido, and R. de Carvalho. 2015. CINTIL-DependencyBank PREMIUM handbook: Design options for the representation of grammatical dependencies. Technical Report DI-FCUL-TR-2015-05, University of Lisbon.

de Marneffe, M.-C., T. Dozat, N. Silveira, K. Haverinen, F. Ginter, J. Nivre, and C. Manning. 2014. Universal Stanford dependencies: A cross-linguistic typology. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*, 4585–4592.

Silva, J. 2014. *Robust handling of out-of-vocabulary words in deep language processing*. PhD thesis, University of Lisbon.



# Resource | **CINTIL-DependencyBank**

Written by **João Silva**

CINTIL-DependencyBank is a corpus of Portuguese utterances annotated with the representation of grammatical dependency relations,[3] a kind of linguistic information that, roughly speaking, captures the fact that for a sentence to be grammatical the occurrence and position of words depends on, and is constrained by, the occurrence and position of other words in the sentence. The annotation is represented in a machine-readable tabular format, as was depicted in Figure 3.

Such annotated corpora are important resources for the study of natural languages and for the development of natural language processing tools. In the former, they support, for instance, concordancing and the search for syntactic patterns in corpora, which are necessary to check whether the theory fits the observed data; while in the latter, they are used, for instance, as training and evaluation data in the development of machine learning parsers (such as LX-DepParser, also presented in this Volume).

The developmental process of CINTIL-DependencyBank is worth noting, as it sets it apart from other dependency corpora. Generally, the manual annotation of corpora is a very time-consuming process that requires expert knowledge and, for large corpora, it is easy for errors and inconsistencies to occur. Because of this and because of a general lack of expert annotators, many corpora are automatically annotated and then manually corrected. While this can help, inconsistencies can still easily occur and the amount of effort required for correcting the annotation depends on the quality of the annotation tool.

In the NLX-Group, which is the Natural Language and Speech Group of the Faculty of Sciences of the University of Lisbon, we have developed LXGram, a symbolic deep processing grammar of Portuguese ("deep" in the sense that the analysis goes all the way to the semantic representation of meaning) under the HPSG framework (Pollard and Sag 1994), which we have used to support the annotation process. Figure 4 gives a striking impression of the amount of data and the complexity involved in a full deep analysis of what is a relatively simple sentence.

[3] https://hdl.handle.net/21.11129/0000-000B-D31C-8

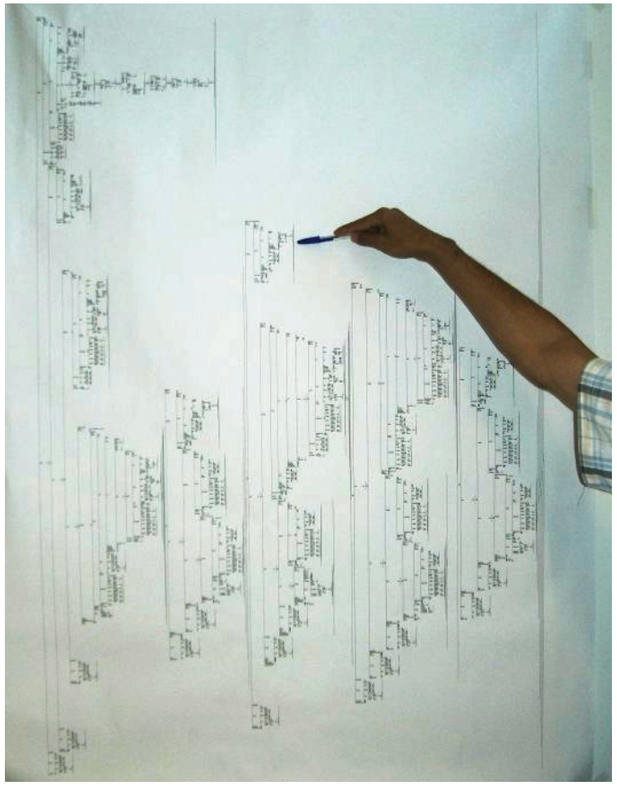


**Figure 4:** Printout in 6 pt font of the output of LXGram for the sentence *Todos os computadores têm um disco* ("All computers have a disk"). The hand holding a pen is provided to give a sense of scale.

To use LXGram for the annotation of corpora we rely on the fact that the grammar produces all grammatically valid analyses of a sentence, the parse forest. Then instead of having to correct the output of the grammar, the human annotator only has to disambiguate by picking which of the possible analyses in the parse forest is the valid one in that particular case. Note that the annotator does not need to scan all possible analyses one by one, which would be tiresome and error prone, as parse forests often contain many hundreds of analyses. Instead, the annotation platform that was used, [incr tsdb()] (Oepen 2001), automatically provides a set of discriminants (something like "does this constituent attach at point A or at point B?"), which are binary decisions that, when the annotator makes a choice, allow cutting the size of the forest in half. Through this process, a parse forest can be reduced to a single analysis with only a few discriminant choices, greatly speeding up the process. This is done by two annotators, following a process of double-blind annotation with adjudication of disagreements by a third annotator, thereby greatly reducing errors and improving the consistency of the annotation.

From the rich deep semantic representation (DeepBank) produced by LXGram we have extracted sub-representations, which we term "vistas" (Silva and Branco 2012). We have thereby extracted not only the DependencyBank presented here, but also a TreeBank (with constituency trees), a PropBank (with semantic roles) and a LogicalFormBank (with semantic representations); importantly, all these vistas are consistent and aligned among themselves, a major advantage in studying, for instance, the correspondence between constituency and dependency analyses. Figure 5 shows an example of a sentence whose constituency and dependency analyses have both been extracted from the same deep analysis provided by LXGram.

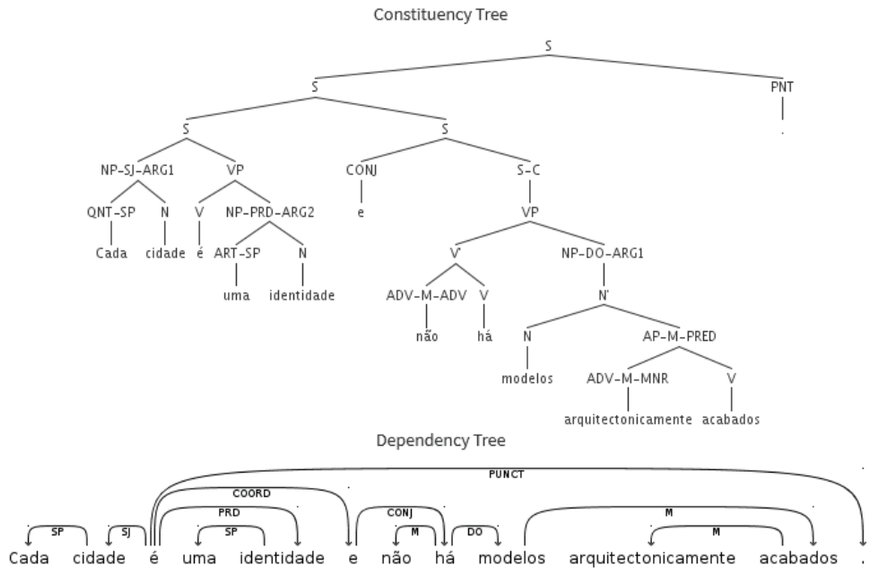



**Figure 5:** A view of a sentence in the CINTIL corpus ("Each city is an identity and there are no architecturally finished models"), showing the constituency and dependency vistas.

**References:**

Oepen, S. 2001. [incr tsdb()] – competence and performance laboratory. User manual. Technical report. Saarland University: Saarbrücken.

Pollard, C., and I. Sag, 1994. *Head-driven phrase structure grammar.* Chicago University Press and CSLI Publications: Stanford.

Silva, J., and A. Branco. 2012. Deep, consistent and also useful: extracting vistas from deep corpora for shallower tasks. In *Proceedings of the Workshop on Advanced Treebanking at the 8th International Conference on Language Resources and Evaluation (LREC'12)*, 45–52.

# Event | **Master Class at NOVA FCSH**

Written by **João Silva**

On 16 April 2021, João Ricardo Silva of PORTULAN CLARIN held an online master class for PhD students in Linguistics from NOVA FCSH, the School of Social Sciences and Humanities of the NOVA University of Lisbon, with the purpose of presenting PORTULAN CLARIN and its services, in particular the language processing tools it makes available for language research in the online Workbench. Roughly 30 students and a few professors were in attendance.

The class began with a presentation introducing CLARIN and PORTULAN CLARIN, followed by a live demo of the several language processing tools PORTULAN CLARIN has accessible in its Workbench. The presentation of CLARIN and PORTULAN CLARIN took about 15–20 minutes and served to provide some context to the students, as none of them were familiar with the infrastructure. The live demo that followed took about 90 minutes and walked the audience through the main parts of the site of PORTULAN CLARIN, with the major focus being on the Workbench. The online interface of almost every tool was shown, and a few examples were run to exemplify the sort of output produced by the tool. An example was already shown in Figure 2, as produced with the LX-DepParser. Figure 6 shows another example, this time produced with LX-TimeAnalyzer, a tool that, given a text, identifies all events mentioned in it and finds relations between them (given two events A and B, the possibilities are for A to precede B, A to follow B, A to overlap with B, or for no relation to be established).

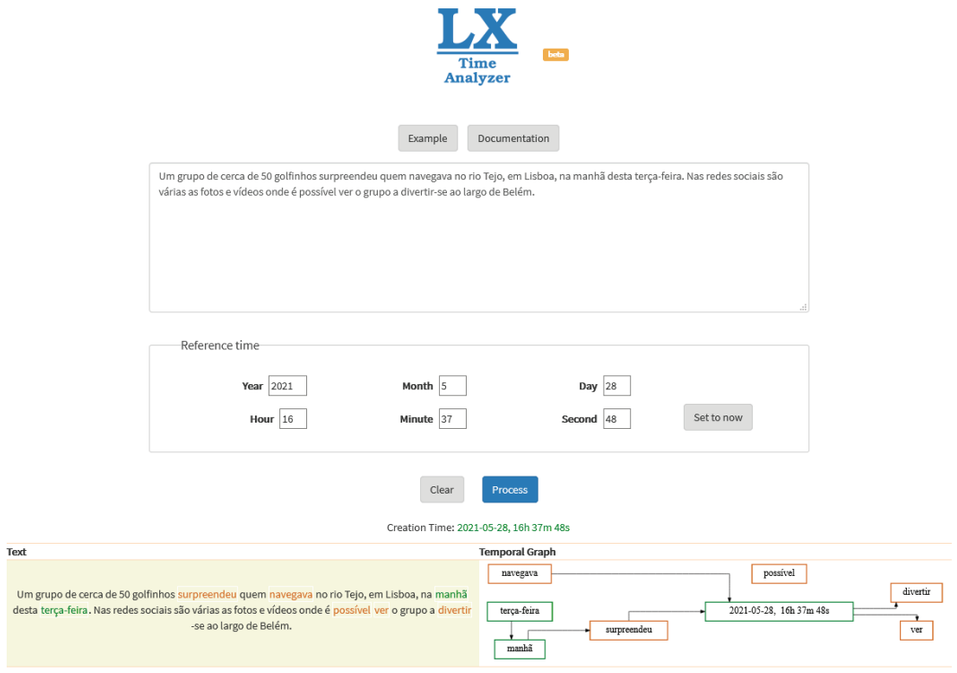



**Figure 6:** The LX-TimeAnalyzer service running in-browser.

At the end of the class there was time for the students to pose questions about the tools. For example, a student asked whether the tools could be applied to languages other than Portuguese. As the tools are language-specific, the student was directed towards the main CLARIN site and its Virtual Language Observatory where relevant results might be found. Another student wanted to know more regarding the CINTIL Concordancer service, namely about which corpora were searchable. As the name of the service implies, currently only the CINTIL corpus is indexed, but this question has motivated us to work towards the inclusion in the Workbench of a concordancing service that can work on a corpus submitted by the user.

# Interview | **Pilar Barbosa**

Pilar Barbosa is an associate professor of general and Portuguese linguistics who has benefitted from PORTULAN CLARIN tools in the annotation of a spoken-language corpus that she used in her syntactic research.

**Please introduce yourself – your academic background and current position.**

<

I am an Associate Professor of General and Portuguese Linguistics at the Department of Portuguese and Lusophone Studies, Universidade do Minho. I hold a PhD in theoretical linguistics from the Department of Linguistics and Philosophy at the Massachusetts Institute of Technology (MIT). My 1995 thesis, entitled *Null Subjects*, was written under the supervision of Noam Chomsky and Alec Marantz. My research interests are formal syntax, the interfaces between syntax and morpho-phonology and semantics, comparative syntax, experimental syntax, language variation and Portuguese and Romance linguistics. I am currently Director of the master's program in Linguistics at the University of Minho and coordinator of the research group Theoretical and Experimental Linguistics at the Centre for the Humanities at UMinho (CEHUM).

>

**With Cristina Flores, you are a part of the so-called "network of implementation partners" in the context of PORTULAN CLARIN. How did you get involved in the network? What are the goals of your involvement (e.g., what will you contribute to PORTULAN and how will you benefit from PORTULAN)?**

<

I got involved through Professor António Horta Branco, whom I've known since 2000. In 2003, I participated in the project GramaXing, which he coordinated, and then later he invited CEHUM to join the PORTULAN CLARIN network. CEHUM is a Humanities research centre and there are several ways in which the PORTULAN CLARIN tools can be helpful. As a linguist, I am particularly interested in spoken corpora, but as part of the projects developed by CEHUM researchers there are a number of literature and theatre digital databases that will certainly benefit from the tools developed by the PORTULAN CLARIN network.

>

**Which PORTULAN CLARIN tools have you used to annotate your corpora? Could you describe the annotation process? Which features made these tools especially valuable for your purposes?**

<

I have used the LX-tagger, developed by the NLX-Natural Language and Speech Group, which is the coordinating centre of PORTULAN and is led by Professor António Horta Branco. I had the transcriptions of the spoken corpus *Sociolinguists Profile of the Speech of Braga* in the EXMARaLDA format and I wanted the text to be annotated so as to facilitate automatic searches for particular syntactic constructions. The NLX team ran the LX-tagger and then we hired two (half-time) research assistants, who manually verified the part-of-speech annotation, here at the Universidade do Minho. Now the annotated corpus is available for public use in the repository of PORTULAN CLARIN.[4]

>

**Could you briefly present these corpora? For what kind of research are they (or will they be) intended?**

<

The *Sociolinguists Profile of the Speech of Braga* is a Portuguese speech corpus with 80 hours of recorded spontaneous speech, aligned with its transcription in the EXMARaLDA format.
It is composed of one-hour interviews with speakers from Braga, Portugal, randomly selected

[4] https://hdl.handle.net/21.11129/0000-000D-F935-F

and stratified according to sex, age and level of education. Thus constructed, the corpus is representative of contemporary Portuguese spoken in the city of Braga and allows for the study of variation and change in European Portuguese.

>

**Your research primarily involves theoretical syntax, which does not often involve the exploration of linguistically annotated corpora. How do you think your field in general can benefit from syntactically parsed corpora?**

<

In Principles and Parameters theory (Chomsky 1986), intra-linguistic and cross-linguistic variation are conceived in the same way, i.e., in terms of a differentiated application of parameters (pre)determined by universal principles. In more recent developments of this research program, such as the Minimalist Program (Chomsky 1995 and subsequent work), the parameters are in the functional lexicon, more particularly, in the feature content of the functional inventory of the language. This framework has been explored in the study of macro-parametric variation and micro-parametric variation within a range of different languages, including Portuguese. However, in order to study intra-linguistic variation and identify ongoing processes of language change, we need to carry out quantitative analyses and these require the collection of data samples that may be considered representative of a linguistics community. This is why I have become interested in corpus data. Annotation enables faster searches for particular constructions.

>

**Have you used such corpora or any other tools (developed by PORTULAN) in your syntactic research? Could you briefly present the main results?**

<

The results of our research on the corpus *Sociolinguists Profile of the Speech of Braga* have been published in a John Benjamins volume, entitled *Studies on Variation in Portuguese.*[5] I have contributed two papers. In one paper, written in collaboration with Cristina Flores and Ana Bastos-Gee, we studied a particular case of variable syncretism found in the region of Minho in Portugal, involving the 1st and 3rd person singular forms of "strong" preterites. In the speech of some speakers, these forms can be levelled and levelling can be obtained in two ways, by shifting the 3rd person to 1st person (as in the sentence *Não sei se* ***ele.3sg fiz.1sg*** *aquilo...*"I don't know if he did that ..." where the preterite verb *fiz* "did" has first person singular features despite the third person subject pronoun *ela* "she") or, alternatively, by shifting the 1st person to the 3rd (as in *E então que* ***fez.3sg eu.1sg****?* "And then what did I do?" where the preterite verb *fez* "did" has third person singular features whereas the pronoun subject *eu* "I" has first person features).

Our statistical analysis of the corpus data identified education level among the predictors for levelling. In addition, we discovered that there is a consistent use of a given form per verb within the speech of the same speaker, which led us to the conclusion that variation is not random. In particular, there is inter-individual variation in the choice of the form used for paradigm levelling. Since each individual speaker alternates between the use of the standard form and syncretism, there are two different kinds of variation: intra-and inter-individual. We developed an account of these paradigm levelling effects that is based on the interaction between the internal syntax of strong preterites and the Late Insertion of underspecified functional Vocabulary Items, as proposed in the framework of Distributed Morphology (Halle and Marantz 1993). We proposed a derivation of the different forms in the standard dialect and then offered an analysis of levelling where intra-speaker variation is tied to the probabilistic application of feature-deleting Impoverishment operations along the lines of Nevins and Parrott (2010). Inter-speaker variation is attributed to different choices as to which feature sets are subject to Impoverishment: the features for Person or Tense. This paper is a good example of how corpus data can be used to inform formal theories of morphosyntax.

The other paper, written in collaboration with Maria da Conceição de Paiva and Kellen Cozine Martins, focused on clitic climbing, that is, structures in which clitic pronouns can (optionally) be attached to the highest verb. To briefly illustrate this phenomenon, let's compare the placement of the clitic *se* (in bold) in the corpus example

> *Passa essa ponte, vai deparar-****se*** *com outra via rápida e os sinais.*
>
> "You go past that bridge and then you will come across another highway and the signs."

with its placement in the example

> *Ora, o autocarro sai de paragem, vai-****se*** *deparar com uma rotunda…*
>
> "Well, the bus leaves the bus stop, it will come across a roundabout…"

[5] https://benjamins.com/catalog/ihll.14

Both sentences contain the finite auxiliary *vai* "will" complemented by the infinitival verb *deparar* "come across". In the first example, the clitic is attached to the lower infinitival verb, thus forming the complex head *deparar-se*. In the second, it is attached in a higher position, namely to the finite auxiliary, forming the complex head *vai-se*. It is this latter phenomenon that is called clitic climbing, the idea being that in such "climbing" structures the clitic adjoins to the finite verb by moving out of the lower position next to the infinitival verb, in which it was originally inserted as a thematic argument of the infinitival.

According to previous work on the topic (Magro 2005), clitic climbing is more productive in the northern varieties of European Portuguese. In this paper we discuss the findings of a comparative corpus analysis of Braga and Lisbon oral speech, and conclude that this is not the case. Our main claim is that clitic climbing is a case of stable variation in both varieties. By means of a multivariate analysis, we show that clitic climbing is more frequent than attachment of the clitic to the infinitive in the two varieties. Moreover, we discuss evidence in favour of the claim that this syntactic variation presents the same configuration in both varieties: it is lexically constrained (as shown by the fact that not all verbs that take infinitival complements allow for clitic climbing) and not socially marked. These results indicate that the phenomenon of clitic climbing is a stable property of the grammar of European Portuguese and should be studied as such.

>

**What are your hopes for PORTULAN in the near future (e.g., what can PORTULAN do to help your research community)?**

<

**I believe that PORTULAN CLARIN will be of great help to linguists interested in modelling variation and change in Portuguese. Hopefully, corpora from other contemporary varieties of Portuguese will be included as well as texts covering the diachrony of the language. But scholars in other fields will also certainly benefit from what PORTULAN CLARIN has to offer to research in the Humanities.**

>

**References:**


Barbosa, P., C. Paiva, and C. Rodrigues. 2017. *Studies on Variation in Portuguese*. Amsterdam/Philadelphia: John Benjamins.

Barbosa, P., C. Paiva, and K. Martins. 2017. Clitic Climbing in the speech of Braga and Lisbon. In *Studies on Variation in Portuguese*, edited by P. Barbosa et al., 200–217. Amsterdam/Philadelphia: John Benjamins.

Barbosa, P., A. Bastos-Gee, and C. Flores. 2017. Variable strong preterites in European Portuguese. A sociolinguistic and theoretical approach. In *Studies on Variation in Portuguese*, edited by Barbosa et al., 154–175. Amsterdam/Philadelphia: John Benjamins.

Chomsky, N. 1986. *Knowledge of Language*. New York, NY: Praeger.

Chomsky, N. 1995. *The Minimalist Program. Cambridge*, MA: The MIT Press.

Halle, M., and A. Marantz, A. 1993. Distributed morphology and the pieces of inflection. In *The view from building 20: Essays in linguistics in honor of Sylvain Bromberger*, edited by K. Hale and S.J. Keyser, 111–176. Cambridge, MA: The MIT Press.

Magro, C. 2005. Introdutores de orações infinitivas: O que diz a Sintaxe dos clíticos. In *Actas do XX Encontro Nacional da Associação Portuguesa de Linguística*, edited by I. Duarte and I. Leiria, 649–664. Lisboa: Associação Portuguesa de Linguística.

Nevins, A. and J. K. Parrott 2010. Variable rules meet impoverishment theory: Patterns of agreement levelling in English varieties. *Lingua* 120: 1135–1159.

## K-Centres featured in this volume:

SAFMORIL, the Knowledge Centre for Systems and Frameworks for Morphologically Rich Languages

CORLI, the French Knowledge Centre for Corpora, Languages, and Interaction

CLASSLA, the Knowledge Centre for South Slavic languages

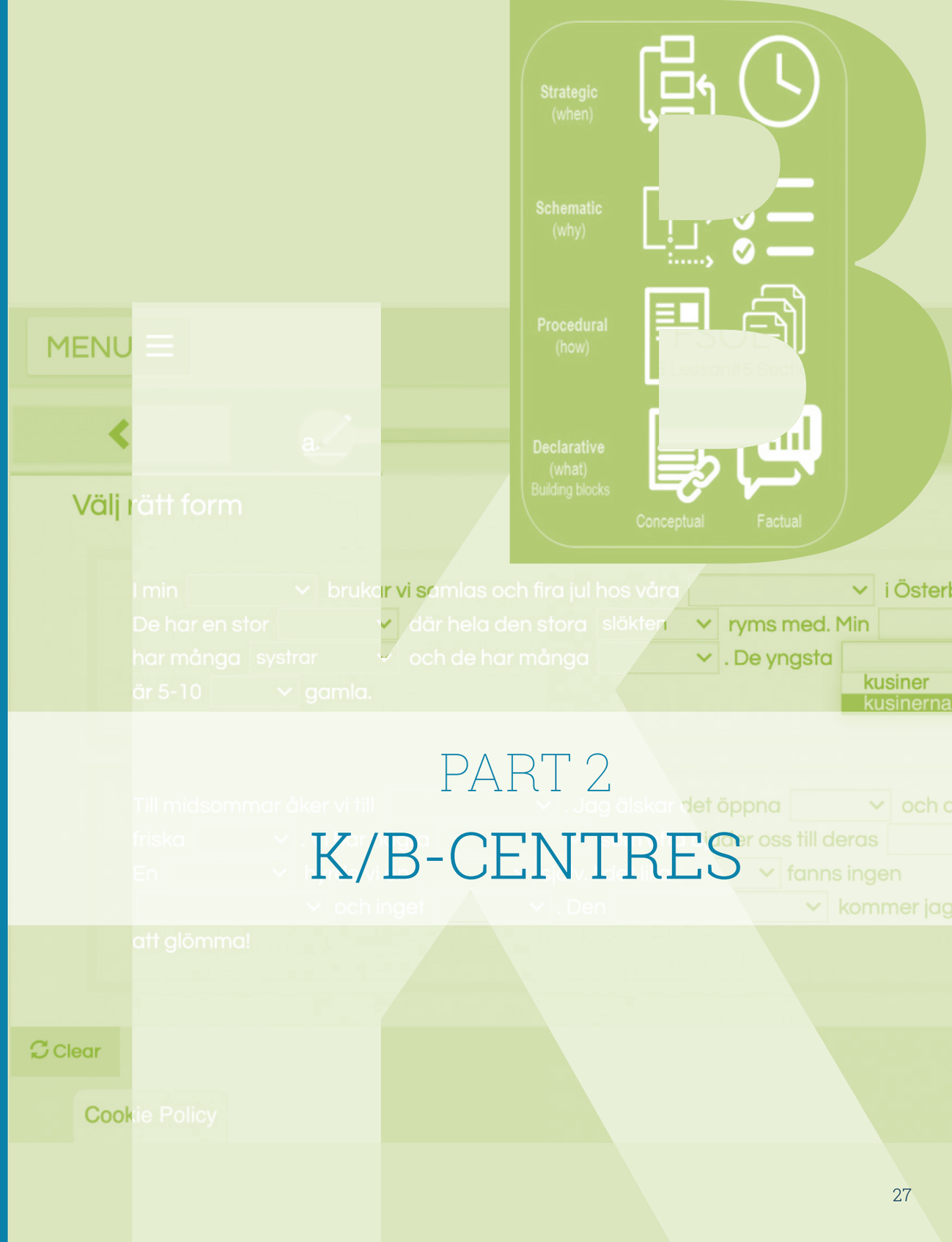
NLP:EL, the Knowledge Centre for Greek

## B-Centres featured in this volume:

ARCHE, the Austrian B-Centre for Digital Humanities and Cultural Heritage

The CLARIN-PL B-Centre






# PART 2
# K/B-CENTRES

# SAFMORIL, the Knowledge Centre for Systems and Frameworks for Morphologically Rich Languages

## Introduction

Written by **Erik Axelson, Sjur Moshagen, Jurgita Vaičenonienė, Inguna Skadina, Therese Lindström Tiedemann,** and **Krister Lindén**

SAFMORIL was officially recognized as a K-centre by CLARIN on 19 June 2019.[6] The K-centre operates as a distributed virtual centre supported by the following CLARIN member institutions:

- FIN-CLARIN: University of Helsinki and CSC, Finland
- CLARINO: University of Tromsø, Norway
- CLARIN-LV: Institute of Mathematics and Computer Science, University of Latvia
- CLARIN-LT: Vytautas Magnus University, Lithuania

SAFMORIL brings together linguists as well as researchers and developers in the area of computational morphology and its application during language processing. The focus of SAFMORIL is on actual, working systems and frameworks based on linguistic principles and on providing linguistically motivated analyses and/or generation on the basis of linguistic categories. Such systems are relevant in particular for languages with rich morphologies, as is the case of Nordic and Baltic languages (such as Finnish, Swedish, Norwegian, Latvian, Lithuanian as well as the Sámi languages) and more generally Fenno-Ugric languages, Inuit languages, Canadian First Nation languages and Babylonian languages.

[6] https://www.kielipankki.fi/safmoril/

SAFMORIL offers online courses for developing and teaching morphologies, tokenizers and spell-checkers, a repository for storing morphologies, and an environment for creating tokenizers and spell-checkers. SAFMORIL serves linguists and computational linguists developing and adapting morphologies as well as digital humanities scholars, linguists, and computer scientists processing language data. Researchers are welcome to get in touch with SAFMORIL regarding any matters related to morphology (computational or otherwise) via the SAFMORIL Helpdesk (*safmoril@kielipankki.fi*).

The four member institutions of SAFMORIL – that is, FIN-CLARIN, CLARINO, CLARIN-LV, and CLARIN-LT – each offers its own unique technologies and services for working with morphology. The Finnish member FIN-CLARIN focuses on creating novel morphology systems and frameworks. The two main tools that FIN-CLARIN contributes as a member of SAFMORIL are Mylly, which is used for analysing and visualizing data sets, and HFST – Helsinki Finite-State Technology, which is a compilation and runtime software, with some source morphologies.


**Figure 7:** An exercise from Finland Swedish Online which focuses on morphology.

FIN-CLARIN offers online tutorials for XFST-based Morphology Development (provided by Erik Axelson, Kimmo Koskenniemi and Mathias Creutz at the University of Helsinki) and Morphology Construction (developed by Jack Rueter at FIN-CLARIN) as well as documentation for experimental two-level rule compilation using Python HFST (provided by Kimmo Koskenniemi at FIN-CLARIN). Lastly, FIN-CLARIN offers Finland Swedish Online, which is a free online course in Swedish as spoken in Finland. It is designed based on the model of the Icelandic Online course and includes a variety of texts, videos, sounds clips and exercises to help you learn Swedish. Morphology is practised implicitly through reading and listening to Swedish, where we take care to repeat forms and patterns that are being practiced, and it can be practiced in self-correcting exercises. Finland Swedish Online currently consists of two courses, but a third course is soon to be launched and soon there will also be a special course designed for librarians.

CLARINO contributes to SAFMORIL mainly via the Arctic University of Norway (UiT), which is one of the institutions comprising the Norwegian CLARIN consortium. UiT offers the GiellaLT infrastructure, which hosts language resources and tools for more than 140 different languages in more than 180 repositories. The infrastructure includes a development environment for morphologies and morphology-based tools, morphology teaching service GiellaLT ICALL, and offers tutorials for making computer tools for your language. In addition, the aforementioned HFST (Helsinki Finite-State Technology) toolkit has been applied extensively in the GiellaLT infrastructure, and is also a core part of the proofing tools provided by it.

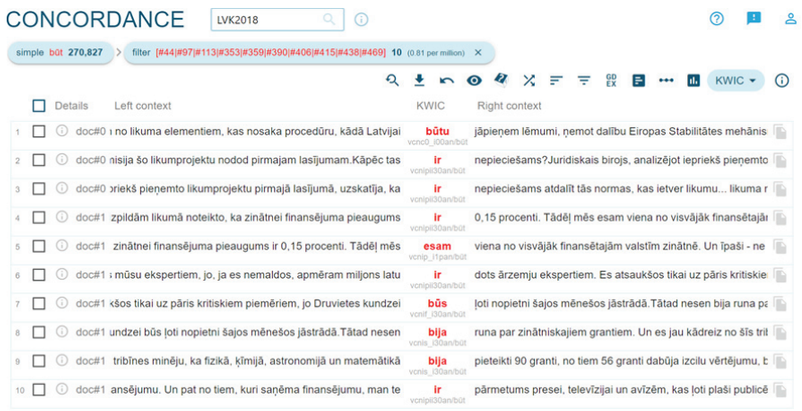



**Figure 8:** Find forms of the verb *būt* ("to be") together with its morphological annotation in the LVK 2018 corpus (http://nosketch.korpuss.lv/#dashboard?corpname=LVK2018).



As part of SAFMORIL, CLARIN-LV aims to provide support not only for Latvian as a morphologically rich language, but also for other morphologically rich languages spoken and researched in Latvia (e.g., Latgalian). The CLARIN-LV repository includes not only LKV2018, a morphologically annotated 10-million-word corpus of modern Latvian; the Saeima corpus of parliamentary proceedings; and Senie, a 900-word-corpus of historical Latvian texts from the 16th to 18th centuries; but also the Latgalian language corpus MuLa. To support users of digital resources for the Latvian language, CLARIN-LV organizes practical workshops and hands-on sessions on different topics (e.g., on regular expressions, morphological annotation and how to search in syntactically annotated corpora). All materials from seminars are available on the CLARIN-LV website. These materials are actively used in different courses at the University of Latvia and Liepāja University, as well as at Digital Humanities summer schools.

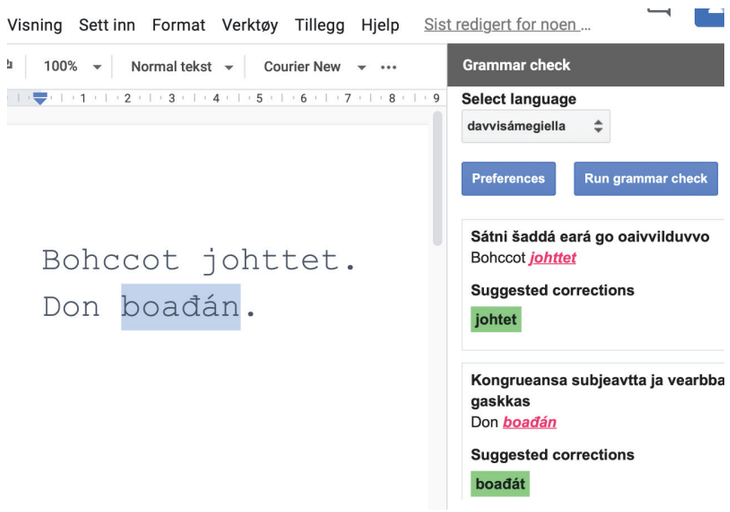



**Figure 9:** A screenshot of the North Sámi grammar checker, highlighting a congruence error and a correction suggestion.

In 2020, another morphologically rich language, Lithuanian, was included in SAFMORIL. Although the CLARIN-LT consortium already had a Helpdesk on corpus linguistics and natural language processing methods for Lithuanian, the team of Lithuanian researchers was very glad to expand their knowledge sharing with regards to the Lithuanian morphology, syntax, semantics and tools for linguistic analysis (e.g., those produced within the project SEMANTIKA-2) with an international audience. As a member of SAFMORIL, the CLARIN-LT team looks forward to exchanging experiences, opening new opportunities for cooperation, and the further development of resources and tools relevant for the analysis of morphologically rich languages.

# Interview | **Jack Rueter**

Jack Rueter is involved with the SAFMORIL Knowledge Centre. He is a computational linguist whose work primarily focuses on finite-state descriptions of highly endangered languages with complex morphological systems.

**Please introduce yourself – your academic background and current position. What inspired you to become a computational linguist?**

<

My name is Jack Rueter, and I have a doctorate in General Linguistics from the University of Helsinki. My earlier studies had helped me to specialize in Uralic languages, and, in fact, I have been able to apply this acquired knowledge analogically to other languages of the world, including indigenous languages of the Americas. Over the past decade I have worked as a university researcher in Digital Humanities, which to a great extent has involved finite-state descriptions of highly endangered languages with complex morphological systems, over twenty languages in all – Komi-Zyrian (1996–), Erzya (1999–), Olonets-Karelian, Livonian, Moksha, Hill Mari, Tundra Nenets (2013–), Skolt Saami (2014), and so on. This work has been made possible through collaboration with the Giellatekno infrastructure at the Norwegian Arctic University in Tromsø, which hosts well over 150 languages, and where much of the Helsinki Finite-State Technology (HFST) has been used in practice and of course funding from the Kone Foundation as well as FIN-CLARIN itself, where the technology has been developed.

Since the time I was a little boy, I was fascinated by foreign languages. The University of Helsinki provided me with the environment and possibility to pursue language learning. It was not until the 1990s, however, that I was introduced to the possibilities of describing languages in a way that could also be used for their facilitation. That was when I made my first description of the Komi-Zyrian language. It all started with creating a lexicon and adding glosses in two languages for a university language class. Subsequently, I introduced regular inflection of nouns and verbs utilizing the two-level model concept behind HFST with the mentorship of Kimmo Koskenniemi.

When I travelled to the Komi Republic and the Republic of Mordovia in the second half of the 1990s, I was encouraged by Pirkko Suihkonen to gather text corpora for the University of Helsinki Language Corpus Server (UHLCS) at the University of Helsinki, which predates FIN-CLARIN, and is now one of the many elements available through FIN-CLARIN and the Language Bank of Finland. Collecting language corpora naturally required gathering releases from the individual authors and publishers. Learning to speak the Komi and Erzya (one of two Mordvin languages) languages made it easier to negotiate work with language corpora. It even helped me to acquire and develop connections with speakers of Moksha (the other Mordvin language) and the Udmurt languages (a close relative of Komi).

It was this work with finite-state description of languages and collecting corpora that brought me to my dissertation on a portion of Erzya regular morphology and strengthened my close relation with HFST and the Giellatekno infrastructure. Regular morphology is where each incremental part of morphology is directly related to one or more increments of semantics, and the meaning of the resulting word form can be deduced from its morphological structure. While regular morphology in the English language might be equated to four regular forms in verbs and four in nouns, e.g. *to talk* to *talk*, *talks*, *talked* and *talking* and *cat* to *cat*, *cat's (name)*, *cats* and *cats (names)*, regular morphology in Erzya might easily take us to figures of over 200 for nouns and verbs alike.

It can be said that my work with HFST tools has contributed greatly to the multilingual facilitation of minority Uralic languages, and projects directly associated include mini-paradigms and derivation information for a new Finnish-Skolt Saami dictionary, the FST behind the Python libraries in UralicNLP, online dictionaries, spell-checkers and keyboards from GiellaLT, lemmatization for corpora at the Language Bank of Finland and the University of Turku, not to mention integration work with regard to HFST in ELAN for Komi language forms and similar work with other languages supervised by Niko Partanen.

>

**What is your current role at SAFMORIL? How did you get involved?**

<

I am currently continuing the development of morphological descriptions for Finno-Ugric languages, as well as acquiring and developing additional resources for research in these languages in general. This involves acquiring new corpus data, developing Jupyter courses for morphological descriptions using HFST, and has lately also involved efforts to make transliterated speech data available. The ultimate goal should of course be to capture the spoken varieties of a language as variations of the HFST descriptions.

>

**Could you briefly present HFST? What does it do? Why is it important for computational linguistics?**

<

The Helsinki Finite-State Transducer (HFST) toolkit is intended for processing natural language morphologies. The toolkit is demonstrated by wide-coverage implementations of a number of languages of varying morphological complexity. HFST can be run via command line tools or a Python API.

HFST can be used for compiling various formalisms into finite-state transducers and then combining these transducers with lexicons. In other words, linguistic processes are described with a set of rules, combined with the lexicon, and the result is then applied to a text stream on the token, morpheme, word, and sentence levels. This approach can be used, for example, for morphological analysis and generation, tokenizing and tagging text, fast look-up of strings in a transducer, and spell-checking and matching/transformation with a Recursive Transition Network system.

From a linguistic perspective, a finite-state description of morphology involves a closed set of headwords or lemmas that are associated with their individual stems and finite sets of affixes (prefixes, infixes or suffixes) to form inflectional paradigms, where regular semantic meaning and morphology are conjoined. The structures of these descriptions predominantly follow canonical synchronic and diachronic treatises of an individual language, and are therefore open to extensive variation.

Since there might arguably be an infinite number of finite-state descriptions for any given language, it is important that a responsible infrastructure be maintained, such as Giellatekno, where testing and development practices with applied HFST tools guide the language modellers to follow and conform to mutual descriptive notations and structuring. At Giellatekno, where HFST tools have been applied most extensively, Northern Saami is the language that has the longest history of development, hence the most extensive set of applications for language research and facilitation of the Finno-Ugric languages. Among these applications are morphological analysis and generation at the word level and contextual disambiguation at the sentence level.

Word-level analysis contributes to morphologically savvy dictionaries. It allows you to find the meaning of a word without knowledge of the headword or even language-specific alphabetical order. You just have to input the form you have found or click on it in a text. When you have a descriptive analyser for the linguist to delve into standard literary, archaic, dialectic and even regular formations of nonce words, you also have the makings of orthographic as well as basic learning tools. Here is where context-based disambiguation comes into play.

Sentence-level disambiguation, also a target of continued research, is what addresses the multi-ambiguity generated by robust regular descriptions of the language morphology. With contextual disambiguation implemented, work with spellers and intelligent, computer-assisted language learning tools can be extended. Disambiguated morphological readings for all words of a sentence mean we can hypothesize text-to-speech readings, work with machine translation, and even introduce user group-specific spell-checkers with correction suggestions specific to the individual context. In a similar vein, context awareness means that language learning can introduce contextual morphological exercises and even chatbots.

In short, a sharing platform of applied HFST tools can and does provide support for language researchers and facilitators in tandem with their language communities.

>

**What makes Finno-Ugric languages particularly difficult for automatic morphological analysis? How does HFST overcome this?**

<

A majority of the Finno-Ugric languages have few annotated corpora, and few or no embeddings. These languages are low-resourced, unlike Hindi or German; they actually have few or no natural language processing resources. There are, however, limited but feasible resources such as lexical, morphological and syntactic descriptions that have been collected in fieldwork and research over the past two centuries. These have yet to be applied to the facilitation of the languages by their relevant communities.

One hidden asset of HFST is that we actually attempt to utilize previous research results and outcomes. This may mean copying word lists with their research glosses as notes and introducing research paradigms as testing materials. Actually, these can be seen as the first steps of a workflow for working with finite-state descriptions. Here even limited research materials for a language can provide for an extensible approach to language facilitation.

We have noted that Finno-Ugric languages are extremely rich morphologically, which is different from many languages often studied in Natural Language Processing. Using finite-state technology for both a challenging morphology and a simple one alike allows for detailed descriptions of these phenomena while they are being modelled. It also allows the modeller a choice: to avoid or not to avoid any problems ensuing from rare forms that might not be found when creating a model based on annotated corpora alone.

>

**What is the importance of applying HFST in other projects focusing on morphological analysis?**

<

HFST is a very useful component in linguistic research, as we customarily need to analyse texts that have not been annotated at all. We want to be able to select our research data on the basis of what is actually useful for our questions and not be limited by what has been annotated. For this reason, technology to annotate materials automatically is very important, and HFST provides means for doing this.
Since developing workflows with HFST might not initially be obvious to all, it is important that we provide an example. When we annotate our materials with HFST, we often find ourselves in a loop where the technology needs to be continuously improved and texts reannotated. In the end, we may also need to do some corrections manually. Hence, how to document and repeat these types of research workflows is also an important question to solve.

This is especially so in projects where HFST is integrated into tools linguists are already using, such as Niko Partanen's integration work with HFST and the transcription tool ELAN. Solutions for tasks such as this have already been developed, and these methods are currently in use at several universities.



**Of course, we see machine learning methods becoming more popular every day, but we think that a variety of approaches can enrich and optimize what we are trying to achieve. Eventually we will be able to train well-working neural taggers and parsers for various Uralic languages. But still, tools such as HFST have proven invaluable when we start creating new resources for a new language, and we believe that even in the future rule-based methods that provide for precise verifiable results and analyses will remain relevant.**

>

**Could you describe your work together with Niko Partanen on Finno-Ugric languages using HFST? Which resources did you build and what is their intended application, both in NLP/computational linguistics and beyond (i.e., wider digital humanities and social sciences)?**

<

As I mentioned, many Finno-Ugric languages do not have very large computational resources. This is of course something we aim to change. One of the larger projects we have been involved in is Universal Dependencies (Czech CLARIN), where we create annotated treebanks in different languages of the world, using essentially the same annotation scheme. This, we hope, allows further comparative work. A treebank simply means that materials also have annotations at a syntactic level, following dependency grammar, where each sentence has the central root element, usually a finite verb, and then other constituents of the sentence are connected to it as leaves.

These materials will be useful not only in computational linguistics, but also for basic linguistic research. Naturally, to show that this can be done is partly our own responsibility, and we are continuously working with the description of various Uralic languages using these tools to prove exactly this.

HFST plays an important role here, because despite the dearth of corpus materials we can make robust descriptions, which as the linguist desires can be utilized or alternatively left out of the analyser. In work with UD, robust analysis allows us to introduce new extensively annotated corpora for morphologically complex languages which have not been derived from previous corpus projects. Many of the larger majority language projects are basically going through code conversion and annotation alignment, whereas we are starting from scratch. At present Niko and I have dealt extensively with Komi-Zyrian, where we have addressed both literary and dialect

language descriptions with one FST. Although the morphophonological description is relatively simplistic from a phonological alternation perspective, Komi does have an extensive combinatory morphology that may even operate on the syntactic level, where verbs can be derived from complex noun phrases, e.g. “the student put on a bright red shirt”.

>

**Aside from yourself, do you know whether HFST has been successfully used by other researchers?**

<

As I mentioned earlier, the Giellatekno infrastructure at the Norwegian Arctic University in Tromsø is a platform where the HFST tools have been applied quite extensively. Since the primary target languages for facilitation are the minority languages of Northern Norway, i.e. Saami languages and Kven, work with other Uralic languages falls into a different category, but by no means is this other category less facilitated. In fact, it introduces collaboration with the Võro Institute in southern Estonia, the Livonian Institute in Latvia, Mari language research in Vienna and the Mari El Republic as well as work with the Komi language in the Komi Republic, and Karelian research in Eastern Finland and Saami languages in Oulu as well as initial steps in work with Indigenous Amazon studies in Belém, Brazil.

Work with the linguistic description of the languages of Greenland also brings contributions and feedback to the development of HFST from the Institute of Language and Communication (ISK) at the University of Southern Denmark, in its work on Constraint Grammar disambiguation of the HFST analyser output.

Research at the Alberta Language Technology Laboratory, headed by Antti Arppe, applies HFST in its development of tools for several indigenous language families of Canada.

HFST has also found its way into the development of the open-source GATE DictLemmatizer developed in the Universities of Sheffield and Duisburg-Essen.

Further networking with FST technologies also draws our attention to the University of Latvia and “Introduction to NLP” courses taught there for computer science students.

>

**Together with Erik Axelson, you have been developing an online self-study tutorial for morphologically rich languages. Could you present this tutorial? What is its target audience?**

<

The tutorial “Morphologically Rich Languages with HFST” demonstrates how HFST tools can be used for generating finite-state morphologies for morphologically rich languages.[7] It is implemented as Python notebooks which use the HFST Python interface. The tutorial is hosted at CSC, the Finnish IT centre for science.

This web course is based on a course by Sjur Moshagen and myself, organized at the University of Helsinki and named “Language Technology for Finno-Ugric Languages – Methods, Tools and Applications”.[8] The course is part of the MA Programme “Linguistic Diversity in the Digital Age”.

This course gives an introduction to mainly rule-based language technology as used in many full-scale, production projects using the GiellaLT and Apertium infrastructures. The technologies and methodologies presented can be used on any language, although the focus is on morphologically complex ones.

The course assumes that the user knows the fundamentals of general linguistics and has basic knowledge of how to use a computer. Some programming experience is desirable and knowledge of NLP is also a plus.

At the moment, access requires a HAKA account. If you do not have a HAKA account, you can contact the SAFMORIL Helpdesk to arrange for local accounts or request a visitor account directly from the CSC service desk. You also need a join code that you can request from the SAFMORIL Helpdesk.

>

[7] https://blogs.helsinki.fi/language-technology/hi-nlp/morphology

[8] https://courses.helsinki.fi/en/lda-t3113/124901269

# CORLI, the French Knowledge Centre for Corpora, Languages, and Interaction

## Introduction

Written by **Eva Soroli**

For most institutions, sharing knowledge means sharing declarative information: sharing data, metadata, tools (the "building blocks" of knowledge) and giving access to repository centres with such resources. However, accessing knowledge is not only about "knowing-what" can be collected, accessed or analysed within a dataset, but also refers to "knowing-how" a specific dataset is built, how it can be processed, for which purposes should it be created, used/re-used ("knowing-why"), and following which steps/procedures ("knowing-when"). The French Knowledge Centre CORLI for Corpora, Languages and Interaction (henceforth CORLI K-centre) is such an attempt: a certified CLARIN K-centre established in July 2020 that has identified through consensus the main steps that need to be carried out in research involving language corpora along with the main principles that should characterize such investigations.[9] The CLARIN K-centre thus provides knowledge that researchers need in their everyday practice in the following four domains (Figure 10) covering both the *factual* aspects of necessary content elements (data, metadata, tools, repositories) and the *conceptual* parts of knowledge (typology of the data, principles of data collection/processing/storage/reuse, usage-based research methods, training, workflows, etc.).

[9] https://corli.huma-num.fr/en/kcentre/

**Declarative knowledge**: involves information about the most important data and metadata repositories; access to technical manuals and tools; external links to other relevant specialized institutions (e.g., B-, C-, and K-centres), etc.

**Procedural knowledge**: is about offering advice about proper citation of existing datasets; best practice recommendations for data collection; solutions and training for data processing, storage and research management; advice about transcription and annotation principles, data anonymization, metadata standardization, file conversion, etc.

**Schematic knowledge**: refers to providing examples of workflows; research questions and options of investigation/adequate types of corpora; guidelines on the establishment of sound research plans; help with the preparation of ethical approval applications; best practice recommendations for making resources Findable, Accessible, Interoperable, Reusable (FAIR), etc.

**Strategic knowledge:** is about providing information on the necessary steps of a research protocol; access to management checklists, recommendations about data collection unfolding; flow diagrams on data lifecycle, etc.

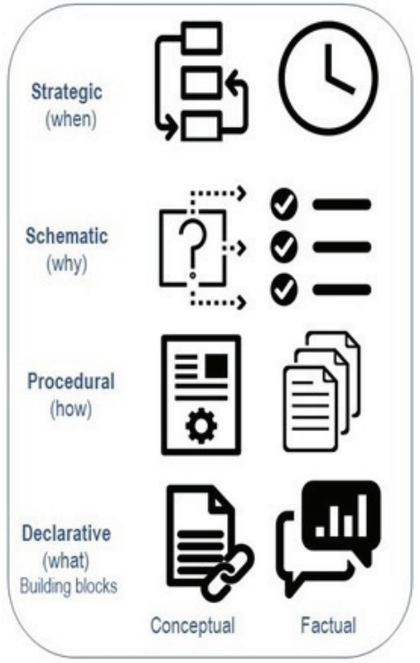



**Figure 10:** CORLI K-centre knowledge domains.

The CORLI K-centre is part of the CORLI consortium, a national consortium of more than twenty Universities and research labs involving more than two hundred researchers around France, and dedicated to consensus-based recommendations and digital solutions in corpus linguistics. The centre is coordinated by the French digital humanities infrastructure Huma-Num and functions as an interactive online platform which centralizes and provides cross-border access to knowledge through both proactive and reactive services.

**More specifically, with respect to proactive knowledge, the CORLI K-centre offers best practice recommendations for:**

1. building corpora and format conversions, accessing existing written and oral corpora repositories;
2. metadata standardization procedures, data storage, assessment and re-use principles;
3. legal and ethical issues related to corpus management and use, among others.

Users with more specific needs in linguistic analysis also have the possibility to access repositories with more specialized information, e.g. manuals for corpus annotation, practical guidelines for the use of corpus annotation and analysis tools, etc.

Some of the most popular actions of the CORLI K-centre include the training opportunities offered every year on the use of digital tools, as well as regular financial support calls for the finalization and transformation of existing corpora to ensure compliance with the FAIR principles. The development of a FAQ (frequently asked questions) page addressing common concerns in these topics as they occur in the questions formulated by the users, further contributes to information access. The users of the CORLI K-centre platform have the possibility to access most knowledge through the website of the centre, and alternatively through the FAQ, where other landing pages offer the possibility to redirect to related content (e.g., to ERIC, CLARIN, other B-, C- and K-centres, etc.) and thus continue the journey ideally without the need for outside assistance.

In cases of requests for further assistance, the CORLI K-centre offers an additional reactive knowledge-sharing service established thanks to a pool of researchers and data specialists who provide further information whenever needed. The way the users interact with the webpage and the provided knowledge is of vital importance to the CORLI K-centre, as these feeds help update the pages and information offered. For this reason, a contact form has been integrated to the platform (easily accessible on a separate page) enabling users who cannot find an adequate answer to their questions to contact the centre directly (see Figure 11 for an overview of the website).

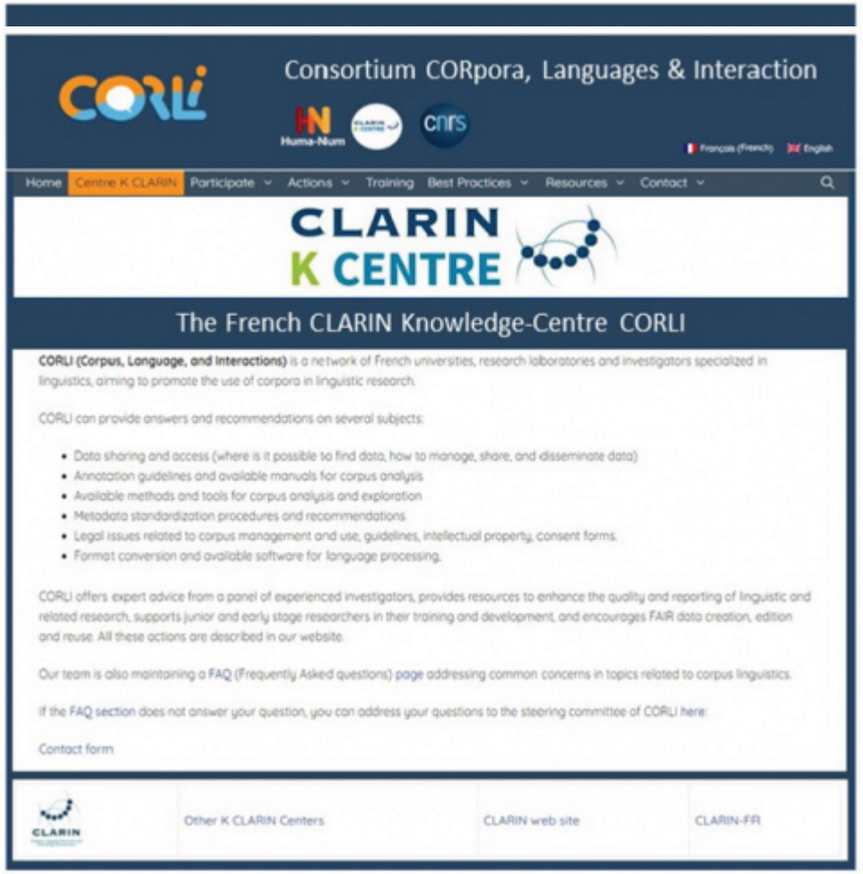



**Figure 11:** Overview of the CORLI K-centre website.

**In addition to this service-oriented line of work, the CORLI K-centre is also a research centre that promotes empirical approaches to language, organized in six working groups:**

1. The Interoperability, Queries and Corpus Exploration Team is focused on theoretical and methodological questions related to corpus exploration and annotation practices, interoperability and metadata standardization.
2. The Multimodality and New Forms of Communication Team is focused on multimodal ways of communication (gestures, posture, gaze, sign language, etc.) and online behavioural techniques of data collection (eye movements, 3D motion capture, EEG, etc.).
3. The Multilingualism Team works on bilingual and multilingual corpora with a special focus on written and oral corpora, language mixing in endangered languages and languages in contact, language acquisition and code-switching, as well as on parallel corpus-based investigations.
4. The Legal Issues and Data Protection Team works on the challenges researchers face in their everyday practice with data management, storage and anonymization (GDPR regulations, open science constraints, sensitive data management, international legal practices, etc.).
5. The Corpus Annotation Team focuses on best practices for the management of corpus annotation projects (project conception, establishment of multi-layer annotation manuals, inter-annotator agreement principles, development/use of annotation tools, etc.).
6. The Corpus Assessment Team works on the evaluation of language resources (data and tools) and the establishment of homogeneous criteria for resource storing, archiving, sharing and resource quality assessment according to the FAIR principles.



The expertise gathered by the pool of specialists involved in the above groups has led to a great number of outcomes that are useful to linguists (at all levels of academic expertise), but also to anyone working with corpora or interested in language use, databases, digital tools for data collection, resource exploration and data management (engineers, data scientists, educators, etc.).

**With expertise in corpus linguistics, multilingualism and multimodality, the CORLI K-centre aims:**

- to enhance multi-level knowledge and practice sharing following the FAIR and Open Science principles among researchers and other actors interested in language and corpus linguistics,
- to become a major platform of communication and a training portal for younger researchers,
- to facilitate the exchange of ideas and collaborations, and support international synergies.

# Interview | **Thomas Gaillat**

Thomas Gaillat is a lecturer in corpus linguistics and a teacher of English for specific purposes. He is part of the CORLI steering committee.

**Could you please introduce yourself – your background and current academic position?**

<

Today I work as a lecturer in corpus linguistics and as a teacher of English for Specific Purposes (ESP) at Rennes 2 University in France. These two roles illustrate my deep interest in understanding how people acquire a foreign language. I have been teaching English for 20 years at all levels of the curriculum, from children to university students. I have always wondered what could make a teaching method better, and research naturally caught my interest. I love the blend between the practitioner's experience and the distant, objective analysis of the researcher.

In 2004, I joined the University of Rennes 1 language department as an ESP teacher where I also acted as director for two years. This position gave me the opportunity to enter the academic world and understand the linguistic needs and interests of learners of English specialised in other disciplines. I got involved in the design of online course materials. I also took charge of a master's class in the design of online courses for language learning. Through these activities I felt the need to go further in the analysis of learner language. My initial experience in the computer industry, back in the nineties, pushed me towards computer methods applied to language learning. And so I embarked on the doctorate odyssey.

I received my PhD in 2016 at the University of Sorbonne Paris Cité, awarded *summa cum laude*. The thesis focused on corpus interoperability as a method to explore how *this*, *that* and *it*, as referential forms, are used by learners of English.



As my work intersected the domains of corpus linguistics and Natural Language Processing my thesis was supervised by Professors Pascale Sébillot (INSA Rennes), specialized in NLP, and Nicolas Ballier (Université de Paris), specialized in Linguistics.

In 2017, I was fortunate to be awarded a postdoctoral position at the Insight Data Centre NUI Galway, Ireland as part of a H2020 project (SSIX) focusing on Sentiment Analysis. As a linguist I joined Brian Davis's Knowledge Discovery Unit and contributed to the development of an AI-based pipeline dedicated to predicting sentiments in financial tweets. This work involved tight collaborations between statisticians, programmers and linguists. This first-hand experience showed the benefits of combining complementary skills for the design of online systems.

In 2019 I joined the Linguistics and Didactics research team of Rennes 2 University. I am now in a position to leverage my previous experience in order to contribute to the research on data analytics related to language learners.

>

**You represent your lab LIDILE in the CORLI steering committee: why have you decided to join CORLI and how does the centre support your research and that of your colleagues?**

<

**CORLI is a French organization that positions corpora in the centre of its activities, very much like our research team. My PhD relied on making several corpora interoperable. Corpora are the fuel for machine learning approaches. Unsupervised and supervised learning methods do not just rely on efficient algorithms but also on linguistic metadata that enrich the source data. In the context of policies fostering AI projects, this message needs to be conveyed and, to me, CORLI fulfils this role in France. I am happy to contribute to this.**

>

**What expertise do you and your laboratory bring to the K-centre?**

<

Our lab includes three main research domains, i.e. translation studies, linguistics and foreign language didactics. Most of our projects rely on the development of multilingual corpora. One of our focuses is learner corpora and their exploitation. We have expertise in the many ways

corpora can be used for i) language training purposes, ii) as sources in the training of future language teachers and iii) as sources for the design of ICALL systems. These are questions tackled by the corpus community.

We are also in the process of designing a database infrastructure supporting dynamic corpus querying. To this end, we are transferring some of our corpora to the Nakala repository,[10] which is a CLARIN-FR Huma-Num node. This architecture will rely on persistent data making corpus items queriable and retrievable. Nakala's APIs will give controlled access to corpus items. Through this experience, we have acquired knowledge regarding architectural design constraints. We have faced issues regarding the classification of corpus files related to specific individuals and their metadata such as L1, L2s or mode. The solutions we found might be of interest for other researchers in the corpus community.

>

**You've been involved in the creation of linguistic resources and the development of machine learning methods related to multilingual learner corpora. Could you introduce some that you feel are most noteworthy? What are their main features?**

<

That's right. Over the course of the last three years I've taken part in three projects based on learner corpora.

First, I was the Principal Investigator of a University of Rennes project which aimed at developing a tool to automatically extract and visualize linguistic profiles in texts written by learners of English. We showed how learner writings in English can be exploited with NLP tools to compute and visualize complexity metrics. An ORTOLANG corpus annotated in terms of the Common European Framework of Reference (CEFR) proficiency was used as a benchmark for comparisons with new learner writings. The system creates visualizations superimposing metric values and proficiency levels. This form of feedback aims at giving teachers and learners tools to give objective measurements of specific linguistic features.

Secondly, I played an active role in a PHC Ulysses Franco Irish programme between the universities of Paris-Diderot and Insight NUI Galway. The project, led by Nicolas Ballier (France) and Manel Zarrouk (Ireland), investigated the use of linguistic complexity metrics for the automatic detection of proficiency levels in English writing.

[10] https://www.nakala.fr/

We implemented a supervised learning approach as part of an Automatic Essay Scoring system. The objective was to uncover CEFR-criterial features in writings written by learners of English as a foreign language. The method relied on the concept of microsystems with features related to learner-specific linguistic systems in which several forms operate paradigmatically. Based on a dataset extracted from the EF-CAMDAT corpus, we trained a multinomial logistic regression model and an elastic net model. Evaluation was conducted on an internal test set as well as an external data set extracted from the ASAG corpus. The results showed that microsystems combined with other features help with proficiency prediction.

Finally, the Corpus InterLangue (CIL) is a resource that has long been developed in our team here in Rennes.[11] It was initiated more than ten years ago and now includes more than a hundred audio recordings and writings from learners of French or English. As explained previously, we are going to make this corpus available online. What's more is that it will be accompanied by a series of R scripts forming a pipeline allowing for data curation, data set creation, and data visualization. The idea is to have a modular approach for flexible data sets depending on the research questions.

>

**How are such resources and methods related to CORLI i.e. have they been involved in their development? If so, how?**

<

By nature, these resources are related to CORLI. The corpus which was developed for the first project is available on ORTOLANG and CORLI provides an inventory of all language-corpus resources available in French labs.

We also intend to make the CIL corpus referenced by CORLI via the *resources* page of the website. The R scripts will also be referenced as part as the *tools* section of the website.

>

**Aside from your developmental work, you have also conducted your own qualitative research on the basis of L2-learner corpora (for instance, the paper "A multifactorial analysis of *this, that* and it proforms in anaphoric constructions in learner English"). Could you describe such work? What were the main findings/results?**

<

The main focus of this work is to better understand Interlanguage. This concept, born in 1972 under the efforts of Larry Selinker, hypothesizes the existence of developmental patterns and learning stages in foreign language learning. By using corpus-based analyses the idea is to try to

[11] https://lidile.hypotheses.org/cil

identify some of these stages and patterns. The publication you mention is one of the outcomes of my PhD in which I showed that learners have troubles in using *this*, *that* and *it* in English. Based on comparisons between two learner corpora (of different L1s) and one native English corpus, I developed different types of models showing the probabilities of using a form compared with its two other competitors. The results showed evidence of the existence of a paradigmatic microsystem in which the forms compete functionally as proforms. This has implications for language teaching, as most of the time *this* and *that* are used in a binary dialectic, i.e. the spatial distinction (nearness vs. farness). And yet learners also hesitate with *it* in some contexts. This shows that the forms' anaphoric value plays an important role and that learners tend to be confused when referring to discourse entities. Teaching materials thus need to adapt to this evidence.

Taking a broader view, the concept of microsystems was initially theorized in the late seventies by two linguists – Yves Gentilhomme and Bernard Py. They showed that learners use such forms in an unstable manner. Learners group forms unexpectedly and the delineations of the groupings evolve. By using corpora we are in a position to explore and capture many microsystems. For instance, as an English teacher I see students hesitate between *may*, *might* and *can*. By applying modelling methods, it is possible to compute the probabilities of use of each of these modals depending on contexts and proficiency levels. This can help to diagnose what learners need to adjust. And to do this, we need annotated corpora that include rich linguistic features.

The broader horizon would be the establishment of what Sylvain Auroux calls the third revolution of "grammatization". Based on quantitative methods applied to linguistics, probabilistic models will be trained on rich linguistic data sets. The models will encapsulate the grammar of a language, and this will support AI applications.

# CLASSLA, the Knowledge Centre for South Slavic Languages

## Introduction

Written by **Nikola Ljubešić, Taja Kuzman, Tomaž Erjavec,** and **Petya Osenova**

The CLARIN Knowledge Centre for South Slavic languages (CLASSLA) was recognized by CLARIN on 19 March 2019.[12] The centre offers support for the automated processing of South Slavic languages, and is operated by the Slovene CLARIN.SI, and by the Bulgarian CLaDA-BG.

CLASSLA recognizes the need for the development of language resources and technologies not only for Slovene and Bulgarian, but also for the other under-resourced South Slavic languages. That is why the centre aims to support researchers from the fields of Computational and Corpus Linguistics, Digital Humanities, as well as interested individuals from other scientific and business areas that use and produce language data for Slovene, Croatian, Serbian, Bosnian, Montenegrin, Macedonian, and Bulgarian.

**Space for productive cooperation of small nations**

The languages supported by CLASSLA are spoken by a small number of speakers. The estimated number of speakers worldwide ranges from less than half a million (Montenegrin, Hlavac 2013), between 1.4 million (Macedonian, Wikipedia) and 2.5 million (Slovene, Krek 2012, and Bosnian, Hlavac 2013), to over 5.5 million (Croatian, Tadić et al. 2012) and around 9 million speakers of Serbian (Hlavac 2013) and Bulgarian (Blagoeva et al. 2012). All seven CLASSLA languages together are used by around 30 million speakers. These seven languages form a dialect continuum with various degrees of mutual intelligibility between neighbouring languages.

[12] https://www.clarin.si/info/k-centre/

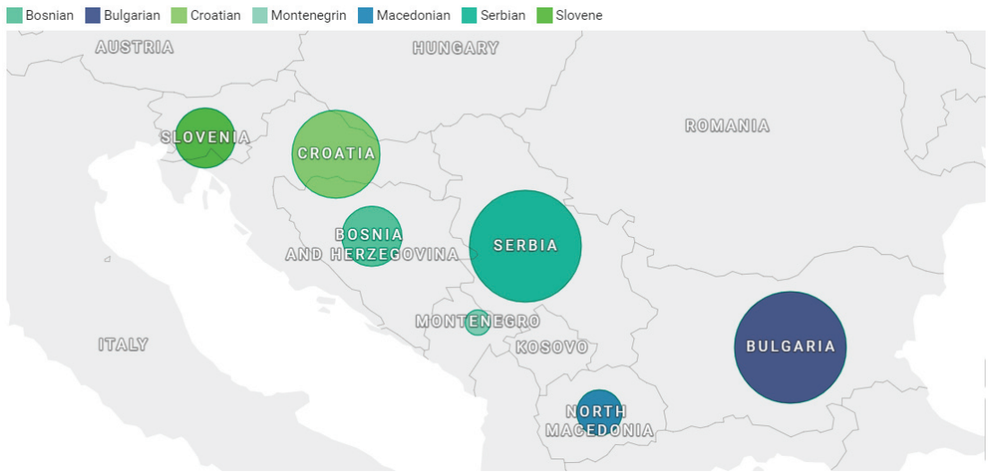



**Figure 12:** Map of the countries in which a South Slavic language is the most prominent. The size of the circles represents the number of speakers worldwide and the colour encodes the level of similarity between languages (created with Datawrapper).

The production of resources and tools for South Slavic languages costs just as much as for global languages such as English with more than a billion speakers. However, despite the small number of speakers and consequently very small language technology communities, it is crucial for the maintenance of an equal status of South Slavic languages in future digital environments that they be supported with the same technologies as global languages. This is where CLASSLA plays an important role. The knowledge centre provides a space for the cooperation of researchers interested in any of the South Slavic languages, as well as for a rational and economical approach to solving common problems, especially in the light of the mutual intelligibility of most of the languages.

To stimulate the development of language resources and technologies, CLASSLA provides information on freely available dictionaries, corpora, concordancers, (manually annotated) datasets, tools, and pipelines. The information is provided in the form of frequently asked questions (FAQ), and it is aimed towards both non-technical and more technically educated audiences. Currently, there are FAQs available for Slovene, Croatian, Serbian, Macedonian, and Bulgarian. The information is regularly updated to encompass all emerging resources and technologies.

**Figure 13:** List of topics covered in the FAQs.

In addition to this, CLASSLA supports researchers in producing resources and technologies for South Slavic languages via its helpdesk, which can be contacted at *helpdesk.classla@clarin.si*. So far, it has provided individual help to more than 50 researchers.

To share knowledge and enlarge the South Slavic language technology community, CLASSLA organizes workshops and raises awareness about its activities at home and abroad. In 2020 the first CLASSLA workshop was organized, which brought together 42 researchers.

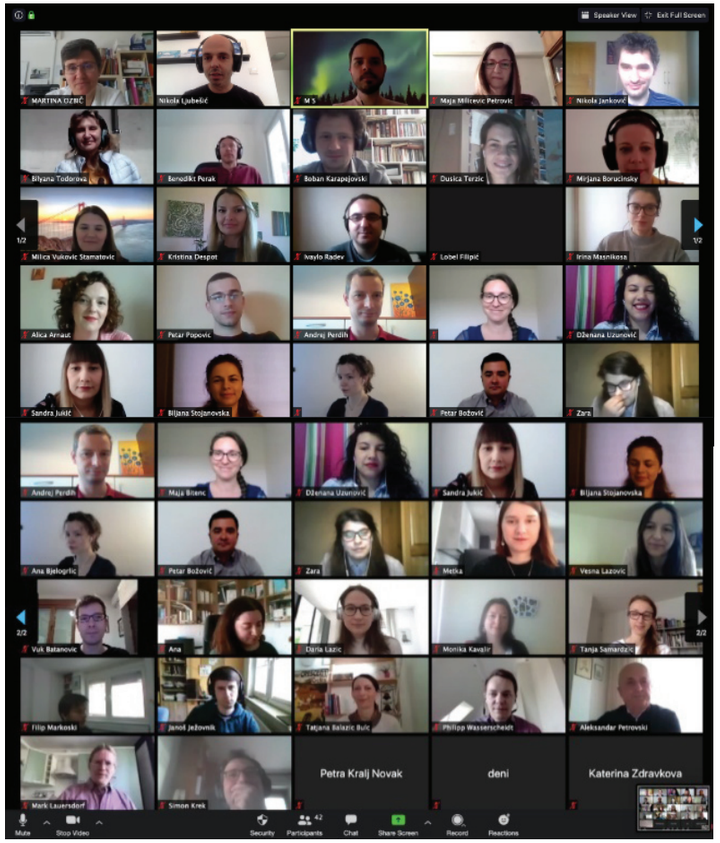



**Figure 14:** The first CLASSLA workshop, which took place in May 2020.

**Developing and providing freely available technologies and resources for under-resourced languages**

The findings of the META-NET White Paper Series “Europe’s Languages in the Digital Age” revealed that although there are some language technologies and resources for Slovene, Croatian, Serbian, and Bulgarian, these languages are only fragmentarily or weakly supported by machine translation, speech recognition, and text analysis (Krek 2012). CLASSLA aims to help close the technological gap between South Slavic languages and Western European languages by enabling researchers to acquire easy and long-term access to language resources via the CLARIN.SI repository, which, for example, comprises literary, news, web, spoken, parallel, parliamentary, and computer-mediated communication corpora. For some of the South Slavic languages, these corpora are the first of their kind ever made. For instance, the first ever linguistically

annotated Macedonian corpus (CLASSLAWiki-mk) was created in 2020 as part of CLASSLA's project of generating Wikipedia corpora in seven South Slavic languages. All the corpora can be downloaded and used locally, or queried via the two CLARIN.SI concordancers, NoSketch Engine and KonText.

Recently, the CLASSLA neural pipeline, an adaptation of the highly popular Stanza package, was built, and it offers state-of-the-art language processing of Slovene, Croatian, Serbian, Macedonian, and Bulgarian. The pipeline encompasses both standard and non-standard language processing, processing from tokenization to syntactic parsing and named entity recognition for most of the supported languages, with semantic parsing being currently added to the pipeline. The CLASSLA pipeline is designed to suit the needs of researchers with various backgrounds, from the non-technical linguists who can simply run the pipeline as described in the documentation, to the more technically sophisticated engineers who can use the pipeline to train their own language models. This year a state-of-the-art transformer model BERTić was also trained that covers Bosnian, Croatian, Montenegrin, and Serbian. Transformer models are large language models consisting of millions, or even billions of parameters that produce a general numerical representation of a portion of text, which is then used for various NLP tasks, from part-of-speech tagging, via text classification and machine translation, to text summarization and question answering.

```
>>> import classla
>>> nlp = classla.Pipeline('mk')
>>> doc = nlp('Крсте Петков Мисирков е роден во Постол.')
>>> print(doc.to_conll())
```

**Figure 15:** The user-friendly CLASSLA pipeline code (example for Macedonian) brings language processing closer to non-technical researchers.

In the two years since the inception of CLASSLA, South Slavic languages have become supported with many new technologies and resources, and many more are planned for the near future. Currently, CLASSLA is a part of the MaCoCu project, which will produce large high-quality monolingual and bilingual web corpora for under-resourced languages, South Slavic languages included. CLASSLA is also aware of the current technological advances in speech technologies, and is working on ensuring a comparable technological coverage of South Slavic languages to their larger counterparts in that area as well. Finally, CLASSLA plans to add a newsflash and other dissemination channels to continue supporting and enlarging the South Slavic language technologies community.

**References:**


Blagoeva D., S. Koeva, and V. Murdarov 2012. *The Bulgarian Language in the Digital Age. META-NET White Paper Series,* edited by G. Rehm and H. Uszkoreit. Berlin, Heidelberg: Springer.

Hlavac, J. 2013. Interpreting in one's own and in closely related languages: Negotiation of linguistic varieties amongst interpreters of the Bosnian, Croatian and Serbian languages. *Interpreting* 15 (1): 94–125.

Krek, S. 2012. *The Slovene Language in the Digital Age. META-NET White Paper Series,* edited by G. Rehm and H. Uszkoreit. Berlin, Heidelberg: Springer.

Tadić, M., D. Brozović-Rončević, and Kapetanović, A. 2012. *The Croatian Language in the Digital Age. META-NET White Paper Series*, edited by G. Rehm and H. Uszkoreit. Berlin, Heidelberg: Springer.

Wikipedia contributors, Geographical distribution of Macedonian speakers, *Wikipedia, The Free Encyclopedia,* https://en.wikipedia.org/w/index.php?title=Geographical_distribution_of_Macedonian_speakers&oldid=1041157079 (accessed 17 September 2021).

# Interview | **Zrinka Kolaković**

Zrinka Kolaković holds a postdoctoral position at the Department of Slavic Languages and Literatures at the University of Klagenfurt. Her current and future research interests are mainly focused on clitics and aspect.

The interview was edited and proofread by Krystyna Kupiszewska.

**Please introduce yourself – your academic background and current position.**

<

In 2008, I graduated from the University of Zagreb with a degree in Educational Science (Pedagogy) and Croatian Language and Literature. The profound theoretical linguistic knowledge of my mother tongue that I acquired during my schooling was further expanded during my time as a PhD student not only at the University of Regensburg but also at the University of Zagreb, where I pursued a *Cotutelle de Thèse,* i.e. joint supervision thesis, which was supervised by Professor Björn Hansen and Professor Zrinka Jelaska.

A vital catalyst that led me to use corpus linguistics for my PhD thesis was a corpus-linguistic course organized at the University of Regensburg by my colleague Edyta Jurkiewicz-Rohrbacher, who is currently a Postdoctoral Assistant at the Department of Slavic Languages and Literatures at the University of Regensburg. Later on, as the main research assistant, I had the opportunity to work with Edyta even closer in the project HA 2659/6-1 *Microvariation of the Pronominal and Auxiliary Clitics in Bosnian, Croatian and Serbian. Empirical Studies of Spoken Languages, Dialects and Heritage Languages* financed by the German Research Foundation (Professor Björn Hansen was the head of the project). Here, I learned how to design corpus linguistic experiments in order to empirically and statistically prove or disprove formulated hypotheses. Also, to overcome the drawbacks of corpus linguistic methods (primarily the problem of negative evidence), Dušica Filipović Đurđević, who now holds an Assistant Professorship at the Department for Psychology at the University of Belgrade, taught me how to create a fully crossed factorial design for a psycholinguistic experiment, to design stimuli and fillers for acceptability judgment tasks, and to conduct experiments. The results of our several corpus studies and seven psycholinguistic experiments, in which we tested 336 Štokavian native speakers from all over Croatia, are already partially published in Slavistics journals like *Rasprave* and *Jazykovedný časopis,* and will very soon be fully published by Language Science Press in a book entitled *Clitics in the Wild: Empirical Studies on the Microvariation of the Pronominal, Reflexive and Auxiliary Clitics in Bosnian, Croatian and Serbian.*

Work on this project, together with my passion for a better understanding of fascinating clitic phenomena, really made me think about the fundamental principles of empirical research in linguistics and led me to adopt a research philosophy that is based on the triangulation approach to methodology. Currently when I try to find answers to my research questions, I combine introspective, corpus-linguistic, and psycholinguistic methods. However, I must emphasize that corpus-linguistic methods always hold a central place in my research.

I presently hold a postdoctoral position at the Department of Slavic Languages and Literatures at the University of Klagenfurt. Although currently on maternity leave, I am planning to work on my own project proposal in which I would pursue not only a synchronic but also a diachronic study of so-called phrase splitting in various Slavic languages. The sentence *Sestrina mi prijateljica sutra dolazi u posjet,* "My sister's friend is going to visit me tomorrow", is an example of the phenomenon, with the pronominal dative clitic mi inserted between the possessive adjective *sestrina* and the head noun *prijateljica.*

>

**Your PhD thesis discussed biaspectual verbs at the crossroads of descriptive theories, prescriptive rules, and actual usage. Could you briefly present the thesis? What was the research question?**

<

In my dissertation, entitled *Biaspectual Verbs: the Difference between Description, Prescription and Real Use,* I investigated a topic in South Slavistics that is empirically poorly studied: biaspectual verbs (BVs) in Croatian on four language levels. The levels were lexical (with a focus on actionality, i.e. (a)telicity, durativity, dynamics and phasality), morphological (with a focus on the formation of new overtly aspectually marked derivatives and factors that influence affixation of BVs),

sentential (with a focus on the usage of BVs and their derivatives in aspectual sentential functions, i.e. the concrete-factual, general-factual, iterative/habitual) and textual (with a focus on the usage of BVs and their derivatives in aspectual taxis functions, i.e. sequence of events and coincidence of events).

I addressed only one research question on the lexical level: whether morphologically stable (without any overtly aspectually marked derivatives; e.g. *apstrahirati* "to abstract", *ilustrirati* "to illustrate", *veljeti* "to say") and unstable (with overtly aspectually marked derivatives; e.g. *karakterizirati – okarakterizirati* "to characterize", *savjetovati – posavjetovati* "to advise/to counsel", *častiti – počastiti, čašćavati* "to pay for dinner/lunch/to honour", *eksplodirati – eksplodiravati* "to explode") BVs differ on the lexical level. That is, are their actional properties significantly different? On the morphological level, I dealt with five research questions, one of which was whether prefixed derivatives of base BVs are equally present in different corpora of the Croatian language, i.e., corpora reflecting the standard and colloquial language use. I also addressed one research question concerning the sentential and textual levels: whether base BVs and their perfective (henceforth, PFV) derivatives differ significantly in respect of their usage on the sentential level (i.e., their distribution in aspectual sentential functions), and whether base BVs and their PFV derivatives differ significantly in respect of their usage on the textual level, i.e., their distribution in aspectual taxis functions.

>

**How did you go about it methodologically i.e., what corpus-linguistic methods did you use? Which corpora did you use for your thesis (any CLARIN corpora)?**

<

The corpus-linguistic data were used as the primary data source in studies on the morphological, sentential, and textual levels. While studying BVs on the morphological, sentential, and textual levels, I used two Croatian CLARIN.SI corpora: the web corpus hrWaC and the reference corpus Riznica. First, these corpora were used to find out which BVs do and which do not form prefixed and suffixed derivatives. Then, if it turned out that they do have such derivatives, I established precisely which derivatives are formed this way and how frequent they are in comparison to their base verbs. In addition, the two corpora were used to sample random sentences with BVs and their PFV derivatives to check the differences in their distribution in aspectual sentential functions (i.e., concrete-factual, general-factual, iterative/habitual) and aspectual taxis functions (i.e., sequence of events and coincidence of events). However, I must emphasize that the results obtained on the sentential and textual levels should be complemented either with more corpus-linguistic data or with experimental data, and preferably with both types to allow greater control over factors that were not studied and might possibly turn out to be confounding. Especially on the textual level, i.e., for the usage of BVs and their derivatives in aspectual taxis functions, the obtained results were inconclusive.

Nevertheless, I can happily say that for the very first time I have managed to demonstrate that factors influencing the prefixation of BVs of both Slavic and foreign origin in Croatian can be analysed on the morphological level with the help of advanced statistical analysis, such as the generalized linear mixed model. The application of statistical methods revealed several exciting results. First, the origin of the base biaspectual lemma has a statistically significant impact as a factor on prefixation: BVs of Slavic origin (e.g., *savjetovati* "to counsel") are more likely to be prefixed (e.g., *posavjetovati*) than biaspectual borrowings (*ilustrirati* "to illustrate"). Second, prefixation of BVs with a synchronic and/or diachronic prefix (e.g., *doručkovati* "to have breakfast") and prefixation of BVs that do not have such a prefix (e.g., *karakterizirati* "to characterize") differs significantly: having a synchronic and/or diachronic prefix, like in the case of *doručkovati,* has a negative impact on prefixation of BVs. Third, BVs for which suffixed derivatives are attested (e.g., *parkirati* "to park", *vezati* "to bind") are more prone to prefixation. Fourth, BVs with different numbers of meanings differ significantly with respect to prefixation, in the sense that the more polysemous BVs (e.g., *častiti* "to 'pay for dinner/lunch'/to honor", *vezati* "to bind") seem to be more prone to prefixation. Fifth, prefixation of BVs is more frequent in corpora containing colloquial and unproofread texts than in corpora compiled from texts written in the standard Croatian variety.

Furthermore, on the sentential level, by analysing the distribution of BVs and their perfective derivatives in aspectual sentential functions, I showed that derivation of overtly aspectually marked derivatives from base BVs could be motivated in some way by aspectual sentential functions. My assumption is that PFV derivatives might be formed to explicitly distinguish between the progressive and the concrete-factual aspectual sentential functions. Specifically, if a BV is uttered in the future or past tense and obvious taxis signals (such as co-temporality) and other cues are missing, the addressee will not be sure whether the speaker's intention was to convey the progressive or the concrete-factual function (e.g., *Analizirat ćemo te tvrdnje* "We will analyse those claims/We will be analysing those claims"). However, the aspectual vagueness is eliminated when a perfective derivative is employed (*Proanalizirati ćemo te tvrdnje* "We will

analyse those claims"). In that case, the addressee can be certain that only the concrete-factual meaning was conveyed.

And finally, in the study of BVs on the lexical level, I was the first to empirically show that the morphological stability of biaspectuality (non-formation of overtly aspectually marked derivatives) is interrelated with actional features, i.e., the lexical aspect of the base BV. Specifically, the meanings of the analysed BVs without attested overtly marked aspectual derivatives have, almost without exception, telic actional properties. Conversely, BVs with overtly marked aspectual derivatives have many more meanings with atelic actional properties.

>

**What were the novel empirical findings that contradicted the prescriptive rules or shed new light on the descriptive theory?**

<

My results strongly suggest that aspectual affixation of BVs cannot be categorized as the formation of mere pleonasms (i.e., redundant forms), as argued in the normative literature. Quite the contrary, the mentioned phenomenon is highly functionally motivated and triggered or suppressed by various factors, of which I only identified the ones mentioned above. The results are (or are about to be) published in the journals *Russian linguistics*, *Suvremena lingvistika* (Contemporary Linguistics) and *Zeitschrift für Slavische Philologie* (Journal for Slavic Philology) as well as in the book *Glagolski aspekt i dvoaspektni glagoli u hrvatskome jeziku: formalno-funkcionalni pristup(Verbal Aspect and Biaspectual Verbs in Croatian: A Formal-Functional Approach).*

>

**How has the CLASSLA K-centre helped you in your research? Have you used any specific CLARIN/CLASSLA-related tools (e.g., NoSketch Engine concordancer)? What makes these tools important for researchers working with South Slavic languages?**

<

The CLASSLA K-centre has helped me in many ways. First, in 2013 when I started working on my PhD thesis, I knew that I wanted to compare prefixation of BVs in corpora which reflect standard Croatian language use, and in corpora which contain texts without any external proofreading. This was necessary in order to establish whether the usage of BVs and their derivatives by average users is significantly different than usage by those who strictly follow the norm or have their text externally proofread and, of course, to obtain non-skewed data, which could give me a more precise picture of which factors govern prefixation of BVs in Croatian. In 2013, the CLARIN.SI hrWaC web corpus with a mixture of proofread and non-proofread Croatian language had already been compiled, and it also contained texts from the forum.hr subdomain with exclusively user-generated non-proofread text. The Croatian National Corpus was also available at that time. Still, I was not entirely happy with it, and I wanted to use an additional corpus representing standard Croatian language usage, Riznica. But at that time, Riznica was neither lemmatized nor morphosyntactically annotated. Not to mention that the concordancer via which it was searchable was not really user friendly, i.e., it did not permit quick queries and sorting of results. Luckily, before I finished and submitted my PhD thesis Riznica became fully annotated and available at the Slovene CLARIN via my favourite NoSketch Engine concordancer. The abovementioned CLARIN corpora were also extremely important and relevant for the project *Microvariation of the Pronominal and Auxiliary Clitics in Bosnian, Croatian and Serbian. Empirical Studies of Spoken Languages, Dialects and Heritage Languages.* Unfortunately, some corpora that could have been relevant for our work, such as Torlak, appeared after we had to stop the empirical phase of the project and concentrate on finishing the manuscript.

Although until March 2022 the CLARIN South Slavic corpora are also available via Sketch Engine, I personally prefer and always suggest that my students use the NoSketch Engine provided by the Slovene CLARIN instead. One of the advantages is that the size of the downloadable example sentences, sorted lemmas, etc. is not restricted to 1,000, whereas in Sketch Engine obtaining more than 1,000 concordances is not free of charge. The COVID pandemic also demonstrated the tremendous value of having corpora available at any time and place. Gathering empirical and big data by linguists is hardly possible otherwise. This was also one of the reasons why, at the Department of Slavic Languages and Literatures at the University of Klagenfurt in 2021, in the first online semester we decided to welcome Tomaž Erjavec and Nikola Ljubešić from CLARIN.SI/CLASSLA for the invited lecture. And finally, I would like to emphasize that online corpora are especially important for those like me who hold PhD, postdoctorate, and similar non-permanent positions and will one day have to compete for new (hopefully permanent) positions, and even more so for those among us with (small) children. I am very grateful for the work of the whole CLARIN.SI team because I can (try to) conduct empirical research from my home even when I am on maternity leave.

>

**Do you feel that South Slavic languages are under-resourced? Why is the CLASSLA K-centre essential for overcoming this issue for these languages?**

<

When I look back at the time when I started working on my PhD thesis or in the project *Microvariation of the Pronominal and Auxiliary Clitics in Bosnian, Croatian and Serbian in BCS*, I have to say that many new interesting corpora have appeared in the meantime and the existing ones are being monitored, enlarged and better annotated. In our book on clitics there is one whole chapter dedicated to the review of existing BCS corpora, which I will summarize here. If I compare, for instance, the corpora available for Slovene and Russian, their tagging accuracy and the concordancers via which they are searchable, well, then it is clear that David has won over Goliath. The Russian National Corpus is way smaller than Gigafida and its concordancer does not allow for powerful CQL queries which make it possible to look for syntactic patterns. The ruTenTen is impressive in size, but as a user I am not happy either with its tagging accuracy or with its MSD tagset description, since it does not correspond to actual tags. So when trying to formulate CQL queries for ruTenTen one cannot rely on the MSD tagset, but instead has to look up the actual tags in the corpus first.

Croatian is still lagging behind Slovene; however, the situation is not entirely bad. For instance, there is a diachronic corpus (CroDi) as well as a corpus of spoken Croatian (HrAL). Still, they are neither fully annotated nor available via a user-friendly concordancer such as NoSketch Engine, not to mention that the last time I checked I couldn't find access to HrAL. Moreover, it seems that CroDi's concordancer Annis does not allow the user to download metadata regarding authorship (name, work), text age and the dialect in which it is written. These kinds of metadata are important in linguistic studies, and even more so in diachronic ones.

In my opinion, Serbian and Bosnian are definitely the most under-resourced South Slavic languages, although the situation with Serbian is slightly better; there is, for instance, the Torlak dialectal corpus (also hosted by CLARIN.SI).

**The CLASSLA K-centre is essential since it maintains the existing corpora and improves them, handles the construction of new corpora and tools, and strives to put all available tools and corpora under one roof by providing access to them via the free-of-charge, user-friendly NoSketch Engine concordancer.**

**You are helping (or plan to help) CLASSLA in building the new generation of web corpora for South Slavic languages. Could you describe the goals of this endeavour?**

<

After email communication with Nikola Ljubešić, a group of Croatian linguists who were keen to help had a kind of brainstorming kick-off meeting where we decided to split into two teams. One team is interested in improving tagging accuracy and the other in splitting hrWaC into subcorpora for different registers. I would like to take part in both groups. Regarding the work of the first team, I am especially interested in the MSD tagging accuracy of clitics, but also other (word) forms which have homonym/homograph (word)forms with other clitics or other parts of speech. However, the current version of the CLASSLA ReLDI tagger is more than satisfactory in dealing with these phenomena, especially in comparison to taggers that were used to tag the Corpus of Contemporary Serbian Language (for instance, all instances of the BCS wordform je are tagged only as a verb and none as a pronoun) and the Croatian National Corpus (where for example the tagging accuracy of the word form te requires checking). As for the other team's tasks, I am most interested in "cleaning" the hrWaC from all user-generated content (not only content from the forum.hr subdomain but also other kinds of comments, blogs, and fora) by building a separate subcorpus containing only the CMC data, so that it would then resemble the Slovene JANES corpus. Ideally, this should also be done for the Bosnian bsWaC and Serbian srWaC web corpora, since their current versions entirely lack this kind of CMC subcorpus (for Croatian, at least the Forum subcorpus exists).

>

**What can be done (by CLASSLA) to facilitate the creation of new comparable resources in these languages?**

<

I already mentioned that for Bosnian and Serbian, or better to say for the CLARIN.SI-hosted bsWaC and srWaC corpora, large-scale subcorpora with user-generated content are missing (I am familiar with the Tweet.hr and Tweet.sr corpora, but these are small in size and due to the nature of tweets the language in them is very specific in its syntax and other characteristics). Such content would be very useful, especially for those linguists who are interested in language variation and generally in establishing differences between the prescribed norm and the actual language use. But for Bosnian and Serbian in general, decent widely available and fully annotated corpora of standard varieties which would be searchable via a user-friendly concordancer are also missing. It would also be nice to have (comparable) dialectal, and diachronic corpora of spoken Bosnian, Croatian, and Serbian searchable via NoSketch Engine.

# NLP:EL, the Knowledge Centre for Greek

## Introduction

Written by **Maria Gavriilidou** and **Iro Tsiouli**

NLP:EL is the CLARIN Knowledge Centre for Language Technology and Language Resources in Greece.[13] NLP:EL is a relatively new CLARIN K-centre. It was established in March 2020 and is hosted by the Institute for Language and Speech Processing (ILSP) of the Athena Research Centre.

ILSP is a research and development organization in the area of Language Technology in Greece, whose activities cover a broad range of the Language Technologies spectrum. Besides research and development, ILSP is actively involved in educational activities in collaboration with universities. To complement and support its research efforts in these areas, ILSP continuously invests in developing its Language Resources Infrastructures, prominent among which is CLARIN:EL, the national research infrastructure on Language Resources and Technologies, which aims to be the central point for Language Technology and Language Resources in Greece.

NLP:EL, the Knowledge Infrastructure for Language Technology in Greece, constitutes an integral part of CLARIN:EL, together with the central catalogue which aggregates resources (datasets, lexical resources, tools, services, and workflows) developed mainly by the national network members, but also significant external resources; it currently hosts 520 resources and more that 40 tools/services.

[13] https://www.clarin.gr/en/kcentre

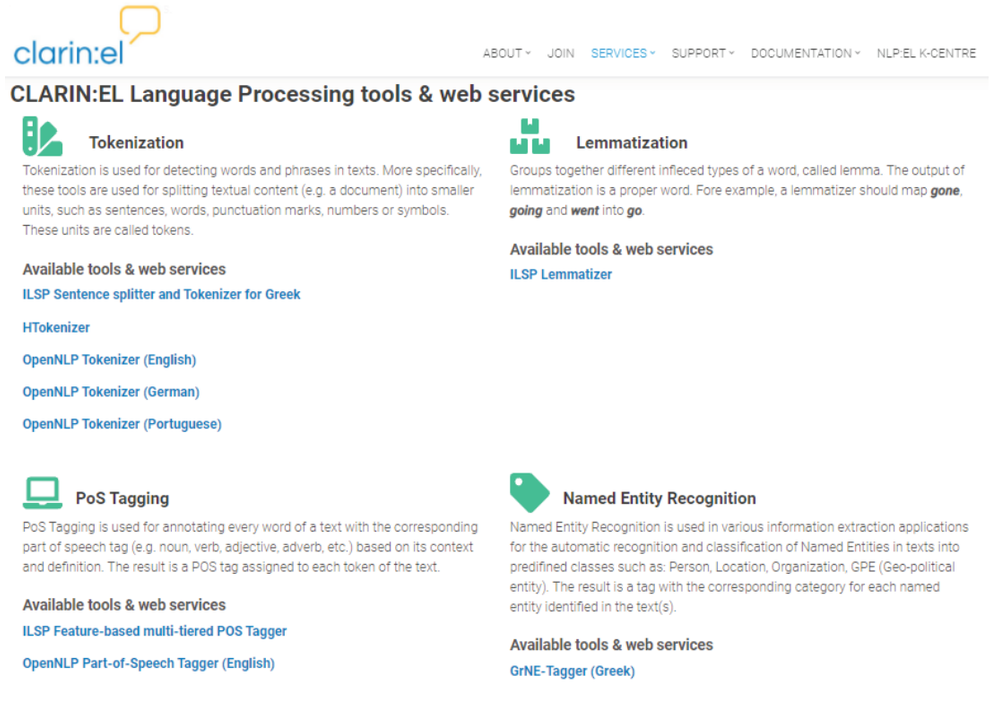



**Figure 16:** CLARIN:EL language processing tools offered as web services.

NLP:EL has the mission to support language technology research for the Greek language, the digital readiness of Greek, and sign language technologies research and development.

Through its web pages NLP:EL provides a plethora of services, **including information about language processing tools and web services.** Some of the most frequently used NLP processes have been selected among those offered by CLARIN:EL, that is, tokenization, lemmatization, part-of-speech tagging, named entity recognition and chunking. For these, NLP:EL provides a brief definition of each process describing their function and examples to clarify the task they perform; then, a list of the respective tools/services offered by CLARIN:EL is provided and, finally, via a link to the landing page of each tool the users are redirected to the CLARIN:EL repository, where they can view the metadata description of each tool/service and get access to use it. Thus, a user who wants to know for instance what lemmatization is and why it is useful is guided through the definition and the list of tools that perform this task, from where, based on the metadata descriptions and the usage conditions, they can select the most appropriate one.

NLP:EL also provides **videos and tutorials** about selected services and applications provided by CLARIN:EL, whether developed by CLARIN:EL members and integrated as web services or developed by third parties. An example of this is a video describing the Named Entity Recognizer (NER) service of ILSP that is available through YouTube.[14] ILSP NER is a tool for the recognition of proper nouns in a text (person names, place names, companies, names of months, days), but also of paralinguistic items like dates, numbers, emails, etc. In addition to the identification of these items, the NER tool annotates them with the appropriate tags (e.g., PERSON, LOCATION, DATE, etc.). The tool is available from CLARIN:EL as a web service. The video, using the NER web service as an exemplifying case, describes the whole procedure a user has to follow when using the CLARIN:EL infrastructure, from the selection of the service from the CLARIN:EL inventory to the uploading of the data to be processed, then to the online running of the service and finally to the downloading of the results and their subsequent inspection and analysis.

As an additional example, there is a webinar recording which presents Voyant Tools, a web-based analysis environment for digital texts (Sinclair & Rockwell) listed in the CLARIN:EL inventory. The webinar was part of a series of training activities whose aim was to introduce various text analysis tools to interested students, educators, language professionals, etc. This webinar was presented by Professor Dimitroulia (interviewed in this issue), and the recording was made available publicly.

There are several **manuals and guides for tools and/or applications** that have been developed by the CLARIN:EL team or that can be accessed through the CLARIN:EL Research Infrastructure. For example, CLARIN:EL hosts an instance of WebAnno, a web service application for collaborative text annotation developed by the Computer Science Department of Technische Universität Darmstadt. Through this platform, users can upload their texts, create their own projects, define their tagsets or use already built-in tagsets, invite other users to their projects in order to annotate texts in collaboration and at the end export the annotation results locally. Naturally, WebAnno has its own documentation; however, documentation manuals have been written in Greek from scratch for CLARIN:EL users, covering all WebAnno user roles (curator, annotator, project manager).

NLP:EL also offers **access to services and products in the fields of dynamic sign language synthesis,** such as the Fingerspelling keyboard, a virtual keyboard for alphanumeric symbols corresponding to the signs for the 24 letters of the Greek alphabet, and the Dynamic Synthetic Signing environment, which allows users to produce new sentences in Greek Sign Language by selecting the components of each phrase to be produced from a glossary of signs; the users can preview every phrase he or she produces through a virtual signing avatar and modify it if necessary.

**Educational materials** offered by NLP:EL include scientific publications and slides of presentations in the relevant fields published by the CLARIN:EL network team as well as a list of university courses in the fields of Language Technology, Data Science, Information Technology and Digital Humanities offered by Greek Universities.

There is also a **dedicated helpdesk,** which supports the users in their quest for information on the above subjects. Personal advice is often sought by individual users on concrete issues that concern them, and these requests are supported accordingly.

Lastly, NLP:EL organizes **training events,** through which knowledge on the domains of expertise is transferred. The training events take the form of webinars, tutorials, hands-on sessions or focus groups, depending on the needs and the audience. Thus, separate events have been organized, dedicated to user groups with different backgrounds (literature studies, social and political sciences, computer science) catering for their special needs; for example, metadata curation and data deposition, use of language processing tools, dockerization of services and integration thereof into the CLARIN:EL repository, etc.

As a prominent example, NLP:EL organized the CLARIN:EL Summer School 2021, which was held online between 6 and 8 July 2021. It was attended by 75 participants, with backgrounds in Library Science, Language Studies, Education, History and Archaeology, Literature, Computer Science, Digital Humanities and Political and Social Sciences. The participants had the opportunity to get acquainted with the basic concepts of Language Technology and Language Resources, to take a deep dive into data collection, curation and processing, to discover LT-based applications, to hear about the role of national and EU Language Resources and Technologies Infrastructures, and to familiarize themselves through hands-on workshops with the CLARIN:EL Infrastructure, the curation of resources and the use of NLP tools.

With more than 16,400 users and 28,400 page views since the establishment of NLP:EL in March 2020 (based on Google Analytics), the K-centre and the CLARIN:EL inventory, as integral parts of the national research infrastructure, are committed to ensure the digital preservation and readiness of the Greek language, by supporting research in the field of NLP and the development of language technologies in Greece.

[14] https://www.youtube.com/watch?v=jve__Tl7MB8&t

# Interview | **Titika Dimitroulia**

Titika Dimitroulia is a professor of translation studies, as well as a translator and literary critic. Many of her literary corpora have benefitted from NLP:EL processing tools and are deposited in the CLARIN:EL central inventory.

**Please briefly present yourself – your academic career and current position.**

<

I am a Professor of Translation Studies at the School of French, Aristotle University of Thessaloniki, a translator and a literary critic. I have studied Classical, Modern Greek and French literature in Athens (National and Kapodistrian University of Athens) and Paris (Sorbonne-Paris IV), and I was acquainted with corpora and digital tools during my PhD research on translation technologies at Panteion University of Social and Political Sciences. Having an IT background, acquired through informal training, and influenced by the French school of lexicometrical/textometrical analysis, I then took the digital approach to literature. My research focused on corpora and text analysis, text representation and visualization and their impact on text interpretation, and I adopted the hermeneutical method, according to which text analysis is always qualitative, no matter if the tools and algorithms are used to provide quantitative data. Apart from text analysis, I am very interested in the digital textuality from a communicative, semiotic and literary point of view, and I have published a book on digital literary studies (with Katerina Tiktopoulou). Being a founding member of the Semiotics Laboratory at the Faculty of Philosophy, as well as of the Digital Humanities Lab, where I was the first director, I have integrated digital methods and tools in all my graduate and postgraduate courses on literature, translation and semiotics and created two specific courses on computer-assisted literary translation and literary studies. I am currently working on the translation of Geoffrey Rockwell and Stefan Sinclair's book on the philosophy and methods of hermeneutical text analysis in the humanities, *Hermeneutica,* which will be published in Greek soon, and I am preparing an introduction to the digital humanities, aiming to contribute to the consolidation of this emerging field in Greece.

>

**How are you involved with CLARIN:EL and its NLP:EL Knowledge Centre?**

<

I have been collaborating and sharing experiences on the use of language technologies in literature and translation with researchers from ILSP/ATHENA, in particular Maria Gavriilidou and Stelios Piperidis, since the 1990s. So, when they set up the Greek CLARIN network they contacted me, and since 2014 I have been part of the network on behalf of my university. Since CLARIN and DARIAH came together in Greece in the new infrastructure, Apollonis, I have been coordinating Aristotle University's Clarin-Apollonis team in the network.

We deal mainly with language technology in the humanities and social sciences and focus mostly, but not exclusively, on text (and discourse) analysis. Our goal is twofold: to create and host resources, especially corpora, in the field of the humanities and social sciences, as a base for electronic analysis; and to use these resources in seminars and presentations in order to encourage colleagues and students to join the language technology community, through precise case studies dealing with specific domains. For me, presenting easy-to-use, open-source tools to the community is a deliberate and conscious choice, so that colleagues and students can immediately see the benefits of language technology in their own research. Until now, we have had more than 1,000 registrations for our seminars, and for the new academic year we plan to focus systematically on the more complicated functions of tools such as Voyant Tools and CATMA, or Orange Textable. All information about our seminars and their videos can be found on our website.

I personally use the CLARIN:EL infrastructure to host and analyse my corpora. I also supervise Aristotle University's repository and coordinate my team's work and contribution to the infrastructure as well as our training and research activities. I take part in the training, by organizing and giving seminars and talks, and develop our collaboration with NLP research communities abroad. Our current training activity is our contribution to the NLP:EL Knowledge Centre, as some of the videos and presentations of our seminars and talks are also available through their interface. Finally, I support the users of Aristotle University's repository in their questions concerning Language Technology in humanities research. I am happy to say that, as result of our activities, the first community on text analysis has been built and new courses have

been introduced in the departments' curricula, such as the one on quantitative and qualitative text and discourse analysis in the School of Political Sciences at AUTh.

>

### How do you use the Greek CLARIN infrastructure? What are the main characteristics of the resources that you have contributed?

I use the CLARIN:EL infrastructure to host and analyse the monolingual and parallel corpora that I need for my own projects in the fields of Modern Greek, French literature and translation. In relation to literary texts and translations, I would like to stress a permanent problem, which is the restrictions we face due to copyright. Although raw and annotated corpora are "bags of words", which the typical reader can hardly read, we always face licensing restrictions, which prohibit us from sharing available resources with the community, while one of greatest problems of the NLP community in Greece is the scarcity of resources. This is the reason why I always try to strive for open access; for example, the parallel corpus of French literary works translated into Greek (FREL), which I initially created and which can be used both in translation teaching and practice, is accessible through a dedicated interface (Figure 17). However, some of my corpora are not in the public domain and can only be accessed upon demand.


**Figure 17:** The interface of FREL.

Some of the texts contained in FREL, though, are included in the public translation corpus GLTC that is accessible through the CLARIN:EL Central Inventory.

All the corpora that I have created and are hosted publicly on the repository, like the GLTC, are raw literary corpora, which means that initially they are not linguistically annotated; however, they can be processed with CLARIN:EL (as well as external) tools that are integrated with their repository. The corpora can be used individually and in combination with other resources for purposes of linguistic, stylistic, stylometric, thematic, pragmatic and other analyses. We have already used the corpus of the literary works by the writer Alexandros Papadiamantis[15] together with a corpus of his signed translations, and a corpus of his contemporaries (all of which are hosted by CLARIN:EL), in an authorship attribution project, where we tried to identify if some anonymous translations were made by Papadiamantis.

One of my key resources that is not available publicly is the corpus of the literary works of the eminent Greek author Melpo Axioti, whose only novel written in French is hosted in the AUTh repository. This corpus is the result of my collaboration with the text and corpus linguist Dionysis Goutsos, professor at the University of Athens, as part of a wider, collaborative project on exile literature, in which we combined both close and distant reading. This project aims to explore the methods and practices in interdisciplinary "big humanities data".

In such an interdisciplinary approach, researchers need to resort to geomapping in order to effectively visualize the text. To visualize the data in the Axioti corpus I have used Dreamscape, a geomapping service provided by Voyant Tools. Dreamscape is, according to its creators, "a preliminary attempt to explore how texts might be represented geo-spatially. The tool tries to identify locations (especially city names) mentioned in texts, and suggests patterns of recurring connections between locations; patterns that might help identify travel of people, ideas, goods, or anything else". Figure 18 gives an example of geospatial visualization of Axioti's novel *Το σπίτι μου* ("My home").

[15] http://hdl.handle.net/11500/CLARIN-EL-0000-0000-6785-6

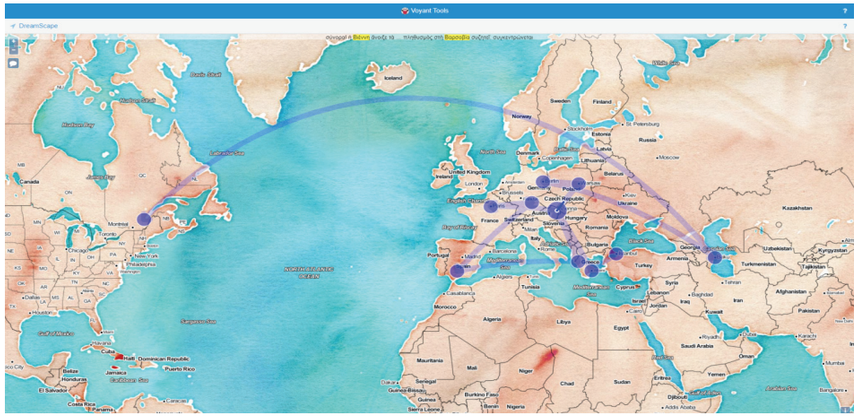



**Figure 18:** Geomapping applied to the Axioti corpus. The blue arc shows the occurrences of named entities as they appear in the text.

Although Dreamscape is currently in its beta-release stage, which means that the identification of locations is not really fully reliable, it still gives us an idea of how easily such tools could be used in the near future, as well as of their impact on macro-analyses. Of course, accurate geospatial representation of texts is still possible today through a combined use of existing tools – e.g., in the case of Greek texts, I think that CLARIN:EL Annotator of Named Entities can be of great help.

>

**In preparing such resources, has NLP:EL helped you with any kind of linguistic or structural annotation?**

<

Linguistic and structural annotation is very important in literary and translation studies, but what is crucial for qualitative research is to combine them with extra-structural markup. In the Melpo Axioti corpus, we have first used annotation tools provided by NLP:EL through CLARIN:EL – specifically, the CLARIN:EL Annotator of Named Entities and the GrNE-Tagger – in order to tag the recurrence of persons and places in Melpo Axioti's works, and the ILSP Feature-based multi-tiered POS Tagger and ILSP Lemmatizer to study the morphological changes in her language, which are connected with her ideological choices in different periods of her work. Afterwards we used CATMA, which is a tool for the extra-linguistic annotation of culture-specific concepts associated with a particular word in the text. For instance, we have annotated the occurrences of the word *house* which is the most frequently recurring lexical word in Axioti's literary works, and constitutes their central thematic element. We have thereby created a double tagset, in which the house is tagged as either a positive or a negative space, with subsets defined by positive (e.g., "home", "friendly") and negative ("disgust", "loneliness") connotations of the word, as seen in Figure 19.

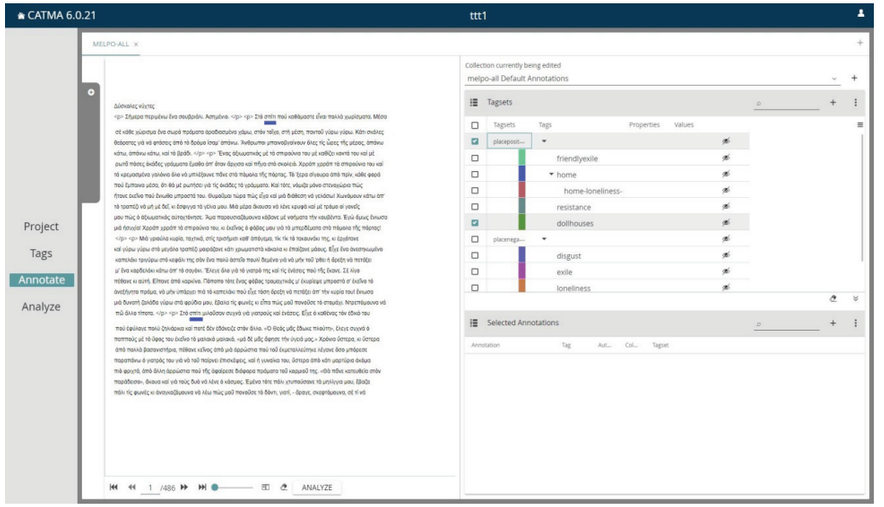



**Figure 19:** CATMA applied to the Melpo Axioti corpus.

On the basis of this, we then linked the presence or absence of the occurrences and particular meanings and connotations of the term *house* to broader thematic and stylistic choices, as well as to different periods, in Melpo Axioti's work. This sort of extra-linguistic annotation is a good example of how a researcher is able to work with qualitative research questions using a digital approach, especially in humanities, as such questions can only be formulated accurately if the user knows the corpus well and tries to shed new light on it by applying both linguistic and extra-linguistic tools offered by different developers.

>

**Your research has quite a large and interdisciplinary scope, combining corpus linguistics with approaches like post-structuralist discourse theory, as well as literary studies. Could you present a result or paper in which your research has benefitted from NLP:EL's involvement? How concretely did your research benefit from NLP:EL?**

<

I was involved in the POPULISMUS project, a research initiative focusing on populism and anti-populism in the Greek press that involved collaboration with colleagues working in political science. This led to several Greek and English publications. My contribution in this project was

mostly methodological: after a workshop on text analysis for political scientists in 2014, a team was created that I trained in corpus-based analysis. They got acquainted with the digital approach and applied quantitative and qualitative methods to confirm their theoretical hypotheses on populism, based on the so-called Essex School analysis. This project proves that hermeneutical text analysis as a method can apply to all kind of texts and research, on the condition that the specialists formulate and then test clear and solid hypotheses. This team is now exploring new paths in quantitative and qualitative analysis, integrating different methods of analysis and AI. An NLP:EL seminar, focusing especially on political discourse, was held recently at AUTh, with more than 250 participants from all over Greece, and we are now organizing a follow-up that will focus on the practical results of the project.

Concerning my work on translation and literature, I would like to present two examples in which my research has benefitted a lot from electronic text analysis and annotation. The first one relates to the translation of poetry and prose of Greek authors, like Andreas Embirikos and the aforementioned Melpo Axioti, who wrote in French (three poems by the surrealist poet Andreas Embirikos and some recently discovered short stories written in French by Melpo Axioti, to be published soon). When translating an author's discourse in a language other than their mother tongue, you face a major dilemma: you need to decide if you are going to translate simply what you read or try to reproduce the author's style in the original (in this case, Greek). I opted for the second strategy, so I needed to have the most accurate idea of what the author's style corresponds to and which devices it consists of. In my effort to define it, both textometrical analysis and thematic, morphological and syntactic annotations were of great help to me, in particular the ILSP Feature-based multi-tiered POS Tagger and ILSP Lemmatizer, which gave me an idea of the general use of the author's vocabulary and their morphological choices, and the ILSP Dependency parser for the prose, which was of great help in studying Axioti's syntax and identify recurrent patterns representative of her particular style. For example, when exploring synonymy I chose words actually used in the works, focusing on their morphological particularity, and I also reproduced, in the case of Axioti, her particular syntax. Since then, I have relied on similar approaches for all my translations, using linguistic-analysis tools like the aforementioned ILSP POS Tagger and Dependency parser together with Voyant tools and CATMA, with which I try to obtain concrete stylistic evidence as a basis for translating the text. Such a method of combining tools can be of great help not only for researchers, but also for professional translators.



The second example concerns Modern Greek poetry, in particular the work of the poet Kostas Papageorgiou. While studying his work for an essay on his poetry, I had the feeling that the frequency of comparisons in his poetry collections is linked to how his poetic vision had evolved through time. To confirm my hypothesis, I digitalized his collections, loaded them into Voyant Tools (which is also accessed through the CLARIN:EL repository), defined the words introducing comparisons and studied them in term of concordance. The quantitative data confirmed my hypothesis, as it turned out that comparisons occur less and less in his three last collections, becoming replaced by metaphors, where poetic discourse is represented in terms of somatic (i.e., bodily) functions. My study revealed another important role that comparison plays in his work – that is, the different status and function that this trope progressively acquires: at first, it is used to establish an analogy between realities of different nature, while over time it starts to signal the analogy itself, since the second element of the analogy is no longer spelled out and the word that signals the comparison (such as *like* in English similes) gets suspended. My findings on the frequency, function and transformation of comparison were the basis for an article I published on the poetry of Papageorgiou.

>

**What makes NLP:EL especially invaluable in facilitating corpus linguistic research in the aforementioned fields?**

<

It is researchers who hold the most important position in digital research, and particularly in text analysis. Their deep knowledge of the field and its theoretical framework is a prerequisite, as even the choice of the texts and the compilation of the corpus is crucial for the validity of the hermeneutical analysis. So, there is no digital analysis in vacuum; rather, it starts with a question formulated more or less clearly by a specialist, on the basis of the relevant theoretical premises. Digital analysis permits the confirmation or refutation of these premises, with authentic, quantitative data, which in the hermeneutic cycle of the texts have qualitative value.

**NLP:EL is really invaluable as it trains researchers how to effectively collaborate and share their methods and resources, thereby enhancing interdisciplinarity in many different ways. It also provides opportunities for exchanges and consultation and generally works as an invaluable forum and a training hub for the Greek NLP community.**

>

# ARCHE,
# the Austrian B-Centre for Digital Humanities and Cultural Heritage

## Introduction

Written by **Martina Trognitz**

The Austrian Centre for Digital Humanities and Cultural Heritage (ACDH-CH), founded in 2015, is an institute of the Austrian Academy of Sciences in Vienna.[16] The ACDH-CH pursues a dual agenda of conducting digitally enabled research and providing technical expertise and support to the research communities at the Academy, on the national and international levels. The institute has undergone a dramatic development in the last two years, evolving from the predecessor Institute for Corpus Linguistics and Text Technology (ICLTT) with its focus on language data, to an institutional and national centre of expertise in the broader field of digital humanities and cultural heritage.

ARCHE (**A R**esource **C**entre for the **H**umaniti**E**s) is central to ACDH-CH's mission of fostering the change towards the digital paradigm in the humanities. ARCHE is the successor of the 2014 initiated CLARIN Centre Vienna / Language Resources Portal (CCV/LRP). CCV/LRP's goal was to provide depositing services and easy yet sustainable access to digital language resources created in or related to Austria. In 2017 ARCHE replaced CCV/LRP extending its mission by offering advanced and reliable data management and depositing services open to a broader range of humanities disciplines in Austria. CCV/LRP was already featured in Tour de CLARIN, Volume I.

ACDH-CH and ARCHE have together contributed to the CLARIN infrastructure as a Service Providing Centre (CLARIN B-centre) since 2017. They are embedded within the European infrastructure consortia CLARIN ERIC and DARIAH-EU, which in Austria are jointly represented by CLARIAH-AT.

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



**Services by ARCHE**

The main aim of ARCHE (**A R**esource **C**entre for the **H**umaniti**E**s) is the long-term preservation of research data and related resources. In addition to providing a set of further services and activities revolving around the deposition process and data management in general, ARCHE's curators stand by with advice and assistance for researchers. A preformatted suggestion for citation includes a Handle link and allows researchers to download the reference in BibLaTeX format for persistent referencing.

ARCHE's extensive metadata, which aids researchers in finding and understanding the data, is stored in a dedicated metadata schema described with OWL. All metadata is freely available under CC0, for example via ARCHE's OAI-PMH endpoint, which provides a variety of metadata formats, including CMDI. Via OAI-PMH, all language resources are harvested by the Virtual Language Observatory (VLO).

For data, individual licences and one of three access modalities (public, academic and restricted) can be selected. Academic access is granted via an institutional login which is provided by the CLARIN ERIC and eduGAIN identity federation.

Data in ARCHE covers a wide range of humanities disciplines. In addition to linguistic resources such as dictionaries, Arabic corpora and audio recordings, the ever-increasing collections of ARCHE also include documentation archives from archaeological surveys, 3D scans of ancient objects, TEI annotated historical data or protocols, and born-digital data from the Digital Humanities. When it comes to languages currently represented in ARCHE, the archive stands out not only because of materials in English, French, German and Spanish, but also because of a large collection of resources in Arabic and some of its varieties (Algerian Saharan Arabic, Egyptian Arabic, Mesopotamian Arabic, North Levantine Arabic and Tunisian Arabic) and a few other languages like Dagbani, Persian and Yue Chinese.

Depending on the file type, a growing set of bespoke dissemination services allows researchers to preview, download, serialize or disseminate the file contents or its metadata. The dissemination services for each resource in ARCHE are displayed with clickable buttons.

Examples of dissemination services that visualize data include an online viewer for 3D files, which is based on the 3D Heritage Online Presenter (3DHOP) framework, and one for images, which is based on the International Image Interoperability Framework (IIIF). These two services are shown in Figure 20 with two example resources.

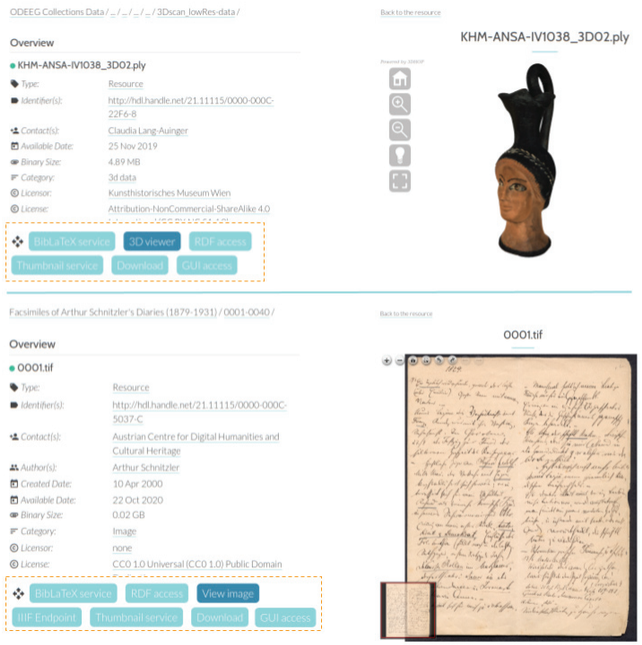



**Figure 20:** Dissemination services for viewing the file content online. Above: 3D viewer for an anthropomorphic Attic vase with the PID http://hdl.handle.net/21.11115/0000-000C-22F6-8. Below: IIIF based image viewer for a scan from Arthur Schnitzler's diaries with the PID http://hdl.handle.net/21.11115/0000-000C-5037-C.

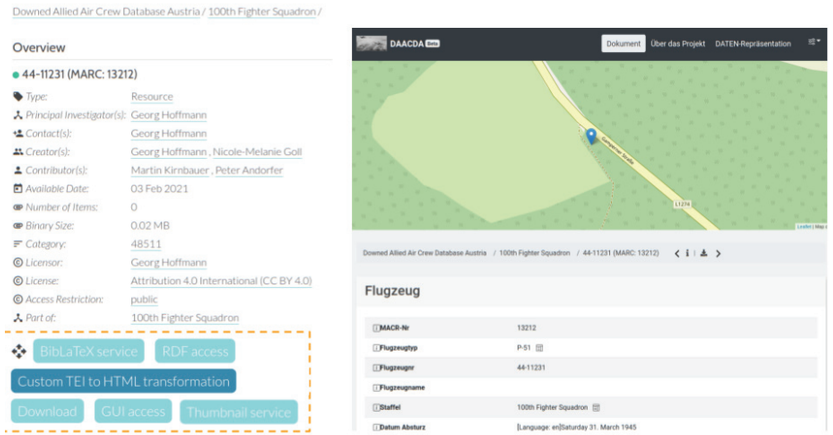



**Figure 21:** A TEI resource with information about a crashed WW2 airplane (https://arche.acdh.oeaw.ac.at/browser/oeaw_detail/104494) in ARCHE and its view with the dissemination service "Custom TEI to HTML transformation".

The dissemination services allow the re-use of data stored in ARCHE in dedicated web applications. In this way, data that is stored for the long term can not only be accessed via the ARCHE GUI, but can also be displayed in a project with a custom web application.

**Examples for such web applications re-using data from ARCHE include:**

- The facsimiles of the diaries of Arthur Schnitzler are only stored in ARCHE and can be accessed from the dedicated web application; for instance, in the entry for 1879-03-19, the link *Faksimile* displays an image of the diary facsimile directly fetched from ARCHE via the IIIF endpoint (Figure 22).
- The photographs, drawings and 3D models of ancient Greek vases are all fetched from ARCHE and then displayed along custom metadata in a dedicated web application for the ODEEG project.
- With the dissemination service that transforms TEI into HTML with a custom stylesheet, it is possible to create XML-based web applications with data solely held in ARCHE. This is still in its beta phase, but the Downed Allied Air Crew Database Austria can already give a first impression.[17] For instance, when the TEI resource with information about a crashed WW2 airplane (Figure 21) is viewed with the custom transformation from TEI to HTML, it is possible to navigate between the preceding and following entries within the collection without needing to go back to ARCHE.

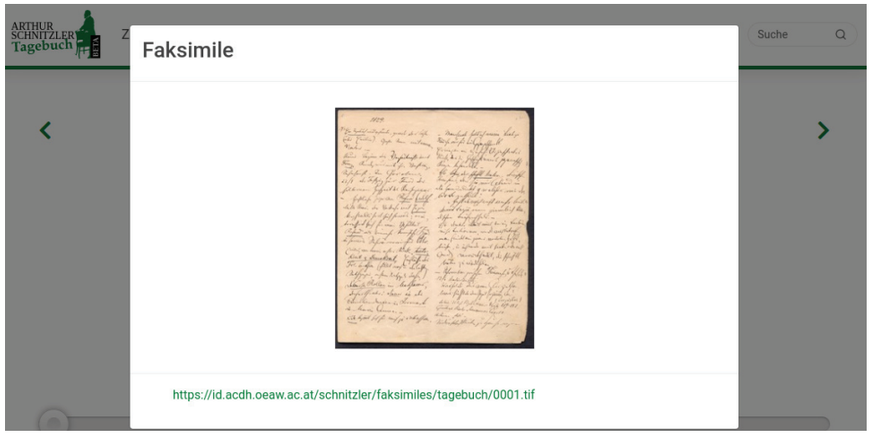



**Figure 22:** An image of a diary facsimile in a dedicated web application. The image is directly fetched from ARCHE via the IIIF endpoint.

[17] http://hdl.handle.net/21.11115/0000-000D-CA69-A

Some dissemination services forward data from ARCHE to external services, such as the Language Resource Switchboard by CLARIN. The example in Figure 23 shows how the TEI annotated protocol of the 34th session of the congress of Achen (1818) from the data collection Mächtekongresse 1818–1822. The digital edition can be sent to the switchboard and then be analysed with the Voyant Tools.

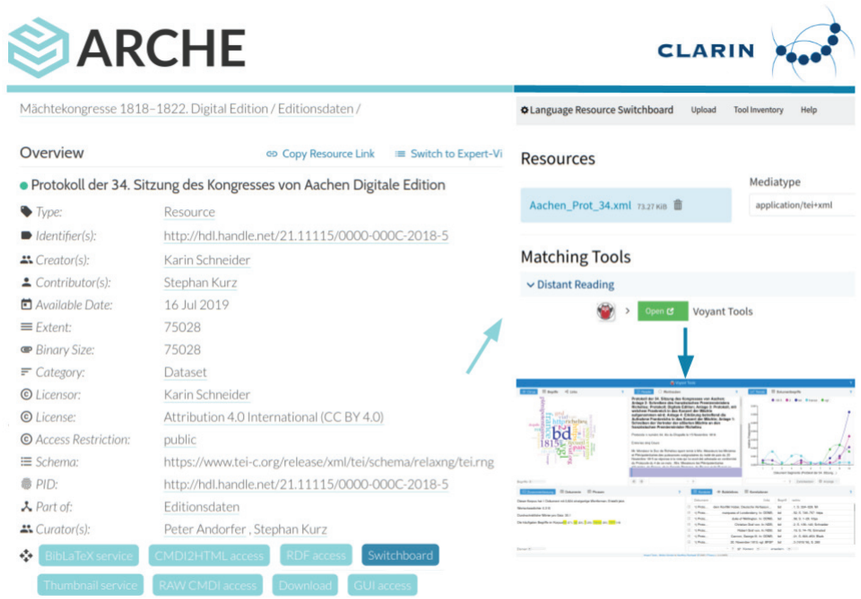



**Figure 23:** TEI Resources in ARCHE can be viewed in the CLARIN Language Resource Switchboard, which offers further tools for processing or visualization, such as Voyant Tools.

### Services by ACDH-CH

While ARCHE provides support and services revolving around the long-term preservation of digital resources, its host institution, the Austrian Centre for Digital Humanities and Cultural Heritage (ACDH-CH), provides a wider palette covering all phases of the data lifecycle. The ACDH-CH also pursues a digital humanities research agenda on its own. Current projects cover a wide range of humanities domains, investigating technical standards, infrastructure components, semantic technologies and text technological methods.

In addition to ARCHE, services offered by the ACDH-CH include the development of bespoke web applications and research software as well as re-usable tools for recurrent tasks. The portfolio of services is completed with personal consulting that aims at offering advice and guidance to the research community.

Prominent examples for tools developed at ACDH-CH include services for the wider CLARIN and DARIAH community, such as the CMDI Curation module to validate CMDI metadata records in the VLO, the DARIAH ELDAH – CONSENT FORM WIZARD to support humanities researchers in obtaining valid consent for data processing, the ACDH Vocabularies for persistent hosting of SKOS vocabularies, the open-source database OpenAtlas with automatic mapping to CIDOC CRM, and SACHA – Simple Access to Cultural Heritage Assets for viewing historical books, newspapers and postcards.

# Interview | **Peter Andorfer, Stephan Kurz,** and **Martin Anton Müller**

The following Tour de CLARIN interview is about ARCHE, the Austrian B-centre that is run by the Austrian Centre for Digital Humanities and Cultural Heritage (ACDH). The interview features Peter Andorfer, who is a research software engineer at ACDH; Stephan Kurz, who is a German studies scholar working with digital scholarly editing; and Martin Anton Müller, who is a German philologist.

L–R: Peter Andorfer, Martin Anton Müller, Stephan Kurz

### Please introduce yourself (your academic background and current position).

<

**Peter Andorfer:** I am a historian by training, and since 2015 I've been working as a research software engineer at the ACDH-CH – with a focus on developing data-driven applications, data wrangling and archiving.

**Stephan Kurz:** A German studies scholar by training, I developed a genuine interest in digital scholarly editing, also in connection to making available less well-known epistolary novels in the context of my dissertation. At the Austrian Academy of Sciences, I work with digital and hybrid scholarly editions – from archival sources to the long-term preservation of digital representations. The Institute for Habsburg and Balkan Studies currently has two ongoing major digital edition projects that I am involved in: the Minutes of Ministers' Councils of Austria and the Austro-Hungarian Monarchy, and an upcoming edition of sources of Ottoman–Habsburg diplomatic exchanges.

Martin Anton Müller: I am currently leading a research project on the online edition of the Austrian dramatist and author Arthur Schnitzler's (1862–1931) professional correspondence. I studied German philology and art education and then did my doctorate in philosophy. In the course of working on various scholarly edition projects over the last thirteen years – the latest two as project leader – the digital approach has become increasingly important.

>

### How are you involved with ARCHE?

<

**Andorfer:** I'm part of the ARCHE core team, which means most of the time I request new features, play around with those, find bugs and/or more things I'd like to have. On the other hand, I'm also a user of ARCHE, someone who wants to see data archived in some safe haven for eternity.
**Kurz:** ARCHE is already an important infrastructural part of the Austrian Digital Humanities field. Most projects I work with will eventually end up in ARCHE. This is also because of the projected sustainability of this solution in comparison with other long-term data storage solutions. I am simply an end-user of this system who deposits data there. Together with Karin Schneider, we deposited a mid-size scholarly edition of protocols, treaties and other genres related to the follow-up negotiations after the Vienna Congress, entitled Mächtekongresse 1818–1822, which you can find on ARCHE using a practical Handle link.[18] Of course, I am always curious to also discover new additions to the repository – there's lots to discover!

**Müller:** My self-image and my ideas of what I have to do as an editor are shaped by book production. When I make a book, I can assume with some certainty that in the year 2500 it will still be possible to read it. This is crucially different in the digital realm. The integration of my research data in ARCHE is the first time that I am trying to hand over data in such a way that it can be re-used for as long as possible and as well as possible, so that in the best-case scenario it will still be accessible in 2500 as well.

>

[18] http://hdl.handle.net/21.11115/0000-000C-2093-9

**Which ARCHE resource and/or tools have you used? How, concretely, did you integrate them in your existing research?**

<

**Andorfer:** I mainly use the ARCHE API and ARCHE Dissemination Services. A core feature of ARCHE is its granularity. Whereas solutions like Zenodo – which I like and use myself a lot – provide stable and resolvable IDs mainly on the collection level, ARCHE exposes each of its resources, be it collections, binaries, entities such as persons or organizations, through resolvable URIs. This makes it possible to integrate ARCHE resources into other applications or to develop your own clients that interact with ARCHE. To give an example, a previous version of ARCHE used a triple store as a storage layer, which made it possible to query ARCHE through SPARQL. But for recent ARCHE versions we switched to a PostgreSQL database, mainly for performance and maintainability reasons. By doing so we lost the SPARQL feature. But thanks to the existing ARCHE API, with its default response format being RDF, it is quite easy to write an API client in your favourite programming language that would fetch the data and throw it into a triple store.

**Kurz:** ARCHE to me is only part of the digital infrastructure that keeps evolving around our digital editions workflows – by nature and by design, it's concerned with the later stages in the lifecycle of digital objects. For data generation and data curation we use various other tools. But the need to be able to archive data in ARCHE in a meaningful way with metadata that goes beyond the coarse surface level that is normally collected with bibliographic records does in fact already help conceptually at a data modelling step that precedes all data collection. ARCHE's web interface shows exactly what metadata fields you have to fill out on the levels of project, collection and resource. In addition, there are several well-documented API endpoints that return metadata in various formats. It's a satisfying experience to have been tediously generating metadata in one format and having ARCHE return it through its OAI-PMH interface to another web application in another format. Furthermore, the semantic integration of statements (e.g. A said that person B has this other identifier C) makes a lot of sense to me, especially when those statements become identifiable as resources that have their own handle.net handles. With potentially infinite versions of a single data set, dealing with data resources quickly becomes more complex than a human can deal with. That said, ARCHE helps a lot, both before and after creating a digital resource. Yet the most important factor is the ARCHE team, experts who truly thought through what it means to host digital objects for the long term. They truly are there to help, cross-check, and integrate digital resources.

**Müller:** I have only been working with primarily digital data for just under three years. For their display, I have a web app from Peter, the DSE-base app, which simplifies it a lot to make the data – in my specific project, letters, telegrams and postcards – accessible. And Peter with the curation of Schnitzler's diaries has implemented a related use-case that I can track and copy. Frankly, ARCHE to me is more like one of several end goals next to the website and preparations for a possible print edition. I'm learning all the time, and at the moment I know how the data gets into ARCHE, but I'm not yet sure what I'm going to do with it yet.

>

**Are there any specific features of ARCHE tools and resources that make them especially well-suited for Digital Humanities Research, especially for non-technical users?**

<

**Andorfer:** In my opinion, digital humanities research needs some technical understanding or some tiny bit of data literacy. The beauty of ARCHE is its consequent usage of the RDF data model. All data ingested or requested is described in RDF. So in the end I don't need any fancy graphical user interface to read or process ARCHE's (meta)data; I just need to know RDF or at least how to read RDF – a well-established W3C standard.

**Kurz:** I particularly like that ARCHE forces you to think about your metadata schema in a way that makes metadata comparable across the resources that are stored. So my take on this is that the scholarly communities of various disciplines really need to think about the nature and state of their digital resources. This is a tedious process, but it helps to clear up lots of questions that researchers may eventually be asked about the provenance, creation, legal situation, metadata curation, etc. of their data. Kudos to the colleagues at ACDH-CH who have enabled us to rethink our metadata! The need for sufficient and in-depth metadata not only helps users discover and explore one's data, but it already reshapes how data creators and curators think about what they do. This may be a long process, but ultimately it's also about transparency in the research process itself.

**Müller:** What is relevant for me right now at the end of the project is the question of what happens after me. Two things are relevant. One is that ARCHE offers clear guidelines for citing resources that include authorial information as well as a persistent identifier for the resource that lasts for decades. The second is that the research results can be downloaded and further developed.

**Kurz:** Yes, the ARCHE dissemination services that Martin just mentioned are another key feature. This enables data depositors to create custom views of their data on the fly, which works without having a dedicated web application that they have to maintain separately, and allowing access through stable and persistent identifiers.

>

### What makes ARCHE especially important for the Austrian Digital Humanities and SSH research community?

<

**Andorfer:** It is a technical solid data repository hosted by a hopefully quite stable institution. So there is a possibility that ARCHE will not cease to exist soon and therefore can be used by the Austrian SSH research community and everyone else who wants to publish Digital Humanities research data. The ACDH-CH, as its name and mission suggest, is rather focussed only on Digital Humanities, so ARCHE has in a sense a wider scope in relation to data.

**Kurz:** One important thing to note is that, yes, ARCHE is an integral part of the Digital Humanities community, but this is only true in conjunction with other tools and services provided by the ACDH-CH.

**For me, coming from another institute that is also part of the Austrian Academy of Sciences, it is a luxury to have a CLARIN B-centre so close by, with all that comes with it. I suppose the ongoing digital transition of the Social Sciences and Humanities will lead to an ever increasing attention toward digital infrastructure in general, and ARCHE in particular.**

I can only underline what Peter said about the beautiful simplicity of the RDF data model – using such a low-level "grammar" makes it well-prepared for future challenges. ARCHE already stores a wide variety of different data formats, some of them relating to textual sources such as the collections that Martin mentioned, but others referring to 3D geometries, images, etc.

**Müller:** I think ARCHE's solution to have generic interfaces for multimodal data is pretty smart.

>

### Do you integrate ARCHE tools and resources into your university teaching? If so, how exactly do you integrate them?

<

**Andorfer:** The linguistic resources stored in ARCHE are especially regularly used for university teaching, for example in the Department of Near Eastern Studies or the Department of German Studies. Resources include the travel!digital collection (A digital collection of early German travel guides on non-European countries which were released by the Baedeker publishing house between 1875 and 1914), the Facsimiles of Arthur Schnitzler's Diaries (1879–1931), the VICAV – Vienna Corpus of Arabic Varieties or the amc: Austrian Media Corpus. For many of the archived datasets, dedicated web applications developed and hosted by the ACDH-CH exist. For instance, the travel!digital corpus allows researchers to explore the thesaurus and view the facsimiles of the travel guides in combination with their transcribed texts, Schnitzler Tagebuch provides a bespoke view on the facsimiles and the persons, places and dates mentioned in Arthur Schnitzler's diaries and VICAV presents further materials for near-eastern language research. Other use cases of data stored in ARCHE include the virtual Hackathon from 2018. Most of the data in ARCHE is public and open to be used in teaching everywhere around the world, even though we can't tell where… By the way, DH courses can be registered in the Digital Humanities Course Registry (another CLARIN resource ACDH-CH is providing)!

**Kurz:** During the COVID pandemic, the necessity of sustainable data and metadata hosting was becoming clear as all those web services and platforms for e-learning were discussed… Luckily, I did not have any teaching duties [laughs]. In all seriousness: I think the takeaway message for DH teachers, and also for people who design DH curricula, may be to include digital preservation early on – and to involve students in the actual data lifecycles in a way that they really get a feeling how important, but also how tedious, it is to produce meaningful documentation and metadata. From my own experience in making DH a desirable subject area, I see a focus on data creation and data curation, fancy web application development and the like, but the "boring" things that ultimately keep past projects accessible and reusable are often overlooked.

**Müller:** I did not have any teaching duties either.

>

## What's your vision for ARCHE 10 years from now?

<

**Andorfer:** All ARCHE resources can still be requested via their URIs and that their metadata can be read and interpreted by machines as well as human beings. This may not sound like a big vision, but after all, that's what a long term data repository is built for, storing data in a persistent, findable and resolvable manner.

**Kurz:** I'm sure that ARCHE will be alive and kicking in one way or another – that's why it has been created. Web applications may come and go, but data repositories are intended and designed to be around for the longer term. Its backend may change in the future, maybe even to a successor system that could deal faster with the vastly larger amounts of data that I'd expect to be stored in this infrastructure; its frontend and API may change and improve, ARCHE may be adapted to new file and metadata formats – but exactly such processes of regularly updating and curating are core functions of such a system – apart from versioning and redundant data storage under transparent circumstances and policies.

**Müller:** I always find it difficult to predict the future because I always fall back on analogies. In this case, I hope that the handling of research data in ARCHE will become easier in the same way that the internet went from being something for nerds to a mass medium. It is not so important whether ARCHE develops to something simpler or whether more and more users acquire the necessary skills. I imagine both will take place.

>

# The CLARIN-PL B-Centre

## Introduction

Written by **Krzysztof Hwaszcz**

The Polish consortium CLARIN-PL,[19] which is a founding member of CLARIN ERIC, operating since 2012, was already presented in Tour de CLARIN in 2018 (Volume 1). Not much has changed since then in terms of organization and scope: the centre is based at the Wrocław University of Science and Technology (Department of Computational Intelligence) and is still coordinated by Maciej Piasecki. From the user's point of view, the B-centre constitutes the pillar of the Polish CLARIN infrastructure, as it is focused on the development and the maintenance of the existing services from the scientific and technical perspectives. The main objective of the centre is to maintain and make available a unique infrastructure for the computing equipment and data repositories integrated with the European CLARIN infrastructure, as well as natural language processing tools and services. Although CLARIN-PL focuses on Polish, the centre is open for cooperation with other languages as well. Thus far, it has incorporated over 10 languages in our tools and services.

The mission of the B-centre is not only to maintain and promote, but also to develop tools and services that are already available within the CLARIN-PL infrastructure, such as the Polish wordnet plWordNet, the valency lexicon Walenty, and the topic modelling tool Topic, and to create completely new tools and resources. The development model is bottom-up – that is, the directions of activities at the centre respond to user demand. For instance, there has been a recent increase in demand for processing texts that were previously transcribed using optical character recognition. Automatic conversion typically introduces various misspelling errors. As a result, the centre provided a set of new standardization tools which are designed to correct such documents. Similarly, there is an increasing demand to work with texts originally containing personal data – in order to avoid legal difficulties, these texts need to be anonymized; consequently, the Anonimizer service has been developed, which aims to overcome these issues.

Some other tools and resources that have been recently developed may be categorized into four groups: (i) corpus tools, (ii) tools for sentiment analysis, (iii) text standardization and (iv) multilingual use. Let us present them briefly below:

**New tools for working with text corpora**

- Topic for topic modelling
- ComCorp for comparing the linguistic features of corpora
- Cat for simple text classification

**Sentiment analysis**

- Sentemo for determining the text sentiment
- Wydźwięk for the analysis of emotional overtone
- MultiEmo for multilingual sentiment analysis in 11 languages

**Tools for text standardization and correction:**

- Paragraph for dividing the text into sentences or paragraphs
- Symspell for removing redundant spaces
- Wordifier for abbreviation expansion
- Txt Clean for document cleaning
- Speller for improving the spelling
- Punctuator for improving the punctuation
- Anonimizer for removing sensitive data

[19] https://clarin-pl.eu/

**Implementation of new languages in existing services**

- WebSim for detecting text similarity and clustering (Polish, English)
- InterLem for processing literary texts (Polish, English, German, Spanish, Russian)
- Topic ML for topic modelling (Polish, English, German, Spanish, Hungarian, Russian)
- WebSty ML for stylometric analysis (Polish, English, French, German, Spanish, Hungarian, Russian, Hebrew)
- MultiEmo for multilingual sentiment analysis (Polish, English, Chinese, Italian, Japanese, Russian, German, Spanish, French, Dutch and Portuguese)

An integral part of the CLARIN B-centre is the Knowledge Centre for Polish Language Technology (PolLinguaTec) founded in 2017. It has been involved in the implementation of language services in research projects. The main objective of PolLinguaTec is to provide knowledge on the application of tools and systems for natural language analysis. The scope of research conducted with the use of the CLARIN-PL infrastructure includes such areas as Economics, Political Science, Sociology, Social Psychology and Linguistics, to mention only a few. A more detailed description of the CLARIN K-centre was published in the 3rd volume of Tour de CLARIN. PolLinguaTec provides experts able to solve problems related to the use of language processing resources. Apart from that, a number of instructions, guidelines and tutorials have been created to enable or to facilitate the autonomous application of the CLARIN tools by our users. To maximize the efficiency of the CLARIN initiative, both centres (B and K) are oriented towards operating in close collaboration. Overall, since November 2019, PolLinguaTec has helped to plan the implementation of the CLARIN infrastructure in about 35 research projects; the authors of 15 of the projects have also benefited from our support in preparing their grant applications.

For instance, Agnieszka Hess collaborated with PolLinguaTec in the project on the identification of social representations of civil dialogue in Poland and the description of intended and unintended consequences of activities of civil dialogue participants.

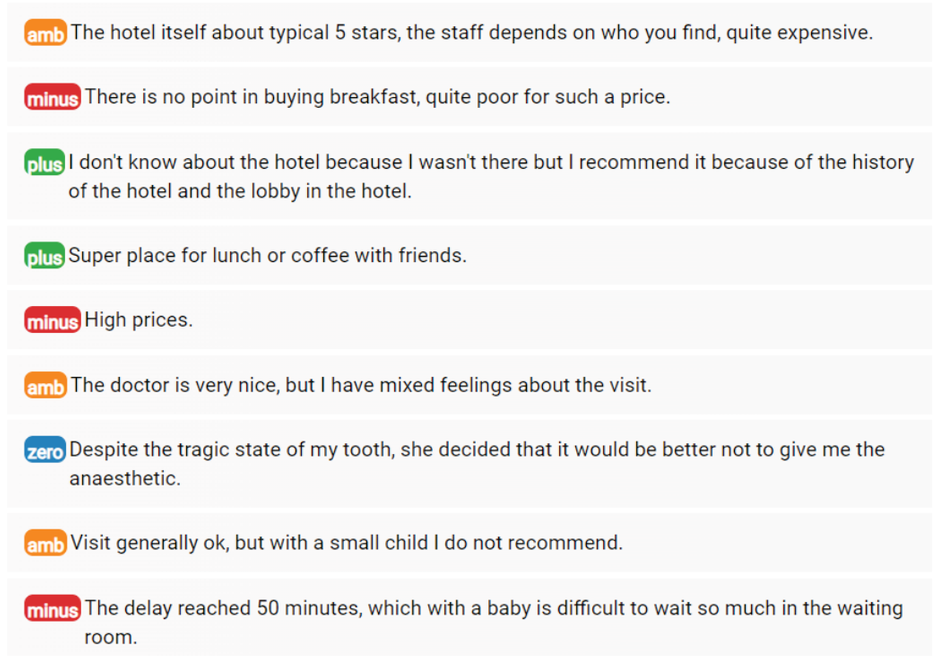



**Figure 24:** Sentence-level sentiment analysis using the CLARIN-PL tool MultiEmo, which detects 4 types of sentiment – negative, neutral, ambivalent, and positive.

Over 42,000 plenary sessions, commissions and interpellations were extracted from the Polish Parliamentary Corpus. PolLinguaTec provided the tools (Topic, TermoPL, MeWeX, Sentemo, Wydźwięk, MultiEmo) and assistance with the analysis of the frequency of certain terms, with topic modelling, with the identification of domain-specific terms and their contexts and with the sentiment analysis of participants' statements in civil dialogue.

Another research project was conducted in collaboration with the Educational Research Institute (IBE), which carries out interdisciplinary research into the functioning and effectiveness of the education system in Poland. It was assumed that the descriptions of learning outcomes can provide a basis for comparing qualifications. The research team needed a tool which would automatically compare the qualification with the use of NLP techniques. The work of PolLinguaTec was divided into the following stages: (i) developing a qualification classifier; (ii) creating labels for fields common in the entire data set; (iii) describing learning outcomes; (iv) clustering; (v) preparing the interface for target users. The results of the cooperation between PolLinguaTec and both Agnieszka Hess as well as IBE team were presented during the conference "CLARIN-PL-Biz – language technologies for learning and business II".

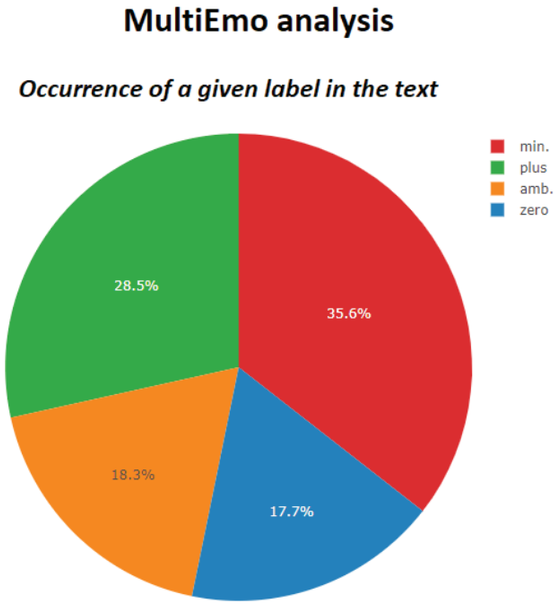



**Figure 25:** Visualization of sentiment distribution using MultiEmo, where 35.6% of the occurrences are negative, 28.5% positive, 18.3% ambivalent, and 17.7% neutral.


**Figure 26:** Querying the historical ChronoPress corpus, which includes texts from the time of the Polish People's Republic, using the CLARIN-PL tool ComCorp.

In relation to Sociology – or Social Media, to be more precise – PolLinguaTec helped in the ComAnCE (Combat Anti-Semitism in Central Europe) project supervised by Viera Žúborová, which aimed to present the phenomenon of discrimination in the Visegrád Group countries (Czechia, Hungary, Poland, and Slovakia) to European organizations, state authorities, police forces, fellows researchers and institutions. The participating countries developed a unique categorization of anti-Semitic statements classified according to the keywords collected from Facebook users. The concordancer (KonText) and programming tools (CMC and Morphodita, used to correct errors) from our centre and from the CLARIN infrastructure were among the software used in the implementation of the project.

# Interview | **Olga Czeranowska**

Olga Czeranowska is a sociologist working in the field of migrant studies. She and her team have performed sentiment analysis of Polish Twitter posts with the cooperation of CLARIN-PL.

**Please introduce yourself – your academic background and current position.**

<

I did my PhD at the University of Warsaw in the Institute of Applied Social Sciences. The topic of my thesis was occupational prestige from an individual perspective, and I was interested in how people with highly prestigious positions on the labour market feel their prestige and how it affects their professional and private biographies. Since April 2020 I have been working at the SWPS University of Social Sciences and Humanities.

My research interests lie primarily in the area of the sociology of work and migration studies. In the near future, I plan to focus more on the interconnections between geographical and occupational mobility, migrants' careers, and the concept of success in the migrants' occupational trajectories.

As for research methodologies, I have experience with both qualitative and quantitative data, as well as using a mixed-methods approach. I am very much interested in trying new methods of gathering and analysing data in order to see social realities through new lenses.

>

**Your research is rooted in the social sciences and focuses on the field of migrant studies. What are some of the topics you are currently exploring in this field? Could you briefly present them?**

<

We are currently analysing data obtained in the project *(IT)Mobility. Immobility of the mobile, mobility of the immobile – migrants in the times of pandemics and new information/communication technologies*.[20] In this project, we are examining the effects of the COVID-19 pandemic on the different spheres of the lives of Polish migrants. The project is based on the broad understanding of mobility, including not only geographical mobility but also virtual mobility. We are also analysing occupational mobility in connection with geographical mobility. Since the beginning of the pandemic, geographical mobility has in many ways been restricted in a way that is unprecedented in recent decades. Some of this "mobility energy" has been shifted towards virtual mobility as many spheres of everyday lives have moved online. This applies not only to private lives, such as meeting family and friends via Zoom or Google Meet, but also to the work sphere and services, such as online classes, working from home, and telehealth. Many of those possibilities existed before, but we can safely assume that the pandemic was an accelerator of their development and more general use.

We assumed that migrants are a kind of extreme case for the analysis of different forms of mobility and immobility during the pandemic – not only did they move to another country, but they generally had more experience with virtual mobility as well. They were keeping in touch with their family and friends or used services like Zoom even before the pandemic, in order to participate in cultural events or use services in their country of origin. In addition, the migrants' situation during the pandemic itself, especially with lockdowns and other restrictions, was unique, as many of them were crucially relying on mobility between their country of residence and country of origin to visit their loved ones or use certain services.

>

[20] The project is financed by the Ministry of Science and Higher Education in Poland under the 2019–2022 program "Regional Initiative of Excellence", project number 012 / RID / 2018/19.

**You are currently performing sentiment analysis of Twitter posts with the cooperation of CLARIN-PL. How did you start collaborating with CLARIN-PL?**

<

Our team took part in a CLARIN-PL online workshop on Natural Language Processing in November 2020 that included general information about their tools as well as some real-life examples of their application in other research projects. After the event, we contacted CLARIN-PL via their online form for researchers who need consultation. We then met with the CLARIN-PL team and discussed what would be the best way to approach our data. It was perfect timing for us, because our project started in September 2020 and at this point, within the qualitative component, we were just beginning to collect the Twitter data while simultaneously looking for the best way to perform the quantitative analysis. Using CLARIN-PL tools gave us a solid methodological foundation for this part of our project.

>

**How are you performing the sentiment analysis? Which CLARIN-PL tools are you using for this task? How is the sentiment labelled? How are you sampling the tweets?**

<

For our first paper, which is currently being prepared, we are using CLARIN-PL's MultiEmo.[21] MultiEmo is a tool for sentiment analysis that is available in 11 different languages including Polish. It uses a manually annotated corpus of consumer reviews as its training set and labels tweets in terms of four sentiment values: positive, negative, ambivalent, and neutral.

We are working with two datasets. The first one consists of tweets gathered on the basis of hashtags that were identified as being relevant for the project. The hashtags mostly relate to pandemic concepts, such as *#lockdown*, *#stayhome*, and *#workfromhome*, and are both in Polish and English. The second dataset consists of tweets by users who were identified as Polish migrants.

Sampling was one of the biggest challenges for our project, as we specifically wanted to access the tweets of Polish migrants. Unfortunately, only a small percentage of Twitter data is geotagged, so we had to find another way to filter them. Our strategy was then to gather tweets from users whose location (given by the user and annotated manually by our team) is in a country other than Poland, but they are using Twitter in Polish. Additionally, we filtered the database to exclude bots. We are aware that this is not a perfect solution, but we think that, taking into consideration the database-related constraints (missing values, people giving fictional places as their location, etc.), it still gives us a solid sample of at least one kind of Polish migrant – a person who considers their stay abroad permanent enough to change their Twitter location, but still has some ties with the home country (for instance, such a migrant uses Polish because of who will be reading their feed) or did not learn the receiving country's language well enough to feel comfortable using it on social media.

>

**What does the sentiment of tweets reveal about migrant mobility in relation to the ongoing COVID-19 pandemic?**

<

We are still at the stage of data analysis, but what we are hoping to see are some longitudinal patterns. With data covering a period that is relatively long in relation to social media (we started gathering tweets in January 2021), we want to analyse how the Twitter discourse has changed over the last year. We will be looking both at the popularity of particular hashtags connected with the pandemic and its consequences, such as the aforementioned *#lockdown* and *#workfromhome*, and also at the sentiment connected with those hashtags.

In further steps, I hope that the two (IT)mobility datasets will also be analysed together with some project-external datasets that include the number of COVID-19 cases in particular countries, the number of people vaccinated or data on mobility restrictions in Poland or the receiving countries. This would give us an opportunity to analyse how changes in the offline world, such as the introduction of the COVID-19 vaccine, influenced the attitudes of the online Twitter community.

We would also like to compare the Twitter presence of our subsample of Polish migrants with that of the general population of users, as well as – if the subsamples turn out to be numerous enough – between the subsamples of Polish migrants in various locations.

>

[21] https://clarin-pl.eu/index.php/en/multiemo-en/

**How is CLARIN-PL supporting you in this task? Who are you collaborating with?**

<

**The CLARIN-PL team is supporting us with the use of the sentiment analysis tool MultiEmo by adding annotations to the tables in which our data is stored. Further analysis and data visualization is carried out by our team with Power BI.**

**We are currently in the process of preparing the first paper based on the sentiment analysis of the Polish migrants' tweets, with the help of Krzysztof Hwaszcz, Jan Kocoń, Piotr Miłkowski and Jan Wieczorek. We certainly hope to work together more in the future, both with the (IT)mobility dataset and within other, upcoming projects.**

>

**Why is it important to take a computational approach, such as sentiment analysis (or more broadly text analytics/natural language processing), in the social sciences?**

<

From a practical standpoint, sentiment analysis methodologies are extremely useful in dealing with big datasets, such as social media datasets. Social media is now a crucial part of the everyday lives of people living in contemporary society, so, naturally, they are becoming a more and more important source of data for social researchers. Social media datasets are real-time reactions to important events, so they can both provide us with data relatively quickly and enable us to analyse social phenomena over time. In the case of migration studies, what is very useful is the possibility of gathering the data internationally and within the context of the various locations of the social media platforms (although this can still be difficult because of the location issues that I have mentioned before). However, this kind of data comes with its own set of challenges, such as the large size of the datasets. Luckily, with sentiment analysis tools such as MultiEmo, we are able to overcome such problems and analyse social media discourse in a standardized way.

>

**What can CLARIN-PL do to further support digital humanities and social sciences researchers working with topics in migrant studies?**

<

I think that further events such as the December webinar are crucial so that social researchers know that such possibilities exist and are within reach. The (IT)mobility team took part in a CLARIN-PL online conference "CLARIN-PL-Biz – language technologies for learning and business II" in July 2021 to present our project as an example of research using CLARIN-PL tools. This conference was aimed at presenting academic and commercial use of the CLARIN-PL infrastructure. I was very happy that we could contribute to this event, and I hope that our example may have inspired some other migration researchers to take a look what CLARIN-PL has to offer in terms of both tools and research support.

>

# COLOPHON

Coordinated by
**Darja Fišer** and **Francesca Frontini**

Edited by
**Jakob Lenardič, Francesca Frontini,** and **Darja Fišer**

Proofread by
**Paul Steed**

Designed by
**Tanja Radež**

Cover image
**National and University Library of Iceland** (row 2, image 3)

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

Publication number
**CLARIN-CE-2021-1975**
**November 2021**

ISBN/EAN
**9789082990935**




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 2021