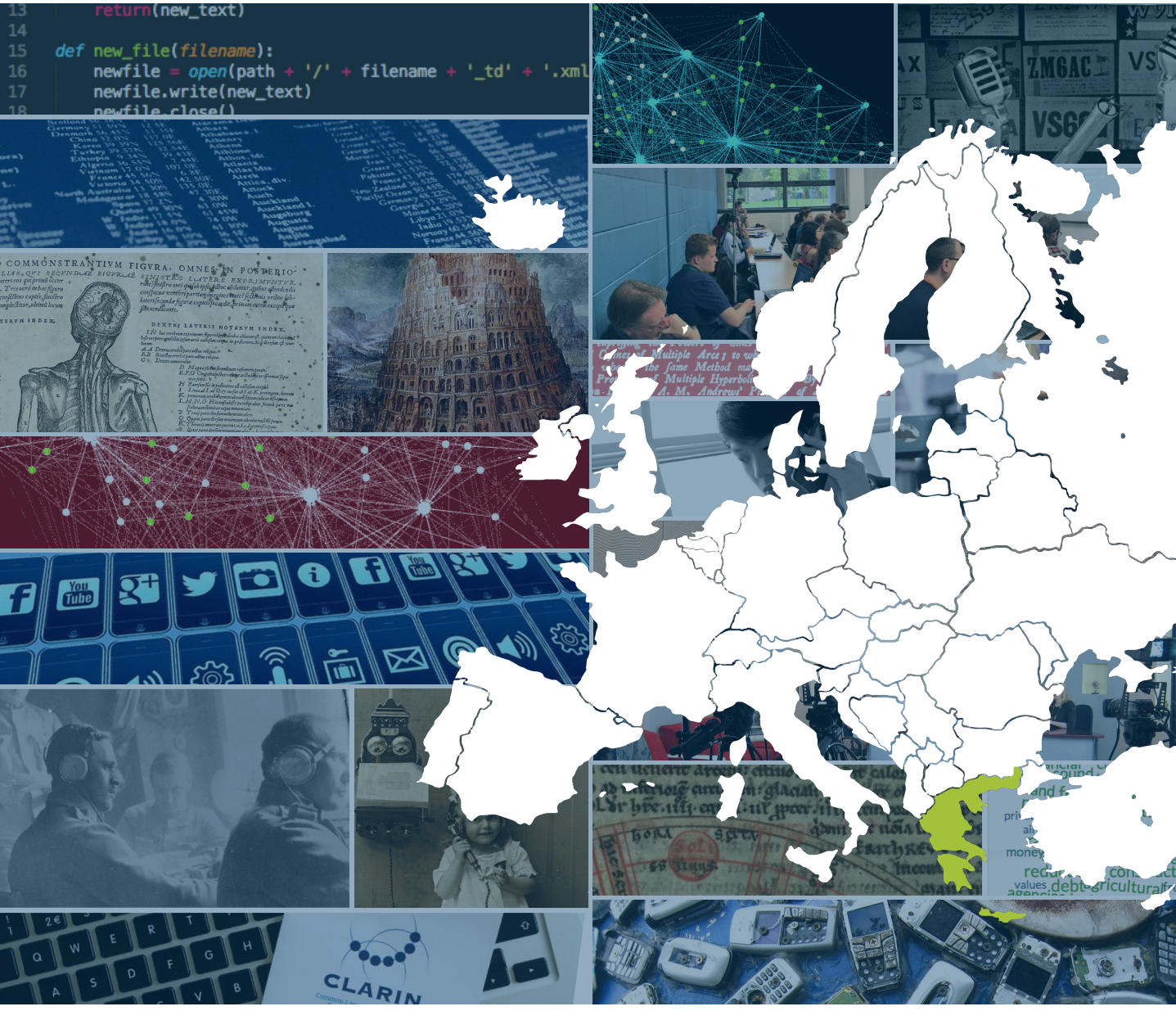
CLARIN

Common Language Resources and Technology Infrastructure

Tour de CLARIN

Greece



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

Foreword

Tour de CLARIN highlights prominent user involvement activities of CLARIN national consortia with the aim to increase the visibility of CLARIN consortia, reveal the richness of the CLARIN landscape, and display the full range of activities throughout the CLARIN network that can inform and inspire other consortia as well as show what CLARIN has to offer to researchers, teachers, students, professionals and the general public interested in using and processing language data in various forms.

This brochure presents Greece and is organized in five sections:

- Section One presents the members of the consortium and their work
- Section Two demonstrates an outstanding tool
- Section Three highlights a prominent resource
- Section Four reports on a successful event for researchers and students
- Section Five includes an interview with a renowned researcher from the digital humanities or social sciences who has successfully used the consortium's infrastructure in their research

Perimetriki Patron, Achaia, Greece
photo by Jason.Blackeye | Unsplash

Greece

Written by Darja Fišer, Maria Gavriilidou and Jakob Lenardič

The Greek network clarin:el[1] has been a member of CLARIN ERIC since February 2015. It was founded by three Greek research institutions:

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

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

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

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

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

Clarin:el currently contains around 500 language resources and 35 tools, which can be accessed by registered and non-registered users, in full compliance with the licence terms defined by the resource providers. Many of the language tools in the inventory are offered as web services for processing content in Greek as well as other languages, which means that researchers can use them to process their data directly through the inventory: users can either select resources from the clarin:el inventory to process, or they can upload their own data for processing. The outcome of the processing constitutes a new resource which can directly be added to the inventory, accompanied by automatically created metadata. Statistics related to the use of the resources (such as number of views and downloads) as well as dynamic recommendations of related resources and services (such as similar resources viewed by other users) are available to all users. Organisations that are members of the clarin:el network have the ability to set up their own repository within the infrastructure, which can then be accessed through the central inventory. Individuals who join the clarin:el network may store their resources at a dedicated repository, the so-called Hosted Resources Repository, which is also available for the storage of resources provided by organisations which do not wish to maintain their own repository.

The Clarin:el Team

The Greek consortium also actively promotes user involvement. Recently, on 27 June 2018, the Greek consortium organised an event intended to deepen the dialogue with digital humanities and social sciences researchers, better understand their requirements and familiarise them with the clarin:el infrastructure. On the one hand, the event featured an interactive session where the researchers had the opportunity to present their own research questions and experiences with using language technologies to clarin:el experts, while on the other, a hands-on session was organised, where the researchers were able to familiarise themselves with the clarin:el inventory and the use of its resources and tools. You can read more about the event on page 9.

Kos, Greece | photo by Mico59 | Pixabay

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

GrNE-Tagger

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

The GrNE-Tagger[2] is a tool available through clarin:el that automatically recognises proper names (Named Entities) in Greek texts and classifies them into one of the following five category types:

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

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

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

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

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

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

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

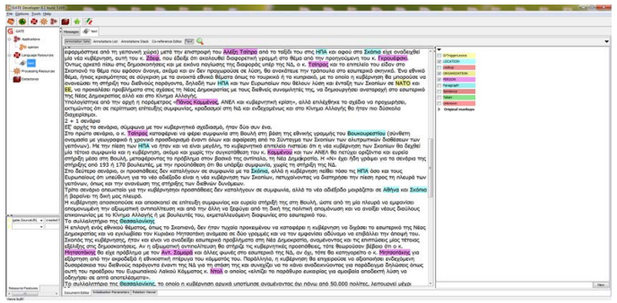


Figure 1: The output of GrNE-tagger (using GATE as a visualisation tool), in which different NEs are marked with different colours.

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

The Hellenic Parliament Sittings and Hellenic Parliamentary Corpus H-ParCo

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

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

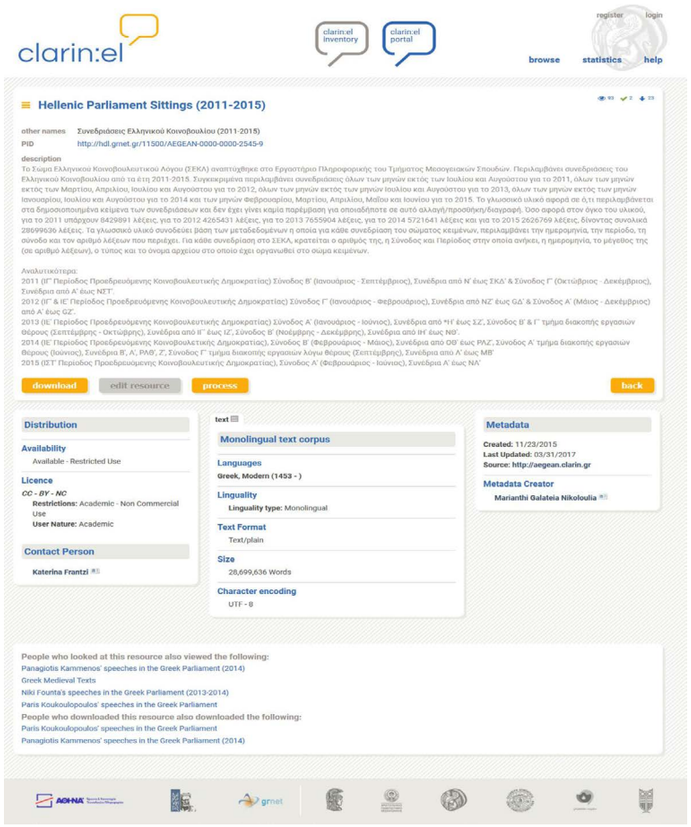


Figure 2: Greek parliamentary corpus in the clarin:el repository.

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

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

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

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

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

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

References:

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

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

Language Data and Technologies in Social and Political Sciences

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

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

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

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

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

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

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

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

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

Vassiliki Georgiadou

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

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

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

My research interests focus on political behaviour, far right parties, populism, radicalism and political extremism. I was principal investigator of the XENO@GR research project,[4] which examines xenophobia in Greece during the economic crisis, based on a computational social sciences approach; currently, I am co-investigator of a research project that examines the different expressions of violence in Greece from 2008 onwards (London School of Economics (LSE)-Hellenic Observatory Grants).

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

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

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

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

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

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

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

Figure 3: The workflow for creating the event database in the project.

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

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

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

Figure 4: The main targets of verbal aggressiveness in Greece in terms of national/ ethnic background, based on the analysis of Greek newspapers during the XENO@GR project.

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

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

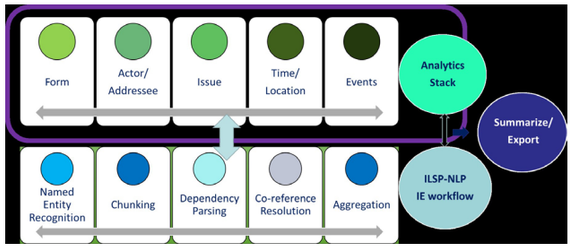


Figure 5: The NLP tasks performed on the XENO@GR data.

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

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

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

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

8. Could you share your experience with the recent clarin:el event for social sciences and humanities researchers (presented on page 9)? How are such events valuable to your research community?

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

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

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

COLOPHON

This brochure is part of the 'Tour de CLARIN' volume I (publication number: CLARIN-CE-2018-1341, November 2018).

Coordinated by
Darja Fišer, Jakob Lenardič and Karolina Badzmierowska

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

Edited by
Darja Fišer and Jakob Lenardič

Proofread by
Paul Steed

Designed by
Karolina Badzmierowska

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

Publication number
CLARIN-CE-2018-1341
November 2018

Contact
CLARIN ERIC
c/o Utrecht University
Drift 10, 3512 BS Utrecht
The Netherlands
www.clarin.eu

CLARIN
Common Language Resources and Technology Infrastructure

CLARIN 2018