CLARIN

Common Language Resources and Technology Infrastructure

Tour de CLARIN
BULGARIA

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

BULGARIA

Foreword

Tour de CLARIN highlights prominent user involvement activities of CLARIN National Consortia and Knowledge Centres with the aim to increase their visibility, 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 and knowledge centres 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.

The brochure presents Bulgaria 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 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

Bulgaria | Introduction ____ 4
Tool | **BTB-Pipe: a Language Pipeline for Bulgarian** ____ 6
Resource | **Bulgarian Child Language Corpus** ____ 9
Event | **CLaDA-BG Dissemination Activities** ____ 11
Interview | **Aneta Nedyalkova** ____ 13

BULGARIA

Introduction

Written by **Petya Osenova** and **Kiril Simov**

Bulgaria has been a founding member of CLARIN ERIC since 2012. In 2014, following the strategic plan of the Bulgarian Government and Ministry of Education and Science, the CLARIN and DARIAH Infrastructures merged into a single infrastructure called CLaDA-BG (CLARIN and DARIAH in Bulgaria) and obtained funding in 2018.[1]

In Europe such models have already proved to be successful in the Netherlands, Austria and Greece. The CLaDA-BG consortium is very heterogeneous; its members come from universities, other academic institutions, museums, libraries, non-government organizations and companies. It includes a group of language and semantic technology oriented partners, on the one hand, and expert and content oriented ones, on the other.

The first group includes: the Institute of Information and Communication Technologies at the Bulgarian Academy of Sciences (Coordinator for CLaDA-BG and CLARIN-BG), Institute of Mathematics and Informatics at the Bulgarian Academy of Sciences,

[1] http://clada-bg.eu/

Ontotext AD (Sirma AI), Sofia University “St. Kliment Ohridski” (Coordinator for DARIAH-BG), New Bulgarian University, Konstantin Preslavsky University – Shumen, and Bulgariana – an NGO promoting CH in Bulgaria. The second group includes: the South-West University “Neofit Rilski” – Blagoevgrad, Sirma Media, the Cyrillo-Methodian Research Centre at the Bulgarian Academy of Sciences, Institute of Balkan Studies and Centre of Thracology “Alexander Fol” at the Bulgarian Academy of Sciences, Institute of Ethnology and Folklore Studies with Ethnographic Museum at the Bulgarian Academy of Sciences, Burgas Free University, “Ivan Vazov” Public Library – Plovdiv, and Sofia History Museum.

The mission of the infrastructure is to build a scientific ecosystem for supporting research in Social Studies and the Digital Humanities. The main goal is to construct a Bulgaria-Centric Knowledge Graph (BGKG) – repository where all types of linguistic and encyclopaedic knowledge are stored and linked. Thus, they will be used for extracting content with respect to particular tasks.

In their first year of operation the consortium worked on: structuring of the various resources, extending and building contemporary and old corpora, and modelling cultural objects, contextualizing the knowledge through connecting events, artefacts, and descriptions.

Some of the main resources to mention are: the syntactic corpus BulTreeBank (215,000 tokens), the BTB-Wordnet that is integrated with Wikipedia (22,000 synsets), the Valency Lexicon (6,000 verb frames), the Inflextional Lexicon (over one million wordforms) (the Institute of Information and Communication Technologies), the large Bulgarian corpus with statistics on collocations with a span of one to six tokens (eight million webpages have been processed) (Ontotext AD), the Corpus of Child Speech (33 hours of records and 355 pages of transcripts) (Shoumen University), the Ethnographical Museum exhibition on 3D representation, the epigraphic collection of ancient inscriptions in Greek – TELAMON[2] (Sofia University), bilingual corpora (New Bulgarian University), and so on.

Among the most important tools for Bulgarian are: the NLP pipeline and the online concordance webclark (IICT-BAS). Several other tools are also in development: an old-to-new spelling transformation tool, a conceptual and keyword search tool over a huge corpus of contemporary Bulgarian, and a semantic annotator of Bulgarian. CLaDA-BG’s plans include the creation of CLARIN B and K centre, and applying for assessment during the second year of the project.

[2] https://telamon.uni-sofia.bg

Tool | **BTB-Pipe: a Language Pipeline for Bulgarian**

Written by **Petya Osenova** and **Kiril Simov**

The BTB-Pipe language pipeline for Bulgarian has been developed incrementally over the last twenty years, starting with the Bulgarian-German BulTreeBank project for the creation of a Bulgarian treebank. The BTB-Pipe comprises the following modules:

- Tokenizer and sentence splitter
- Morphosyntactic tagger
- Lemmatizer
- Dependency parser

Bulgarian is an analytical language with rich word inflection, predominantly in the verbal area. The rich morphology inevitably leads to a lot of morphological ambiguity. Consequently, morphosyntactic tagging is more complex in Bulgarian than in languages like English. BTB-Pipe is a hybrid system combining a rule-based module and a statistical module (Simov and Osenova 2001) and uses the BulTreeBank Morphosyntactic Tagset (Simov, Osenova, and Slavcheva, 2004).

The lemmatizer in BTB-pipe comprises a set of transformation rules that have been developed based on the 1998 inflectional lexicon (Popov, Simov, and Vidinska 1998). Since the rules in the lexicon are implemented through the CLaRK system, they can also be used on unknown words in order to produce some guesses with regard to their word lemmas.

This is an illustrative example of a lemmatization rule:

if pos-tag = Vpitf-o1s then
{ remove -ox; concatenate -a }

When the lemmatizer applies this rule to the verb form *четох* (roughly /četoh/), where the inflection *–ox* encodes the features 1st person singular and the past indefinite tense (“I read”), it produces the lemma *чета* (/četa/).

The parser uses MaltParser and Mate Dependency Parser for training dependency trees. The input is the result from the tagger and the lemmatizer, and the output a dependency tree or trees for the sentences in the text, using either an internal set of dependency relations developed for the CoNLL 2006 Shared Task or the Universal Dependencies.

The current version of the BTB-pipe can be used in three different modes: as a standalone application, as a command line, and as a web service. The output of the pipe can be in the WebLicht standard developed within CLARIN-D (Hinrichs et al. 2010) or in the NAF format (Fokkens et al. 2014). Currently, ClaDA-BG is redesigning and reimplementing some of the modules using spaCy with the goal of improving the performance of the pipeline.

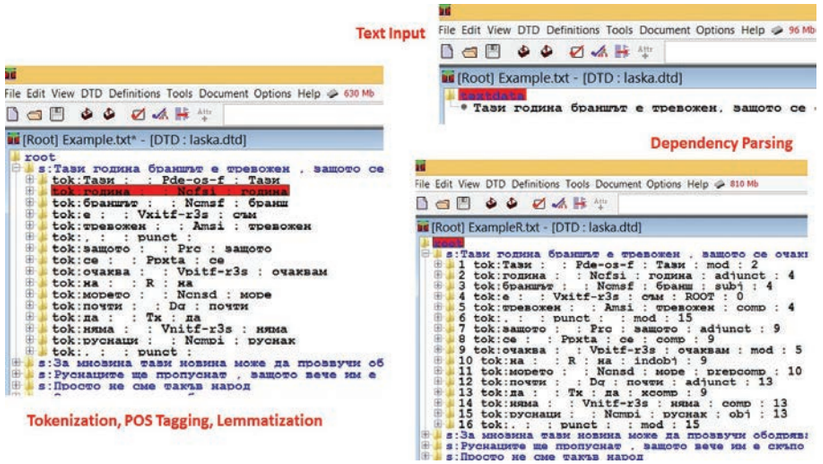


Figure 1: *Linguistic annotation in BTB-Pipe*

Figure 2: *BTB-Pipe annotation in the CLaRK XML Editor*

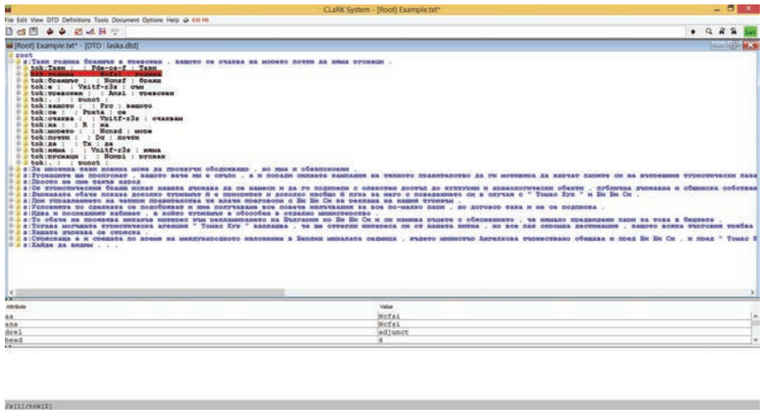


Figure 3: *Tokenization, lemmatization and morphosyntactic tagging with BTB-Pipe*

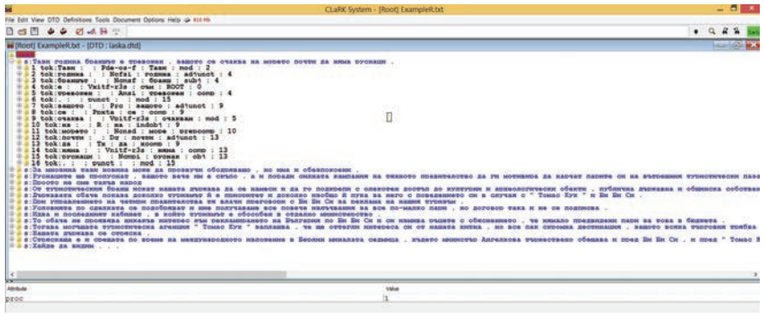


Figure 4: *Syntactic relations annotated with BTB-Pipe*

References:

Fokkens, A., Soroa, A., Beloki, Z., Rigau, G., van Hage, W.R., and Vossen, P. NAF: the *NLP Annotation Format*. Technical Report NWR-2014-3. Version 1.1. NewsReader project: Building structured event indexes of large volumes of financial and economic data for decision making – ICT 316404.

Hinrichs, E., Hinrichs, M., and Zastrow, T. 2010. *WebLicht: Web-Based LRT Services for German. In Proceedings of the ACL 2010 System Demonstrations*, 25–29, Uppsala, Sweden.

Popov, D., Simov, K., and Vidinska, S. 1998. *A Dictionary of Writing, Pronunciation and Punctuation of Bulgarian Language*. Atlantis LK, Sofia, Bulgaria.

Simov, K., Osenova, P., and Slavcheva, M. 2004. BTB:TR03: *BulTreeBank morphosyntactic tagset BTB-TS version 2.0*. Technical Report.

Simov, K. and Osenova, P. 2001. *A Hybrid System for MorphoSyntactic Disambiguation in Bulgarian. In the Proceedings of the RANLP 2001 Conference*, Tzigov Chark, Bulgaria, 5–7 September 2001, 288–290.

Resource | **Bulgarian Child Language Corpus**

Written by **Petya Osenova** and **Kiril Simov**

The first systematic study of child speech in Bulgaria is attributed to the early 20th century philosopher Prof. Ivan Georgov, but interest in child language truly took off in the last decade of the 20th century and the beginning of the 21st, with Yuliana Stoyanova and Velka Popova who worked on longitudinal data. The contemporary systematic study of child speech, based on solid empirical material, is also associated with the creation of the first Bulgarian corpus in the CHILDES framework by the research team of the Laboratory of Applied Linguistics (LABLING) at Shumen University. The corpus is based on an array of longitudinal data from Popova's personal archive.

The CHILDES framework is reputed for its openness and rationality, which are leading factors in the processes of cooperation and globalization in the Humanities. This is a guarantee of both the broad social validity of the research results based on corpora, and their integration into initiatives for exchanging linguistic data and technologies aimed at overcoming the current fragmentation of the research field. Moreover, CHILDES and the sister initiative TalkBank are already integrated into CLARIN as one of the Knowledge Centres. The Bulgarian child language corpus enables cross-lingual research and contributes to a modern, convenient standard for the study of linguistic ontogeny, which, thanks to its universal parameters, enables rapid, accurate and reliable comparison with a large number of languages and the development of solid typologies and modern theories.

The corpus comprises two types of speech resources: CORPUS A (spontaneous speech material of four children at their early age – from one to three years old) and CORPUS B (comprising stories based on a series of pictures with 90 children at pre-school age (from three to six years old). For the sake of integrity and processing, the speech resources are presented in two formats – in Cyrillic as well as Latin. Figure 5 illustrates the encoding of two children.

@Begin	@Begin
@Participants: ALE Alexandra Target_ Child, VEL Velka Mother	@Participants: ALE Alexandra Target_ Child, VEL Velka Mother
@Birth of ALE: 29-JAN-1989	@Birth of ALE: 29-JAN-1989
@Date: 4-JUL-1990	@Date: 4-JUL-1990
@Filename: al10506	@Filename: al10506
@Age of ALE: 1;05.06	@Age of ALE: 1;05.06
@Situation: at home	@Situation: at home
*VEL: Njama li da spish ti?	*VEL: Няма ли да спиш ти?
*ALE: Dzak.	*ALE: Дзак.
*VEL: Dzak li?	*VEL: Дзак ли?
*VEL: Ja da nankash!	*VEL: Я да нанкаш!
*VEL: Kakvo si na mama ti?	*VEL: Какво си на мама ти?
*VEL: Mamino kokiche.	*VEL: Мамино кокиче.
*VEL: Kakvo da ti donese mama – mlechice ili kompot?	*VEL: Какво да ти донесе мама – млечице или компот?
*ALE: Popot [:kompot].	*ALE: Попот [:компот].
……	….....
@End	@End

Figure 5: *Excerpt from the Bulgarian CHILDES corpus*

Future development of the corpus includes annotation with part-of-speech and morphological information, and integration with the WebCLaRK online service, a Bulgarian portal for language services on the web.[3] Video data also exists for the same material, the processing of which is in progress and will be included in one ClassTalk session in the TalkBank database. The data comprises recorded classroom interactions in a number of kindergarten groups. Video transcription of the video data will follow the same basic principles as audio transcription. These Bulgarian corpora could be used not only for research of classroom interactions between the teacher and children, but also as sample material for training students of pedagogy.

[3] http://webclark.org/

Event | **CLaDA-BG Dissemination Activities**

Written by **Petya Osenova** and **Kiril Simov**

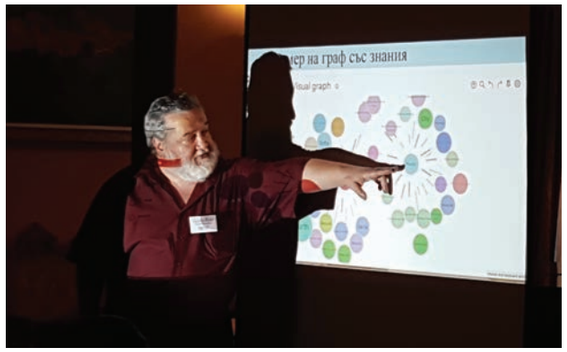
With partners belonging to both CLARIN and DARIAH, the CLaDA-BG consortium is very heterogeneous. For that reason, it regularly organizes seminars and dissemination activities that are aimed at presenting the infrastructure to researchers with backgrounds in the Humanities, such as history, ethnography, library science, and museology. These are 'hosted events', which means that CLaDA-BG experts visit consortium institutions and their teams. During CLaDA-BG's first year of operation we mainly disseminated the goals of the infrastructure to the interested audience and partners. At these meetings we received valuable feedback and also learnt a lot about the needs of the potential users, which are summarized below.

The first awareness raising event was called "Open Science Infrastructures for Big Cultural Data: International Advanced Masterclass". It was organized by UCL Qatar in collaboration with DARIAH-EU and the National Library Ivan Vazov - Plovdiv. The event was held in Plovdiv, Bulgaria, from 13 to 15 December 2018. Members of CLaDA-BG presented the available resources and tools. Dimitar Iliev from Sofia University presented Telamon, which is a corpus of Greek inscriptions found in Bulgaria, while Dimitar Minev, who is the Director of the Plovdiv National Library, talked about how museology datasets can be turned into Linked Open Data resources. A feedback session followed where interesting comments were provided by colleagues from the British Library, especially on the integration of the processing language and image information. The event was concluded by a panel discussion on *the Next Steps in Bulgarian Open Humanities*. It was chaired by CLaDA-BG coordinator Kiril Simov. The panellists were Roumiana Preshlenova (the Institute of Balkan Studies and Centre of Thracology, BAS, CLaDA-BG) and Georgios Papaioannou (UCL Qatar).

On 15 February 2019, a one-day seminar was organized at the Institute of Balkan Studies with a Center of Thracology. At the event, Kiril Simov presented the mission, constitution and organization of the infrastructure, while Petya Osenova presented the resources and tools it offers, after which researchers from the Institute presented their work. In the second part of the seminar, the participants discussed with the lecturers how to make their data and dictionaries machine readable and searchable, and how to OCR and process old books or newspapers at the Institute. The lecturers explained the basic principles of constructing structured data and processing it with the NLP pipe for Bulgarian. In addition, a decision was made to use the data provided by the Institute for the creation of a normalization model to modernize old texts, making them processable by the existing NLP modules.

On 30 and 31 May 2019, Kiril Simov delivered a dissemination lecture at the 12th National Conference “Education and Research in the Information Society”. The lecture was titled “Integrated Language and Knowledge Resources for CLaDA-BG”. The participants were representatives of Bulgarian libraries, universities and educational institutions. The libraries were especially interested in the aspects of improving and speeding up the digitization of their data. In response to their interest, CLaDA-BG started working on the creation of an appropriate normalization model for older texts.

On 23 August 2019, CLaDA-BG experts attended an informal seminar at the Cyrillo-Methodian Research Centre, BAS which has a rich collection of medieval manuscripts in Old Bulgarian, Russian, German and other languages. The main problem of researchers who study Cyril and Methodius is the proper handling of old lexica and their compilation into searchable online dictionaries, a prerequisite for which is OCR and editing, which calls for the reuse of the existing services available in CLARIN centres in other European countries.



Kiril Simov presenting CLaDA-BG to humanities researchers

Interview | **Aneta Nedyalkova**

Aneta Nedyalkova is an MA student of Bulgarian philology. Under the auspices of CLaDA-BG, she is working on an associative dictionary of verbal expressions.

You are a really early-stage-researcher since you are just finishing your MA. How did you get interested in psycholinguistics?

<

My interest was inspired by a course on interdisciplinary approaches in linguistics at Shumen University, especially the project-oriented psycholinguistics practicum which addressed questions that have always excited me. Under the guidance of Prof. Velka Popova I conducted an experiment on word associations with ten people from the Osmar village, Northeastern Bulgaria, where I live. The course assignment then grew into a master’s thesis project and I found myself in the role of a junior but enthusiastic researcher.

>

How did you get involved with CLaDA-BG?

<

I was in an advanced phase of my associative investigation on my master’s thesis when I attended a presentation on CLaDA-BG by the local coordinator Prof. Dimitar Popov and immediately realized that my research interests would be a great fit for CLaDA-BG, as it would give me an opportunity to collaborate with experienced linguists and researchers with similar interests and so learn from them.

>

What are you working on at the moment?

<

I am finishing my master's thesis, titled "Specific features of the contemporary Bulgarian native speakers' dictionary. A psycholinguistic research", which presents a pilot survey on the mental lexicon of non-expert native speakers based on three psycholinguistic procedures: a free associative experiment, spontaneous elicitation of definitions and example sentences for a given word. My results motivated me to extend experimental work and create a dictionary of verbal associations of 100 people from the Osmar village which will be my main contribution to CLaDA-BG.

>

What makes such a dictionary important from a psycholinguistic perspective? Why did you choose this region?

<

One of the key challenges of psycholinguistics is the study of the mental lexicon. A free associative experiment is one of the most popular approaches for this, as it allows psycholinguists to uncover very broad semantic patterns that exist in human consciousness, that is to say, cognitive links between words that are not based on lexical-logical relations, such as synonymy and antonymy, but on more ephemeral associative links. In an associative experiment, you ask participants to write down certain word combinations that they associate with a target notion.

One of the main tasks of CLaDA-BG's Shumen team is the creation of several contemporary associative dictionaries which will serve as a basis for investigations of language awareness in Bulgarian society and for researching the sociolinguistic aspects of the Bulgarian mental lexicon. In the Bulgarian lexicographic tradition, there already exist two associative dictionaries. The first is the *Bulgarian Standards of Verbal Associations from 1984 and the second is The Slavic Associative Dictionary: Russian, Belarusian, Bulgarian, and Ukrainian* from 2004. While the 1984 dictionary is outdated, The Slavic Associative Dictionary, albeit more recent as well as multilingual, has a major methodological flaw in that it includes data from only one social group, students between the ages of 18 and 25. Consequently, the application of the dictionary in research is quite limited.

CLaDA-BG aims to create new associative dictionaries that will be broader in scope with regards to sociolinguistic variables, and will account for differences in territorial origin, gender, age, education, and profession. Consequently, they will be useful resources for a wide range of users across the Humanities and social scientists, such as linguists, psycholinguists, sociolinguists, ethnolinguists, cognitologists, psychologists, teachers, and political scientists. My task to create an associate dictionary of verbal expressions on the basis of the inhabitants of the town of Osmar is just one of CLaDA-BG's associated dictionaries. I have chosen this town for two reasons. First, Osmar is an urban-type settlement, in-between a typical town and a typical village. This is reflected in the inhabitants' specific lifestyle, clothing, and attitude towards technological progress, that is also reflected in the collective features of their mental lexica. The second reason is personal – I live and work as a secretary in Osmar's local library.

>

Could you describe in more detail the compilation of the dictionary of verbal associations and related preparation of questionnaires and experiments? What are the inspiring parts of this work and what are the challenges?

<

The compilation of an associative dictionary involves several stages. First, you need to design the associative experiment that will be the basis for the dictionary. This involves selecting the participants, determining the word-stimulus pairs, setting up the research design (written or spoken) and developing the research materials. Then the experiment is run with every participant separately, which is followed by processing and summarization of the results. Direct contact with the participants is the most inspiring part for me. It is a challenge for me to prepare and motivate them to participate in the experiment. The actual experiment is always interesting, and sometimes very funny or even emotional. The final part of summarizing the data in a systematic way in the dictionary is the hardest and most exhausting, but the curiosity to see the results keeps me inspired and enthusiastic even in this last stage.

>

How did the services and knowledge expertise in CLaDA-BG support you in your work?

<

In general, research is often lonely and challenging for a beginner, so being part of the CLaDA-BG team helps a lot. After the initial training of the young researchers by Prof. Popova, we were offered guidance by the local project coordinator Prof. Popov. In addition, the Student Linguistics Club was established – a small community for like-minded students in which we discuss our research in the infrastructure.

Several different associative dictionaries are currently being developed, some of which have resulted from the joint work of students under the guidance of CLaDA-BG. One of the PhD students is the coordinator and synchronizer of the collected data, which is organized and submitted in separate batches. They are then reviewed by the scientific supervisor Prof. Popova. This way, the data goes through two levels of verification, which provides control and guarantees the objectivity of the results.

The first year of work on the associative dictionaries has shown me the importance of the guidance that CLaDA-BG has offered me related to the compilation of the associative dictionary. Their other language services, such as models and standards for data processing, are also important and useful for us young researchers. The further expansion of these services and the related software environment, which are part of the consortium's future agenda, could be considered as an optimal perspective for the accomplishment of higher quality of the research work.

>

What tools and/or resources of CLaDA do you find most useful for your current and future work and why?

<

For my research, the most helpful CLaDA-BG resources are the corpora of spontaneous speech and the lexicons which provide me with material for the verification of my hypotheses and theoretical models. In addition, I also use the language processing modules, such as the part-of-speech tagger and the sense annotator, because they make my data structured and searchable. Finally, I extensively use the WebClark concordancer for detecting additional contexts that provide explanations for various associations.

>

What do you envisage focussing on after the completion of your MA?

<

After my MA I intend to continue my research in the field of psycholinguistics. I plan to participate in the work on the expansion of associative data and speech corpora by CLaDA-BG, which will be beneficial for theoretical as well as applied activities of a wide range of specialists. In conclusion, I would say that I am very happy to be a member of the CLaDA-BG team, since with my experience and data I am able to help other researchers but also have the opportunity to further develop my competencies.

>

COLOPHON

This brochure is part of the ‘Tour de CLARIN’ volume II
(publication number: CLARIN-CE-2019-1537, November 2019).

Coordinated by
Darja Fišer and **Jakob Lenardič**

Edited by
Darja Fišer and **Jakob Lenardič**

Proofread by
Paul Steed

Designed by
Tanja Radež

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

Publication number
CLARIN-CE-2019-1537
November 2019

ISBN
9789082990911

Contact
CLARIN ERIC
c/o Utrecht University
Drift 10, 3512 BS Utrecht
The Netherlands

www.clarin.eu

CLARIN 2019