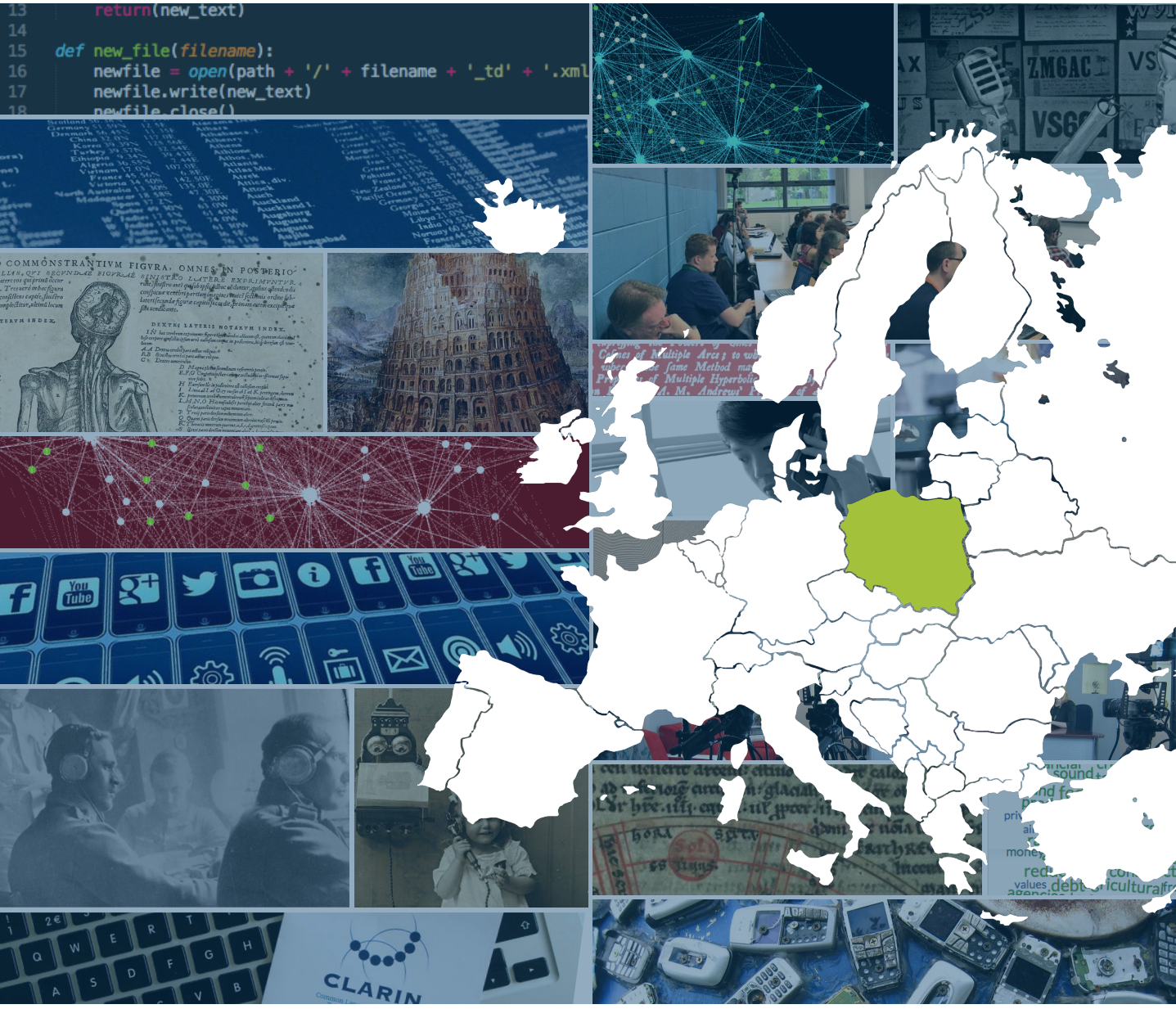
CLARIN

Common Language Resources and Technology Infrastructure

Tour de CLARIN

Poland



Written by Jan Wieczorek, Ewa Rudnicka, Agnieszka Dziob, 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 Poland 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

Wroclaw, Poland | photo by Wilfried Besser | Pixabay

Poland

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

The Polish consortium CLARIN-PL[1] is a founding member of CLARIN ERIC and has been actively involved in its operations since the very beginning in 2005. It comprises six member institutions:

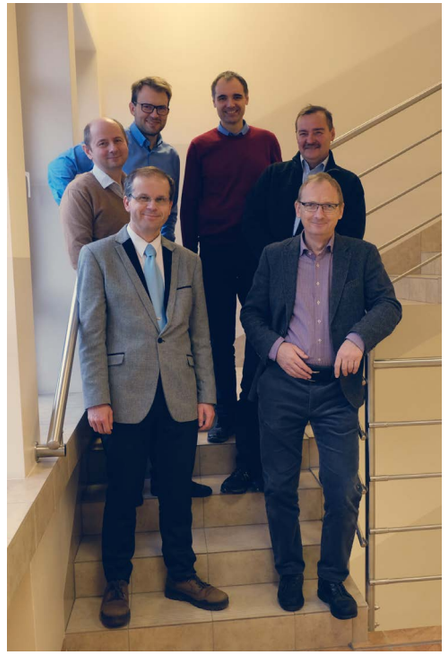
- Wroclaw University of Technology (Language Technology Centre);
- Institute of Computer Science (Polish Academy of Sciences);
- Institute of Slavic Studies (Polish Academy of Sciences);
- Polish - Japanese Academy of Information Technology;
- Lodz University; and
- Wrocław University.

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

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

- plWordNet, which is the biggest wordnet in the world available through the open licence together with its mapping to Princeton WordNet. It includes emotive annotation and was built in close collaboration with the valency dictionary Walenty. Read a more detailed presentation of plWordNet on page 9;
- DSpace repository, which is a large library of linguistic data and tools;
- SPOKES, which is a corpus of conversational data;
- Chronopress, which is a chronological corpus of Polish newspaper texts;
- Websty, which is a tool for the extraction of stylometric data. Most tools and resources work in the user-friendly web service technology (it does not require any software installation on the user's computer). A detailed presentation of WebSty can be read on page 6; and
- various speech recognition tools, such as Align.

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



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

The second goal of CLARIN-PL is to raise awareness and popularise knowledge about NLP among the Polish digital humanities scholars. To this end, the Language Technology Centre has been organising a series of workshops called "CLARIN in research practice" (see page 11 for a detailed description). The consortium is also a strategic partner in many large research projects: employees of the consortium advice on the optimal use of the existing NLP tools and resources and help plan research, which gives them the opportunity to collect opinions and information about researchers' needs. In November 2017 at Wroclaw University of Technology, PolLinguaTec, a CLARIN Knowledge Centre for Polish Language Technology (Clarin K-Centre), was established. Its task is the continuation of user involvement activities.

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

WebSty, an Open Web-Based System for Stylometric Analysis

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

WebSty[2] is a powerful web-based system for stylometric, semantic and comparative analysis of texts. In its current implementation, the system is suited for the quantitative analysis of German, Polish, English, Hungarian, Russian and Spanish texts and is presented as an easy-to-use web interface that enables researchers to simply drag and drop the documents they want to analyse or provide links to uploaded .zip files containing the documents (Figure 1). WebSty is also integrated with the Polish D-Space based repository provided by CLARIN-PL. To ensure fast processing of the documents, WebSty is designed as service-oriented software in which each language tool runs as a separate process with pre-loaded data models. The English version of WebSty makes use of the following tools:

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

After uploading the file to be analysed, researchers can use the Choice of features tab (Figure 2) to specify which linguistic features WebSty takes into account when performing the analysis. Among others, these include the specification of various grammatical classes and a host of features related to named entities. The results of the clustering are primarily visualised in the form of a dynamic dendrogram (Figure 3), which is generated on the basis of the D3.js library and involves an interactive binary tree where each subtree can be collapsed. In addition, WebSty allows researchers to download the results in the .xslx format and also to visualise them with other user-friendly methods, like a heat map, radar chart and multidimensional scaling.

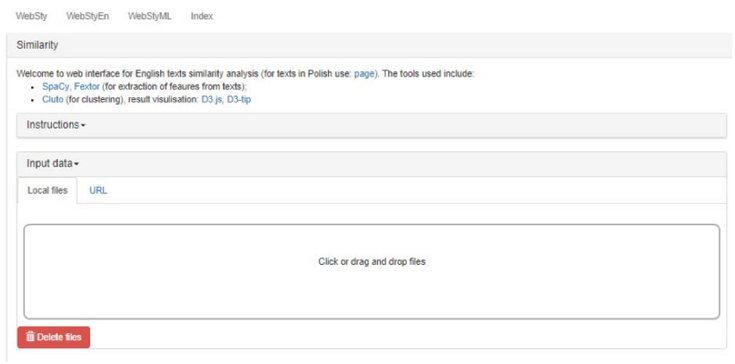


Figure 1: Uploading datasets in WebSty, where in the case of the English version researchers can either upload their own local documents or provide links to online resources. The Polish version is also integrated with the D-Space repository provided by CLARIN-PL.

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

Since WebSty does not require in-depth computational knowledge, it is a crucial tool for fields in the social sciences and digital humanities in that it allows researchers to conduct massive-scale analyses of numerous resources, thus revealing characteristics that have been overlooked by traditional approaches. As an example of a successful application in literary studies, Maciej Maryl, who is Deputy Director at the Institute of Literary Research of the Polish Academy of Sciences, used WebSty to analyse a large collection of blogs with anonymous authorship and thereby detected subtle similarities between documents on the basis of the provided clustering options (the interview with Maryl can be read on page 12). As a successful application in sociology, Marek Troszyński from Collegium Civitas has used the tool in a project for monitoring and documenting manifestations of discrimination against the Ukrainian minority in Poland. In relation to languages other than Polish, WebSty has successfully been used by Palkó Gábor from the Petőfi Museum of Literature to analyse texts in Hungarian (Figure 4). Through cooperation with partners from the same museum, a new version of WebSty will be created with a dedicated interface in Hungarian.

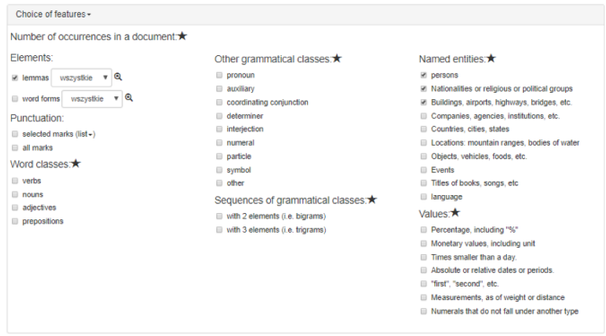


Figure 2: Choosing linguistic features for analysis.

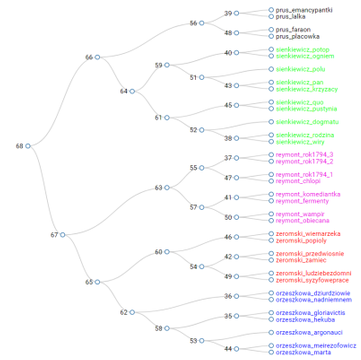
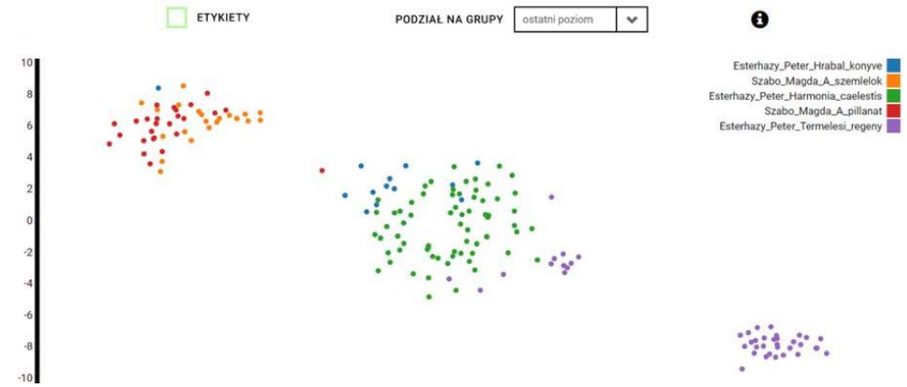


Figure 3: Clustering results (dendrogram and cluster membership) in a form of interactive dendrograms for corpus of Polish books.

Figure 4: Using WebSty to analyse Hungarian text. The visualisation shows clusters of similar texts scaled to 2D space.



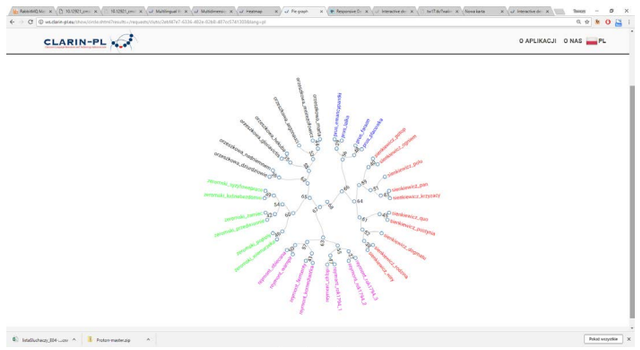


Figure 5: Clustering results (dendrogram and cluster membership) in a form of circle (in the presented results two clusters were selected in contrast to results in Fig: 18 where five clusters were selected).

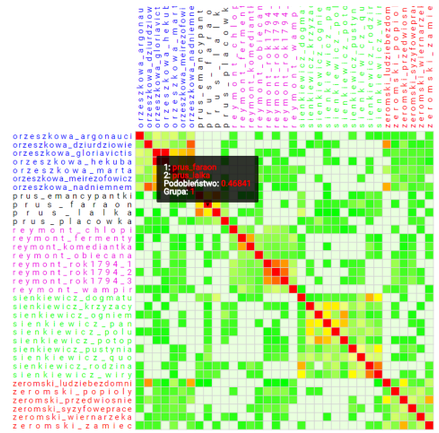


Figure 6: Similarity results in the form of a heatmap.

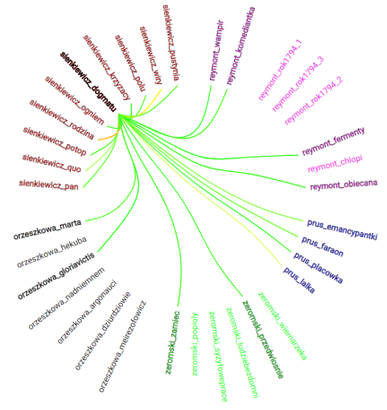


Figure 7: Similarity results in the form of a schemaball.

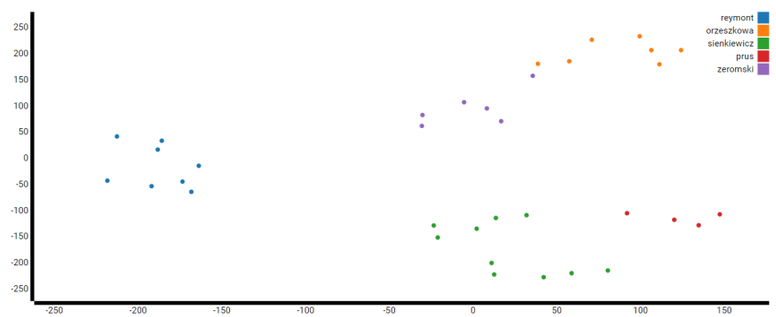


Figure 8: Distance results in the form of 2D plot.

plWordNet

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

plWordNet (Polish Słowosieć)[3] is a (large) lexico-semantic network reflecting (the current content and structure of) the Polish lexical system. It is a kind of dictionary in which word senses are represented by lexical units, linked by relations to create synonym sets – synsets. It is inspired by the Princeton University WordNet – the very first wordnet, which has been in development since the 1980s. Both wordnets are linked via inter-lingual relations, effectively creating a bilingual semantic network. plWordNet has been developed at Wrocław University of Technology by a team of linguists and programmers since 2006.

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

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

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

plWordNet can be browsed online (Figure 9), via a mobile app available on Google Play, or via the WNLoomViewer application. It can be used for linguistic analyses both in Polish as well as in comparative and translation studies. Due to its open licence, which is based on the Princeton WordNet, it can also be used for data mining both in research and commercial projects.

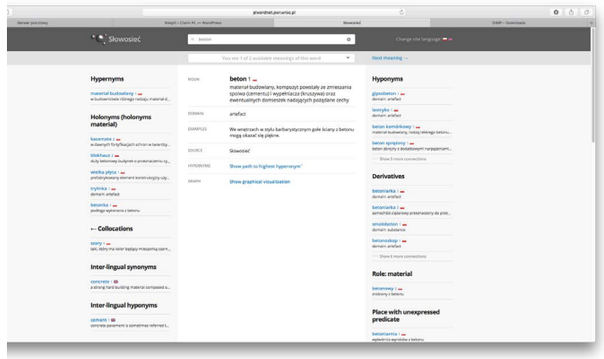


Figure 9: The interface of plWordNet.

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

The WordnetLoom Editor is a Java application that provides a visual, graph-based interactive presentation of the structures of plWordNet and thereby enables browsing and direct editing of lexico-semantic relations and synsets (Figure 10). It is remarkable for its flexibility and adjustability to the needs of individual users. It is currently being used by the Portuguese Wordnet team, and in a project led by Professor Ewa Geller from Warsaw University which aims to describe the Yiddish language and to map senses on the morphological and semantic level from Yiddish to corresponding senses in plWordNet, GermaNet, and the Princeton WordNet.

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

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

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

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

Figure 10: The interface of the WordnetLoom Editor, visualising the wordnet structure of the lemma “złodziej” (“thief”).

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

Written by Darja Fišer and Jakob Lenardič

Since April 2015, CLARIN-PL has organised a series of workshops and lectures titled “CLARIN-PL in Research Practice”. Eight editions of the series have taken place thus far, all of which were well attended, with around 40 participants per event. The latest event in this series took place in Wrocław between 19 and 20 June 2018.[4]

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

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

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

Maciej Maryl

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

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

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

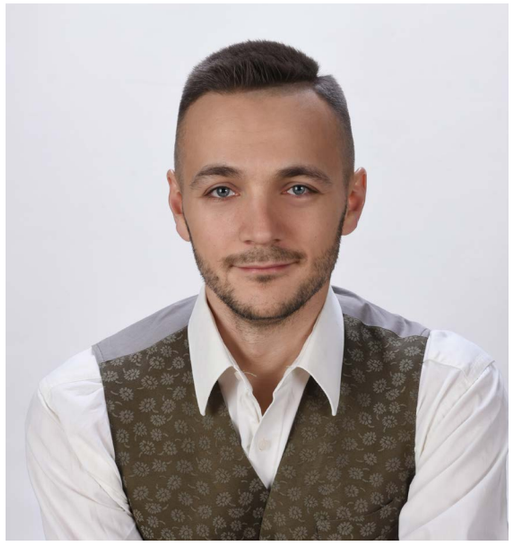
2. How did you get involved with the Polish CLARIN consortium? Are you currently collaborating with them?

In 2013, when I was in the process of setting up the Digital Humanities Centre at the Institute of Literary Research at the Polish Academy of Science, I was introduced to Maciej Piasecki, who is the coordinator for CLARIN-PL. At the same time, the Institute was organising the first THATCamp (The Humanities and Technology Camp) in Warsaw, so I invited him to present the tools developed by the Polish consortium. I was inspired by his talk and wanted to use the tools in my analyses of weblogs that I was conducting at the time. This in turn led to a very fruitful collaboration between CLARIN-PL and our institute, which goes on to this day. We have successfully cooperated on quite a number of projects. To name a few, one ongoing project involves the creation of the Literary Map,[5] in which geographical information that appears in Polish literary texts is mapped onto Google Maps. Another project is LEM (Literary Exploration Machine),[6] an online system that brings together various tools dedicated to processing and analysing literary texts. We have also started two lexicographical projects. One aims at creating a dictionary of Polish Romantic poets, using CLARIN-PL tools and WorldNet, while the other, in cooperation with many institutions, is dedicated to linking together various historical dictionaries of Polish on a single platform. In most cases, we help develop the tools that CLARIN-PL had already created, providing the expertise and needs of our field. This helps to establish a productive feedback loop between developers and users.

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

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

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



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

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

"Blogs as a new form of multimedia writing" is actually the project that marks the beginning of my collaboration with CLARIN-PL. Together with the Polish consortium we worked on the tools used to classify weblogs on the basis of their genre. To give some background, what we did at first at the Institute – that is, before involving CLARIN-PL – was to draft a typology of weblog genres based on a systematic, qualitative analysis of actual texts. We then started the cooperation with Maciej Piasecki in order to corroborate our findings with computational methods. We applied various clustering methods, using tools like CLUTO and CLARIN-PL's stylometric system WebSty (read more about this tool on page 6) to see whether they would group the weblogs together in accordance with our proposed typology. Together with Maciej Piasecki and Ksenia Młynarczyk, we have written an article dedicated to combining close reading with distant reading on the basis of CLARIN-PL tools.

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

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

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

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

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

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

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

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

There are still quite a few scholars who think that using computational approaches shifts your attention from the actual texts to the linguistic surface. I actually believe that this kind of scepticism in Polish academia is quite a widespread phenomenon due to the idea that digital approaches are reductionist and ill-suited for addressing the "big", critical questions of literary studies. But we shouldn't forget that similar reservations have been formulated against empirical approaches in the humanities, probably since the birth of anti-positivism. So, we should have probably got used to it by now. However, in the last five years, digital humanities have begun to flourish in Poland, entering the phase of institutionalisation. Related research centres were established, and researchers established the CLARIN-PL and DARIAH-PL consortia. CLARIN-PL is especially eager to bridge the gap between computational experts and humanities users,

organising hands-on workshops for researchers and translators. So, I expect the body of digital humanities research to grow, but let's not fool ourselves – it is not going to be mainstream. What we need is more digital humanities courses at universities, so the base of practitioners could steadily grow. Obviously, there are many courses in linguistic departments, but we should also reach out to students of history, literature, and cultural studies. My institute has just received a grant to start a graduate program on digital literary studies to enable the study of literature with the help of digital methods and technologies at the PhD level. This program will be carried out in cooperation with the Polish-Japanese Academy of Information Technology, which is also a member of CLARIN-PL.

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

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

Warsaw, Poland | photo by Jacqueline Macou | Pixabay

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
Jan Wieczorek, Ewa Rudnicka, Agnieszka Dziob, 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