

CLARIN Annual Conference 2021

PROCEEDINGS

Edited by

Monica Monachini, Maria Eskevich

27 – 29 September 2021
Virtual Edition

Please cite as:
Proceedings of CLARIN Annual Conference 2021. Eds. M. Monachini and M. Eskevich.
Virtual Edition, 2021.

'Cretan Institutional Inscriptions' Meets CLARIN-IT
Irene Vagionakis, Riccardo Del Gratta, Federico Boschetti, Paola Baroni, Angelo Mario Del Grosso,
Tiziana Mancinelli and Monica Monachini 48

Swedish Word Metrics: A Swe-Clarin resource for Psycholinguistic Research in the Swedish Language
Erik Witte, Jens Edlund, Arne Jönsson and Henrik Danielsson 54

Annotation and Acquisition Tools

Creating an Error Corpus: Annotation and Applicability
Þórunn Arnardóttir, Xindan Xu, Dagbjört Guðmundsdóttir, Lilja Björk Stefánsdóttir and Anton Karl
Ingason 59

ALEXIA: A Lexicon Acquisition Tool
Steinunn Rut Friðriksdóttir, Atli Jasonarson, Steinþór Steingrímsson and Einar Freyr Sigurðsson 64

CLARIN Knowledge Centre for Belarusian Text and Speech Processing (K-BLP)
Yuras Hetsevich, Jauheniya Zianouka, David Latyshevich, Mikita Suprunchuk, Valer Varanovich and
Katerina Lomat 68

Enhancing CLARIN-DK Resources While Building the Danish ParlaMint Corpus
Bart Jongejan, Dorte Haltrup Hansen and Costanza Navarretta 73

Annotation Management Tool: A Requirement for Corpus Construction
Yousuf Ali Mohammed, Arild Matsson and Elena Volodina 77

A Method for Building Non-English Corpora for Abstractive Text Summarization
Julius Monsen and Arne Jönsson 82

Reliability of Automatic Linguistic Annotation: Native vs Non-native Texts
Elena Volodina, David Alfter, Therese Lindström Tiedemann, Maisa Lauriala and Daniela Piipponen
..... 90

Research Data Management, Metadata and Curation

*Seamless Integration of Continuous Quality Control and Research Data Management for Indigenous
Language Resources*
Anne Ferger and Daniel Jettka 95

*The TEI-based ISO Standard "Transcription of Spoken Language" as an Exchange Format within
CLARIN and beyond*
Hanna Hedeland and Thomas Schmidt 100

Curation Criteria for Multimodal and Multilingual Data: A Mixed Study within the Quest Project
Amy Isard and Elena Arestau 105

*Flexible Metadata Schemes for Research Data repositories - The Common Framework in Dataverse
and the CMDI Use Case*

CLARIN Knowledge Centre for Belarusian text and speech processing (K-BLP)

Yuras Hetsevich

UIIP of NASB,
Minsk, Belarus
yuras.het-
sevich@gmail.com

Jauheniya Zianouka

UIIP of NASB,
Minsk, Belarus
evgeniakacan@gmail.com

David Latyshevich

UIIP of NASB,
Minsk, Belarus
david.latyshe-
vich@gmail.com

Mikita Suprunchuk

Minsk State Linguistic Uni-
versity, Belarus
ms@philology.by

Valer Varanovich

Belarusian State University,
Minsk, Belarus
gamrat.vvv@gmail.com

Katerina Lomat

UIIP of NASB,
Minsk, Belarus
katerina.lo-
mat@gmail.com

Abstract

This paper represents CLARIN Knowledge Center for Belarusian text and speech processing (K-BLP) which is based at the Speech Synthesis and Recognition Laboratory, the United Institute of Informatics Problems of the National Academy of Sciences of Belarus, Minsk. The CLARIN Knowledge Centre for Belarusian text and speech processing is part of the CLARIN ERIC, which holds the European ESFRI (European Strategy Forum on Research Infrastructures) certification as a landmark research infrastructure.

1 Introduction

The Speech Synthesis and Recognition Laboratory of the United Institute of Informatics Problems of the National Academy of Sciences of Belarus (<https://ssrlab.by>) established K-BLP center (Figure 1). It provides users with knowledge for text, speech and other data processing for Belarusian, Russian, and English. The K-BLP center proposes tools for text, speech and other data processing for languages, especially for the Belarusian language. The center also offers wide-ranging user support, guidelines and instructions for each service and material.

We are committed to widen the access to Belarusian developments in the computational linguistics environment and popularize our tools within the Republic of Belarus and abroad (Figure 2). It is very important to support available tools and promote them out to improve and facilitate the access for researchers in humanities and social sciences that contributes to wide-ranging user support, guidelines and instructions for each service. The main target audience of K-BLP are researchers in humanities and digital humanities with an interest in different aspects of computational linguistics and natural language processing.

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: <http://creativecommons.org/licenses/by/4.0/>.



Figure 1. CLARIN Certificate of the Belarusian centre

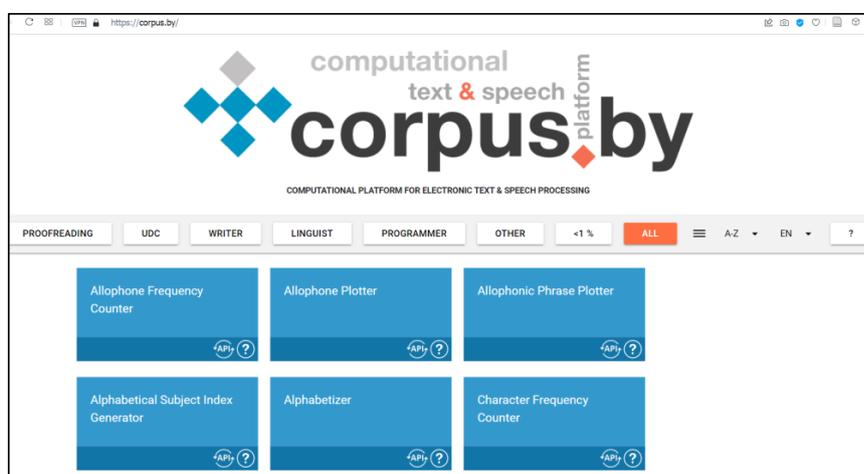


Figure 2. Overview of Belarusian text, speech and other data processors

2 K-BLP center initial activities

K-BLP was formed in September 2020 by the Speech Synthesis and Recognition Laboratory of UIIP NASB. Step by step, it started the process of CMDI metadata creation for all online resources, which means that part of the services is now available via the VLO. Currently, our centre offers Data processing services and tools (the corpus.by platform which includes over 65 services (Dzienisiuk, 2020), a speech intonation analyzer and trainer IntonTrainer (Lobanov, 2019), Belarusian NooJ module for convenient processing of Belarusian language via NooJ linguistic development environment), tutorials and exercises. All provided services can also be accessed through the links directly via <http://www.corpus.by/>. More information is available on the Speech Synthesis and Recognition Laboratory of UIIP NAS Belarus website.

The Laboratory works on such main scientific research directions as digitization of cultural heritage, high-quality text-to-speech synthesis, robust recognition of discrete and continuous word sequences, computer systems for the rehabilitation of people with hearing and vision disabilities. Except this, we work with systems, programs and platforms for processing big data, universal algorithms for stationery, online and mobile platforms for asynchronous input and output storing and issuing information from different platforms, semi-automatic systematization and processing of data by administrators of target programs (Figures 3–5). Our staff also uses the approaches to processing audio and text forms of speech, which is often found in the development of modern systems that work with the input and output of large-size speech (BigData) on different platforms.

We intend to create and maintain user infrastructure to support the sharing, use and sustainability of big data and tools for research in computational linguistics, the humanities and social sciences. Almost all our digital resources are open, free and available to scholars, researchers and scientists from all spheres through single sign-on access.

All products are made to solve the problems of developing algorithms, resources and methods of Internet input and Internet output of speech, saving and systematizing large volumes of speech. The results can be adapted for wide use in applied and practice-oriented research that requires processing large amounts of data at different levels.

https://corpus.by/VoicedElectronicGrammaticalDictionary/?lang=en

корпус	корпус	[кóрпус]	['kɔrpus]	NNIMO	назоўнік	Voice!
корпус	корпус	[кóрпус]	['kɔrpus]	NNIMA	назоўнік	Voice!

According to dictionary sbm2012initial (1)

Word	Transcription	IPA	Category	Voice
корпус	[кóрпус]	['kɔrpus]		Voice!

According to dictionary noun2013 (2)

Word	Transcription	IPA	Tag	Voice
корпус	[кóрпус]	['kɔrpus]	NMN1	Voice!
корпус	[кóрпус]	['kɔrpus]	NMA1	Voice!

According to dictionary asbm2017 (1)

Word	Transcription	Voice
корпус	[кóрпус]	Voice!

Figure 3. Voiced Electronic Grammatical Dictionary

computational linguistics & services
corpus.by Text-to-Speech Synthesizer ? English

Please input a stressed text
Primary stressed vowel must be marked by '+' or '^', a secondary stressed vowel – by '˘' or '˙'. To mark two words as one phonetic word use '' or '˘'.
For example: Паўночна-заходні вятры+нка садзьму+ў+бы ўсе+ лісце на+чы+спе, але+п'ятым калі+сцы. or Паўночна-заходні вятры+ка садзьму+ў_бы ўсе лісце на_воі+пе, але_п'ятым калі+сцы.

Лі+к.
адзі+н.
два+.
тры+.
чаты+ры.
пя+ць.
шэ+сць.
се+м.
во+сем.
дзе+вяць.

Беларуская AlesiaBel Show log information

Generate synthesized speech!

0:00 / 2:57

[download the generated speech file](#)

Figure 4. Text-to-Speech Synthesizer

https://corpus.by/ShortUSpellChecker/?lang=en

Perhaps, here should be «У» or «у»:

There was a letter	Comment
«а у»: ...Мама у трауры....	(«у» after the vowel «а» without a punctuation mark)
«А у»: ...А у іх ёсць пчолы....	(«у» after the vowel «А» without a punctuation mark)
«а» у»: ...«Рама» у краме....	(«у» after the vowel «а» without a punctuation mark)
«а-у»: ...На Украіне паўднёва-усходні вечер....	(«у» after the vowel «а» and a hyphen)
«ау»: ...Сястра ёсць аўсянку....	(«у» after the vowel «а»)
«І У»: ...ЛЮДЗІ УСІХ КРАІН, СЯБРУЙЦЕІ...	(«У» after the vowel «І» without a punctuation mark)

Perhaps, here should be «У» or «у»:

There was a letter	Comment
«т ў»: ...Кот ў ботах....	(«ў» after the consonant «т» without a punctuation mark)
«т» ў»: ...«Брат» ў космасе....	(«ў» after the consonant «т» without a punctuation mark)
«Ў»: ...На Украіне паўднёва-усходні вечер....	(CAPITAL «Ў» IS ONLY ALLOWED IN A TEXT WHERE ALL WORDS ARE WRITTEN IN CAPITAL LETTERS)
«м-ў»: ...Усім-ўсім пра ўсё распавядзем!...	(«ў» after the consonant «м» and a hyphen)
«бў»: ...Тата любіць бульбу....	(«ў» after the consonant «б»)

Figure 5. Non-syllable U Spell Checker: [u] or [w]

One more task is to provide a user-friendly overview of the available tools for researchers as well as to organize the overviews of developed methods and algorithms according to the types of data in the resources and listings sorted by language. Our team has great experience in accumulating big data in different formats and platforms. There are specialists in programming, front- and back-end development, project managers, computational linguists and philologists. We are open to create and develop new resources, tools, algorithms and methods according to users' demands.

3 K-BLP's main aims within CLARIN ERIC Research Infrastructure

The main task of K-BLP Center is to extend our resources and tools of natural language processing and organize them according to the types of data within the CLARIN Resource Families in the examples of other Resource families (cf. Franciska, 2020). Increasing the interest in Belarusian developments in computational linguistics and popularizing available tools and resources are dominant directions of K-BLP. To follow these aims, we should widen the number of scientific organizations of K-BLP (except the UIIP of NASB), add new resources and structuralize our Belarusian services within CLARIN classification. It is very important to promote available resources to facilitate access for researchers. That is why we propose wide-ranging user support, guidelines and instructions for each service. We also plan to create and maintain new tools for electronic text and speech processing in the Belarusian language.

Nowadays K-BLP has main strategic priorities such as:

1. To attract other scientific organizations and institutes with research centers for computer processing of the Belarusian language to widen K-BLP (such organizations as Belarusian State University, the Center for the Belarusian Culture, Language and Literature researches of the National Academy of Sciences and other).
2. To expand K-BLP with such resources as new Belarusian corpora (at least 3), dictionaries (nearly 5-7 items) and other tools for computer processing of Belarusian text and speech information (5-7 tools).
3. To annotate and systematize new resources and tools as consistent with description of all resources disposed in other CLARIN ERIC centers.
4. To optimize existing resources and tools in K-BLP according to CLARIN ERIC classification of resources.

5. To organize the overviews of developed Belarusian tools according to the types of data in the resources and listings sorted by language.
6. To provide a user-friendly overview of the available Belarusian language tools in the CLARIN infrastructure for researchers from digital humanities, social sciences and human language technologies.
7. To create and maintain an infrastructure to support the sharing, use and sustainability of Belarusian language data and tools for research in the humanities and social sciences.

We hope to implement our plans listed above in the near future with the help of CLARIN ERIC.

4 Conclusion

Building and running a distributed knowledge center K-BLP for computational linguistics and natural language processing of Belarusian requires samples, text descriptions, demos, courses and possible contacts with specialists of natural language approaches of Belarusian.

K-BLP provides knowledge about tokenization, morphological analysis, voiced electronic grammatical dictionaries, part-of-speech tagging, frequency counting, spell checking, text classification and other approaches used in speech and text processing. It offers special courses in language processing, data analysis and collecting research data for the fast entrance of humanities and others into the digital world of Belarusian data processing.

We are aimed at collecting Belarusian-language linguistic and computer resources for manual and automatic processing in one unit for popularizing the Belarusian language as much as possible. There is a variety of developments in Belarusian, but they are not in the public domain. For this, we want to conduct research in computational linguistics and modern standard Belarusian language and represent them within the K-BLP Center. The future idea is to participate with other CLARIN centres in joint European projects. The plan is to prepare main services and tools from “Computational platform for electronic text & speech processing www.corpus.by” for CLARIN Virtual Language Observatory.

The Speech Synthesis and Recognition Laboratory organises several courses in universities to educate students and researchers in computer linguistics. Several education online materials in English were prepared, such as “Lab 0 – How to be acquainted with text and speech processing services in 10 days?”. Introduction into CLARIN project could be presented here, too. All this will allow the introduction of different tools for computational processing of Belarusian for all who are interested in it including foreign scientists and partners.

References

- Lobanov, B. and Zhitko, V 2019. Software Subsystem Analysis of Prosodic Signs of Emotional Intonation. *Speech and Computer: 21st International Conference, SPECOM 2019, Istanbul, Turkey, August 20–25, 2019, Proceedings* / Eds. Albert Ali Salah, Alexey Karpov, Rodmonga Potapova. Springer: 280–288.
- Dzienisiuk, D. A., Zianouka, Ja. S., Drahun A. Je. [et al.]. 2020. Platforma dlia apracouki tekstavaj i hukavoj infarmacyi dlia roznych tematycznych damienau bielaruskaj movy. *Yazykovaya lichnost' i ÷effektivnaya komunikatsiya v sovremennom polikul'turnom mire : materialy VI Mezhdunar. nauch.-prakt. konf., posvyashch. 100-letiyu Belarus. gos. un-ta, Minsk, 29–30 okt. 2020 g. / Belorus. gos. un-t ; redkol.: S. V. Vorobyeva (gl. red.) [i dr.]. – Minsk : BGU: 69–74.*
- Franciska, J., Maegaard, B., Fišer, D. [et al.] 2020. Interoperability in an Infrastructure Enabling Multidisciplinary Research: The case of CLARIN. *Proceedings LREC 2020, 12th International Conference on Language Resources and Evaluation, ELRA*. Mode of access: <https://www.aclweb.org/anthology/2020.lrec-1.417>. Date of access: 02.06.2021.