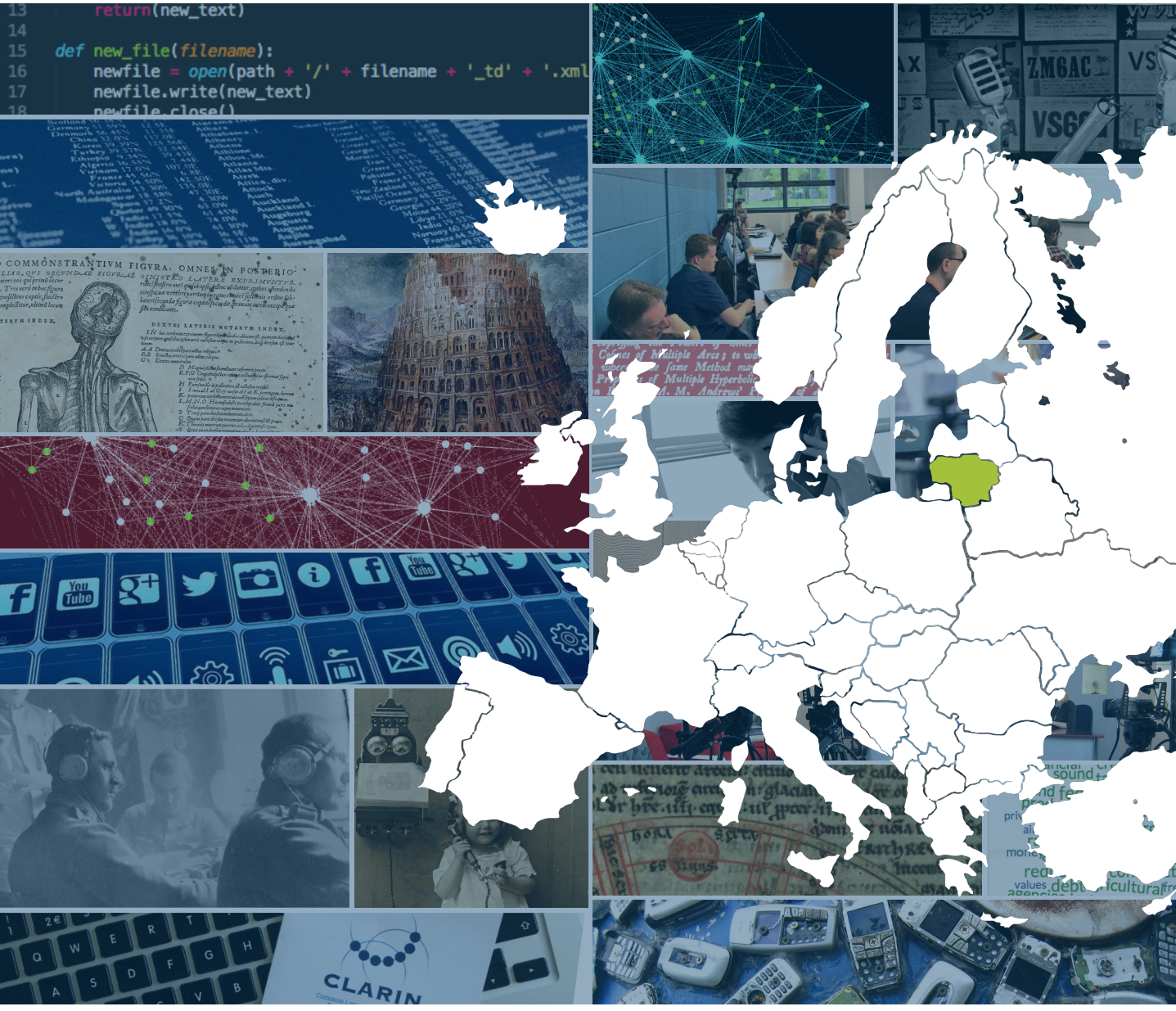
CLARIN

Common Language Resources and Technology Infrastructure

Tour de CLARIN

Lithuania



Written by Tomas Krilavičius, Jolanta Kovalevskaitė, Agnė Bielinskienė, Jurgita Vaičenonienė, 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 Lithuania 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

Vilnius, Lithuania | photo by Peggy und Marco Lachmann-Anke | Pixabay

Lithuania

Written by Darja Fišer and Jakob Lenardič

The Lithuanian consortium CLARIN-LT[1] has been a full member of CLARIN ERIC since 2015. The consortium consists of three partner universities – Vytautas Magnus University, Kaunas Technology University and Vilnius University – and is led by Assoc. Prof. Andrius Utka as the head of the consortium, Dr. Jurgita Vaičenonienė as the National Coordinator.

CLARIN-LT offers a C-certified repository, which provides a host of specialised and well-annotated language resources suitable for research within digital humanities disciplines. The consortium offers dedicated online access to the Corpus of the Contemporary Lithuanian Language. Additionally, a researcher is given access to some important Lithuanian language resources, such as ALKSNIS – the largest Lithuanian Treebank (presented on page 7), LITIS – a corpus of user-generated comments, and the Lithuanian Parliament Corpus for Authorship Attribution, which is especially tailored to authorship attribution tasks and has been successfully used in a variety of interdisciplinary research endeavours.

[1] http://clarin-lt.lt/?lang=en

To promote active user involvement the consortium has set up two help desks, whose experts can be contacted via e-mail or telephone:

- the Lithuanian Language Technology Helpdesk at Vytautas Magnus University provides information and consultations on corpus analysis, terminology extraction, the use of tools for part-of-speech tagging, syntactic parsing, and similar uses of language technology; and

- the Semantic and Conceptual Modelling Helpdesk at Kaunas University of Technology provides information and help on approaches related to database and information system engineering, ontology development methods, the implementation of semantic search processes and other relevant issues.

The consortium has also organised a series of user involvement events. In 2016, CLARIN-LT experts organised a seminar where they presented successful use cases on how the language resources and computational tools developed at the consortium can be applied within digital humanities and social sciences research. In 2017, the consortium organised their biggest event yet – a two-day CLARIN-PLUS workshop dedicated to the creation and use of social media resources, which was attended by some of the foremost computational and digital humanities experts on computer-mediated communication.

CLARIN-LT team (L-R): Andrius Utka, Agnė Bielinskienė, Jurgita Vaičenonienė, Erika Rimkutė, Rūta Petrauskaitė, Jolanta Kovalevskaitė, Loic Boizou.

Colloc

Written by Tomas Krilavičius, Jolanta Kovalevskaitė, Jakob Lenardič and Darja Fišer

Colloc[2] is an experimental tool aimed at the automatic identification of Multiword Expressions (MWEs). MWEs (or multiword units) are fixed word combinations that can be different in their nature: some of them are semantically non-compositional, i.e. their global meaning is different from the sum of their individual parts (idioms or phraseologisms), whereas others are transparent, but have usage-based co-occurrence restrictions (collocations). The tool, developed by a team of researchers working at the CLARIN-LT centre at Vytautas Magnus University and the Baltic Institute of Advanced Technology, covers the whole process of MWE identification, and can also be used for the development of new methods of MWE identification.

The experimental prototype includes all the steps of linguistic analysis, namely:

1. Text pre-processing
2. PoS tagging
3. N-gram generation and calculation of their statistical properties
4. Calculation of Lexical Association Measures (LAMs)
5. Word embedding generation
6. MWE identification using:
   - Filtering (gazetteers, dictionaries)
   - Application of LAMs
   - Application of machine Learning
   - Hybrid methods

The basic user version of Colloc, which is cloud-based and will be available soon, currently supports only Lithuanian and was trained on a 70 million-word corpus, collected from the Lithuanian news portal delfi.lt. The tool has been statistically trained on GloVe Word Vectors and employs artificial neural networks. It is designed to be user friendly, so researchers will only have to upload the text file whose multi-word expressions they want to have analysed (as in Figure 1), and the tool will simply return the annotated document.

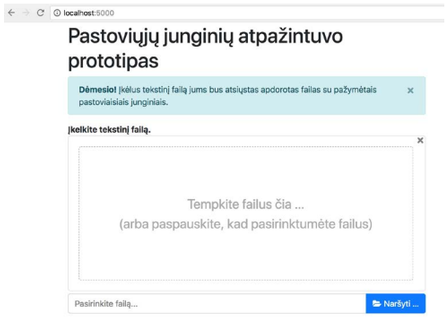


Figure 1: The Colloc user interface.

It is important to have a tool that can extract MWE candidates from particular text(s), since this opens more possibilities not only for terminological, lexicographic and NLP perspectives on language analysis, but also for different areas in applied linguistics, like language learning. The tool will help linguists perform deeper analyses of language, investigate its compositionality, idiomaticity and dynamics. Language technology specialists will be able to use Colloc to improve automatic text analysis, machine translation, information extraction tools, and make chatbots more human.

The tool development is funded by Lithuanian Research Council, Pastovu project.

[2] http://mwe.lt/en_US/

ALKSNIS, the Lithuanian Dependency Treebank

Written by Agnė Bielinskienė, Jakob Lenardič and Darja Fišer

ALKSNIS is a syntactically annotated corpus of Lithuanian, and serves as a gold standard for the syntactic analysis of the language. ALKSNIS currently consists of 2,355 syntactically annotated sentences in the PML (Prague Mark-up Language) format. The format allows researchers to visualise and edit the syntactic trees with the editor TrED.

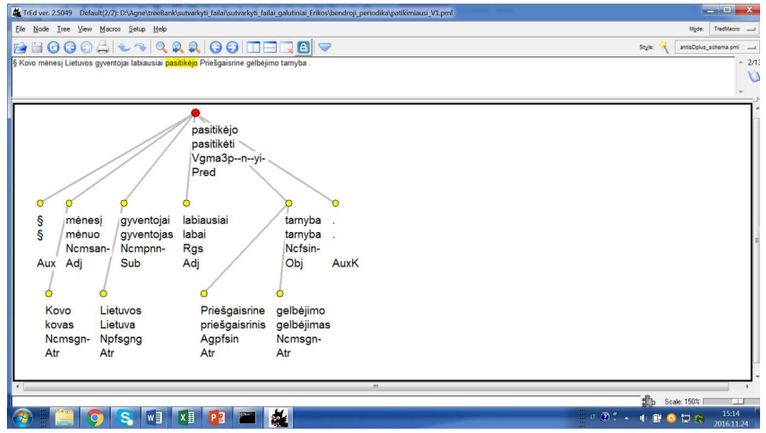


Figure 2: Using TrED to show the syntactic structure of a sentence from the ALKSNIS corpus.

Figure 2 shows the syntactic tree structure of the Lithuanian sentence “Kovo mėnesį Lietuvos gyventojai labiausiai pasitikėjo Priešgaisrine gelbėjimo tarnyba” (“In March, Lithuanian residents mostly trusted the Fire and Rescue Service”), as presented by TrEd. Each terminal node corresponds to a word, a punctuation mark or other text element (symbol, digit, etc.) within a sentence, while the links show the syntactic dependencies. The prepared list of abbreviations for syntactic labels and the presentation of the syntactic relations and dependences were based on the experience of Czech researchers (Hajič et al. 1999). The editor presents the following information for each node:

1. the form used in the sentence (e.g., “gyventojai”, “residents“ in the given example);
2. the corresponding lemma (e.g., “gyventojas”, “resident”, which is the singular form of the plural “gyventojai”),
3. the morphology tag (e.g., “gyventojai” “residents” has the tag Ncmpnn-, which stands for Noun, common, masculine, plural, nominative, non-reflexive, - indistinctive), and
4. the syntactic function (e.g., “gyventojai”, “residents” is the grammatical subject in the given example).

The corpus can also be searched via the ANNIS interface (Krause and Zeldes, 2016). The interface visualises the syntactic dependencies of a sentence and lists its morphosyntactic features, as shown in Figure 3: “Patalpos jau išnuomotos. Taip pat jau rezervuota pusė ploto kitais metais iškilsiančiame statinyje. Dauguma didmenine” (“The premises have already been leased. Also, half of the area of the building to be finished next year has already been reserved. Mostly wholesale”).

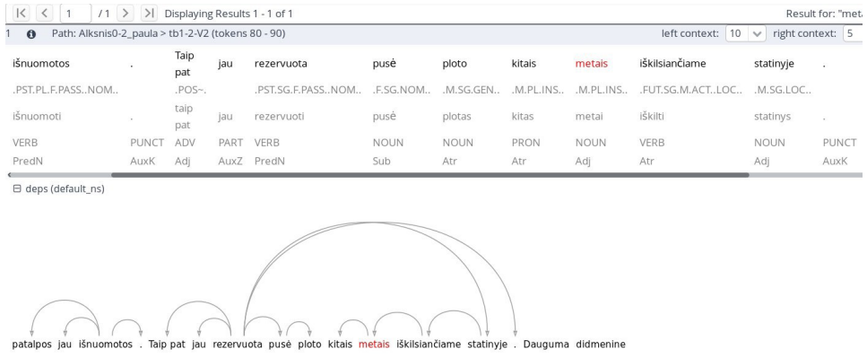


Figure 3: Using ANNIS to parse a sentence in ALKSNIS.

So far, the syntactically annotated corpus has been successfully used by different user groups. For example, at Vytautas Magnus University, students are taught to work with ALKSNIS as part of the curriculum and use corpus data to do various assignments or to develop their theses (for instance, Kristina Brokaitė's Master's thesis used the corpus to analyse grammatical forms of various complex and non-complex predicates in Lithuanian).

The corpus will be enriched with new texts and converted to the Universal Dependency (UD) format. The CoNLL-U format provided by the UD guidelines will serve as the core version of the ALKSNIS treebank. We also plan to annotate the corpus for multiword expressions (also see the description of Colloc, which is a tool for annotating MWEs, on page 6). This will help enhance the usability of the corpus in parsing and in data-driven applications of MWE processing models as well as provide linguists with the information about the syntactic behaviour of Lithuanian MWEs. Finally, a syntactic parser is going to be trained on the Alksnis corpus.

References:

Bielinskienė, A., Boizou, L., Kovalevskaitė, J., and Rimkutė, E. (2016). Lithuanian Dependency Treebank ALKSNIS. In Proceedings of the Seventh International Conference Baltic HLT 2016. Amsterdam: IOS Press, 107–114. http://ebooks.iospress.nl/volumearticle/45523.

Brokaitė, K. (2017). Tarinio raiška gramatinėmis formomis sintaksiškai anotuotame lietuvių kalbos tekstyne ALKSNIS. https://eltalpykla.vdu.lt/1/34649.

Hajič J., Panevová J., Buráňová E., Urešová Z., and Bémová A. Annotations at Analytical Level. (1999). Instructions for Annotators (11.10.1999), UK MFF ÚFAL Praha.

Krause, Th. and Zeldes, A. (2016.) ANNIS3: A New Architecture for Generic Corpus Query and Visualization. In Digital Scholarship in the Humanities 2016 (31). http://dsh.oxfordjournals.org/content/31/1/118.

The Annual CLARIN-LT Seminars

Written by Jurgita Vaičenonienė and Darja Fišer

Since the establishment of the CLARIN-LT centre in 2015, we have been organising different types of events to disseminate knowledge about language resources and language analysis tools deposited in the repository of the national consortium. We help lecturers, teachers and students of the humanities and social sciences to use language resources efficiently in their work and research, contribute their data to our repository, or get involved in various CLARIN related activities. Taking into consideration the needs of different audiences, we offer both recurring and single events with the focus on Lithuanian language resources.

An especially successful initiative which addresses researcher and lecturer communities is our annual seminar series[3] organised at the end of each year. The aims of the seminars are to introduce and give an update about the activities of CLARIN-LT and CLARIN ERIC in general; to present the language resources stored or soon to be added to our repository; and show their possible applications. Most importantly, we also use the opportunity to find out the expectations of the audience related to language resource creation and use.

Every year, we attract about 20 researchers from different Lithuanian universities from different disciplines, such as sociology, psychology, economics, politics, linguistics, journalism, cultural studies, informatics, and history.

Part of the seminar is devoted to the presentations of participants where they introduce their research in digital humanities.

Finally, the most important part of the event is a round table discussion when all participants are given the floor to share their thoughts which helps us to find out the needs of current and prospective CLARIN service users and, at the same time, offer our help or give explanations in response to some of the questions posed.

The outcomes of these discussions are especially helpful when scheduling our yearly activities and help to share our experiences with everyone interested in digital humanities within and outside academia. Each year we receive more and more inquiries to present the mission and goals of CLARIN ERIC and help with various research related questions, which we are always willing to do.

[3] http://clarin-lt.lt/?page_id=458

INTERVIEW: **Erika Rimkutė**

Erika Rimkutė is Senior Researcher at the Centre of Computational Linguistics at Vytautas Magnus University. The following interview took place via e-mail and was conducted by Jurgita Vaičenonienė, edited by Darja Fišer and Jakob Lenardič.

1. Could you briefly tell us about your academic background? What motivated you to apply a computational approach to linguistics?

I studied linguistics at Vytautas Magnus University (VMU). I was inspired to take up computational and corpus linguistics by Prof. Rūta Petrauskaitė, who is the founder of corpus linguistics in Lithuania and was my thesis advisor. The topic of my PhD was morphological ambiguity, which I analysed using a morphologically annotated corpus of Lithuanian. I defended my PhD in 2006 and am now a researcher at the Centre of Computational Linguistics and a lecturer at the Department of Lithuanian Studies at VMU.

I've been a member of the Centre of Computational Linguistics at VMU since my MA studies, which has given me a lot of valuable opportunities to get involved in computational research. I was able to get acquainted with specialists in corpus and computational linguistics working in Lithuania and other countries, to try out different language analysis software and corpora, as well as to observe other developments in language research. The experience gained in this way allowed me to specialise in automatic morphological analysis.

2. You've worked on quite a few important language projects with the Lithuanian CLARIN consortium. What have your contributions been when collaborating with the consortium?

I've had the opportunity to contribute to the creation of some of the key resources in the CLARIN-LT infrastructure. For example, the first version of MATAS[4] was compiled during my PhD studies to analyse the problem of morphological ambiguity, which had previously been very limitedly investigated in Lithuania and abroad. Manual annotation of semi-automatically annotated texts helped me to describe this phenomenon in detail in my dissertation, which contributed to the development of more accurate automatic morphological annotation tools for Lithuanian. The revised version of MATAS was added to the CLARIN-LT repository, so that it is now available for anyone interested in it.

[4] http://hdl.handle.net/20.500.11821/9

3. Would you like to recommend a language resource or tool developed at the consortium that you think is important for the study and analysis of the Lithuanian language?

Since 2016, I've been leading the project Automatic Identification of Lithuanian Multi-word Expressions financed by the Research Council of Lithuania. The project aims to develop a methodology for analysing Lithuanian MWEs by creating or adapting necessary tools and resources. Apart from the MWE identification methodology, we also aim to create MWE extraction tools, a database of Lithuanian MWEs with multifunctional search options, and a corpus-based dictionary of Lithuanian collocations.

CLARIN-LT is a partner on the project, which to me is an example of a successful collaboration between CLARIN-LT and linguists. CLARIN-LT provides me with the technical support and creates the tools necessary for the implementation of the project. In return for their support, we will upload all the results into the CLARIN-LT repository and make them easily accessible for other researchers. For example, at the end of the year, the first dictionary of Lithuanian collocations will be released. Users will be able to access a database of Lithuanian multiword units encompassing over 10,000 lemmas. I think that this is an important contribution both for the development of further lexicographic resources as well as language teaching, especially given that collocation dictionaries don't yet exist for most under-resourced languages like Lithuanian.

4. You are also a teacher at the Department of Lithuanian Studies at the Vytautas Magnus University. How do you integrate the computational approach into your course-work? Do you introduce the CLARIN infrastructure to your students?

I cannot imagine my classes without introducing students to the morphologically and syntactically annotated corpora. Naturally, before starting work with the resources I introduce the main principles of CLARIN and the role of national repositories. I always encourage my students to use the Corpus of Contemporary Lithuanian Language, which was developed by CLARIN-LT, for example. Although corpora do not provide ready-made information, in contrast to dictionaries, I believe it is vital to teach students the importance of making linguistic claims on the basis of authentic language use. The students of BA and MA study programmes of Lithuanian Philology and Modern Linguistics, where I teach, are taught to work with the Lithuanian Morphologically Annotated Corpus MATAS during the lectures on morphology and word formation. The students have to identify the missing node in the collocations extracted from the Corpus of Contemporary Lithuanian Language; identify parts of speech and grammatical categories in extracts from MATAS; analyse syntactic relations in ALKSNIS, etc. Apart from in-class activities, students also write seminar papers and BA and MA theses drawing on data extracted from the mentioned resources, some of them are even invited to work on our research projects. For example, Rūta Brinkutė's MA thesis analyses the distribution of grammatical categories in different genres. I believe that knowing how to work with annotated corpora and tools might be valuable for students in their future work as language editors or researchers.

5. You have been part of the team that created the LILA corpus,[5] which is a parallel corpus of Lithuanian and Latvian. The team included both Lithuanian and Latvian researchers involved with CLARIN. How does the Lithuanian CLARIN consortium benefit from such cross-border collaborations? Do you plan to upload the corpus into the consortium's repository?

The project was part of the EU Cross-Border Cooperation Programme and was conducted in 2011-2012 before either Lithuania or Latvia were CLARIN members. The nine million-word Lithuanian-Latvian-Lithuanian parallel corpus aligned on paragraph and sentence level was compiled by researchers of the Vytautas Magnus University's Centre of Computational Linguistics and the Latvian University's Mathematical and Informatics Institute's Laboratory of Artificial Intelligence (LU MII). It was a really interesting and mutually beneficial experience to work with my Latvian colleagues, as our teamwork not only resulted in the creation of the corpus itself, but also in several joint publications. I believe that if the project was implemented now, when both countries have CLARIN centres, the project aims and results could have been formulated on a much larger scale and more language pairs could have been included in the corpus. I see great value in such collaborative projects, as they allow us to combine a wide variety of research perspectives and approaches, which in turn enhances professional and personal cooperation between the research centres and scientists in different countries.

6. What would you recommend CLARIN to do in order to attract more researchers from the Lithuanian linguistics community?

In relation to my previous comment on the LILA corpus, I think that CLARIN could focus more on promoting joint scientific projects among the CLARIN centres of different countries to create comparable language resources and compatible processing tools. I also think that the fact that there is a consortium like CLARIN-LT which develops tools and resources specifically dedicated to Lithuanian can be very inspiring for new initiatives and research projects that also might want to start working with other less-resourced languages.

Also, I would like to see interoperable lexicographical databases to become available through CLARIN. At the very least, providing more information on the availability of such resources would be very helpful. For example, during the lexicographic project "Automatic Identification of Lithuanian Multi-Word Expressions", we were looking for a database we could reuse for our research. As we found none, we spent a lot of time as well as human and financial resources to create the database ourselves.

[5] http://tekstynas.vdu.lt/page.xhtml?id=parallelLILA

Klaipeda, Lithuania | photo by Åsa K | 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
Tomas Krilavičius, Jolanta Kovalevskaitė, Agnė Bielinskienė, Jurgita Vaičenonienė,
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