# CLARIN Knowledge Centre for Belarusian Text and Speech Processing (K-BLP)

**Yuras Hetsevich**
UIIP of NASB,
Minsk, Belarus
yuras.hetsevich@gmail.com

**Jauheniya Zianouka**
UIIP of NASB,
Minsk, Belarus
evgeniakacan@gmail.com

**David Latyshevich**
UIIP of NASB,
Minsk, Belarus
david.latyshevich@gmail.com

**Mikita Suprunchuk**
Minsk State Linguistic
University, Belarus
suprunchuk@mail.ru

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

## Abstract

This paper represents the CLARIN Knowledge Centre 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. Services for text and speech processing, which were developed by the Laboratory, are presented in the article.

## 1 Introduction

Today, computer technologies are developing rapidly. They capture all new areas of life and fields of knowledge, including those related to language and the transfer of knowledge. For the development of machine dictionaries, translators, search engines and databases, text corpora are increasingly being used. The creation of a corpus can be carried out in different ways, methods, and stages. All of them are quite laborious and require knowledge in linguistics and programming. Proofreading and verification of texts are especially time and human resources consuming stages. In the case of parallel corpora, the problem of sentence alignment is added to this. To solve these and similar problems, a lot of work is being done in the Speech synthesis and recognition laboratory of the United Institute of Informatics Problems of the National Academy of Sciences of Belarus (SSRLab laboratory, https://ssrlab.by).

The SSRLab laboratory established the K-BLP Centre in 2020. It provides users with knowledge for text, speech and other data processing for Belarusian, Russian, and English. The K-BLP Centre proposes tools for text, speech and other data processing for languages, especially for the Belarusian language. The centre 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 1). It is very important to support available tools and promote them 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. Primary 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.

Yuras Hetsevich, Jauheniya Zianouka, David Latyshevich, Mikita Suprunchuk and Valer Varanovich 2022. CLARIN Knowledge Centre for Belarusian Text and Speech Processing (K-BLP). *Selected papers from the CLARIN Annual Conference 2021. Ed. by Monica Monachini and Maria Eskevich.* Linköping Electronic Conference Proceedings 189, pp. 46–55. DOI: https://doi.org/10.3384/9789179294441

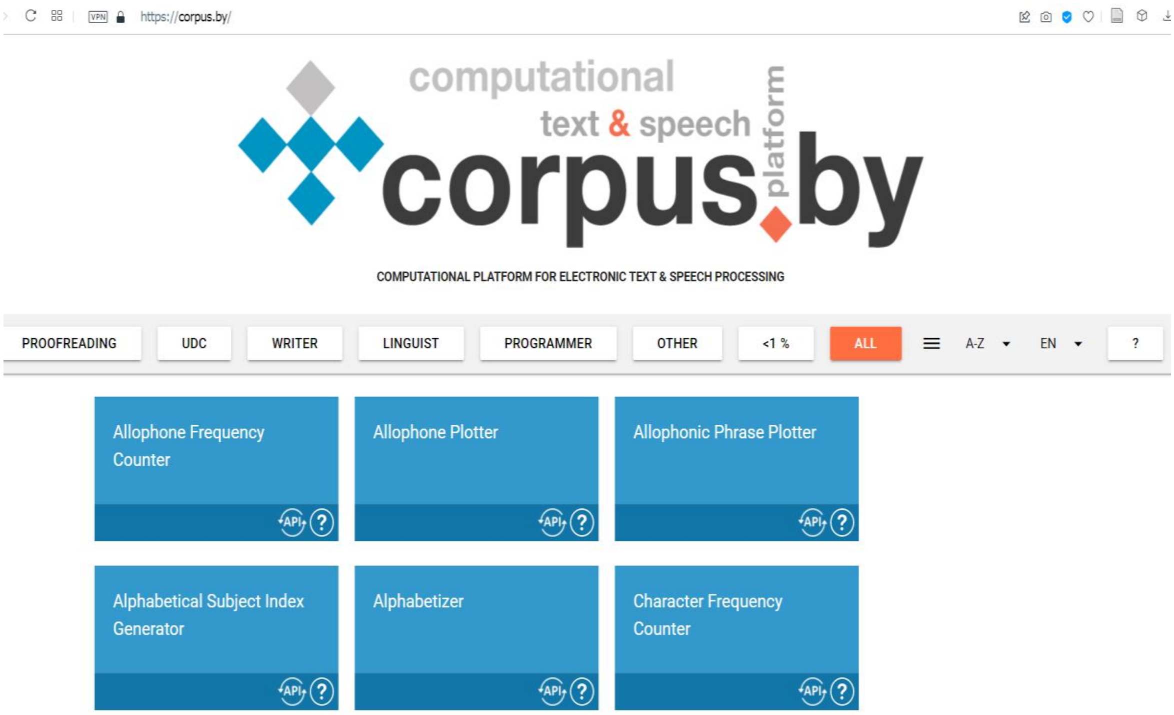

https://corpus.by/
computational
text & speech
corpus.by
platform
COMPUTATIONAL PLATFORM FOR ELECTRONIC TEXT & SPEECH PROCESSING
PROOFREADING
UDC
WRITER
LINGUIST
PROGRAMMER
OTHER
<1 %
ALL
A-Z
EN
?
Allophone Frequency Counter
Allophone Plotter
Allophonic Phrase Plotter
Alphabetical Subject Index Generator
Alphabetizer
Character Frequency Counter
API
?

Figure 1. Overview of Belarusian text, speech and other data processors

Next, we will demonstrate a number of services and tools that are used by SSRLab when preparing text corpora. Most of them are developed in the laboratory. Some programs, such as NooJ (Silberztein 2016; NooJ), were created by other people or organizations, but the laboratory offered it as a tool for collecting and processing Belarusian text information. Developing of several services was supported as part of a new CLARIN project in 2021 “Preparation of available K-BLP tools and resources for the metadata ingestion into CLARIN virtual language observatory (VLO) and representation in CLARIN Resource Families”.

## 2 K-BLP's Main Aims within CLARIN ERIC Research Infrastructure

The main task of the K-BLP Centre is to extend our resources and tools of natural language processing and organize them according to the data within the CLARIN Resource Families in the examples of other resource families (cf. de Jong, 2020). Increasing the interest in Belarusian developments in computational linguistics and popularizing available tools and resources are the main 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 the 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.

At present K-BLP has main strategic priorities such as:

1. To attract other scientific organizations and institutes with research centres for computer processing of the Belarusian language to widen K-BLP (such organizations as Belarusian State University, the Centre 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 (approx. 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 a description of all resources deposited in other CLARIN ERIC centres.

4. To optimize existing resources and tools in K-BLP according to the 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.

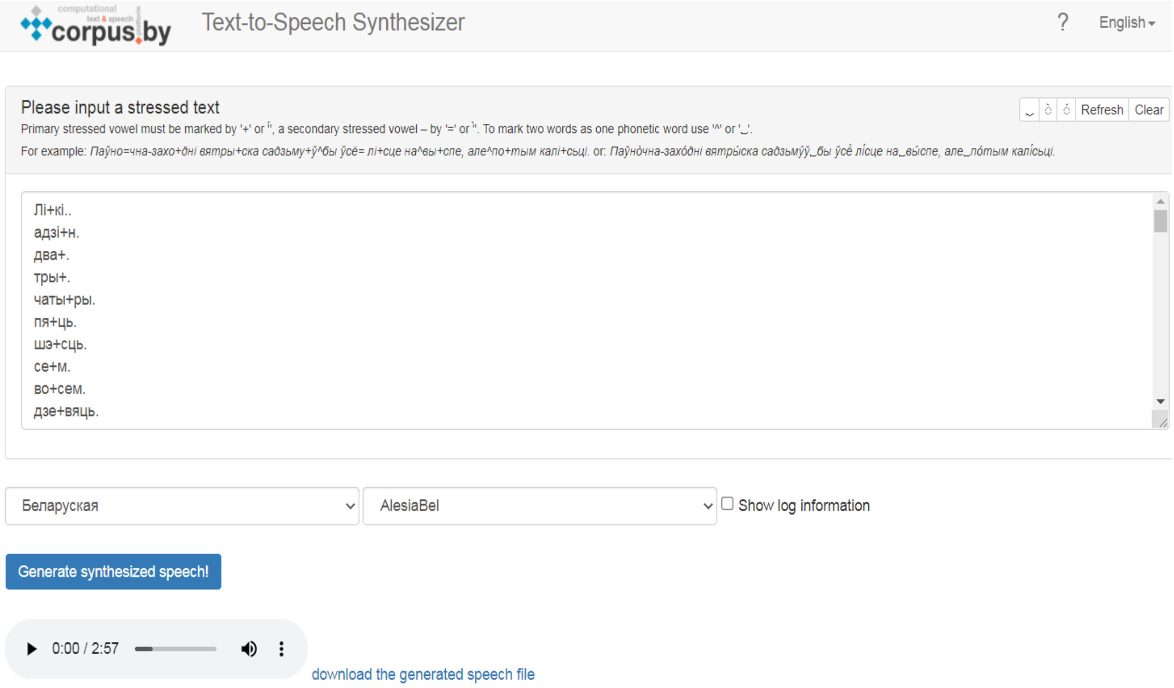
7. To create and maintain an infrastructure to support the sharing, use and sustainability of Belarusian language data and tools for research in the humanities and social sciences.

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

## 3 K-BLP Centre Initial Activities

The Speech synthesis and recognition laboratory of UIIP NASB established K-BLP Centre in September 2020. Step by step, it started the process of CMDI metadata creation for all online resources. So the part of the services is now available via the VLO. Currently, our centre offers data processing services and tools computational platform for electronic and speech processing platform which includes over 65 services (Dzienisiuk, 2020), a speech intonation analyser 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/ link. Detailed information is available on the Speech synthesis and recognition laboratory of UIIP NAS Belarus web-site.

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. In addition, 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 2–4).


computational text & speech platform
corpus.by
? English
‿ ò ó Refresh Clear
Generate synthesized speech!
0:00 / 2:57
download the generated speech file

Text-to-Speech Synthesizer

Please input a stressed text

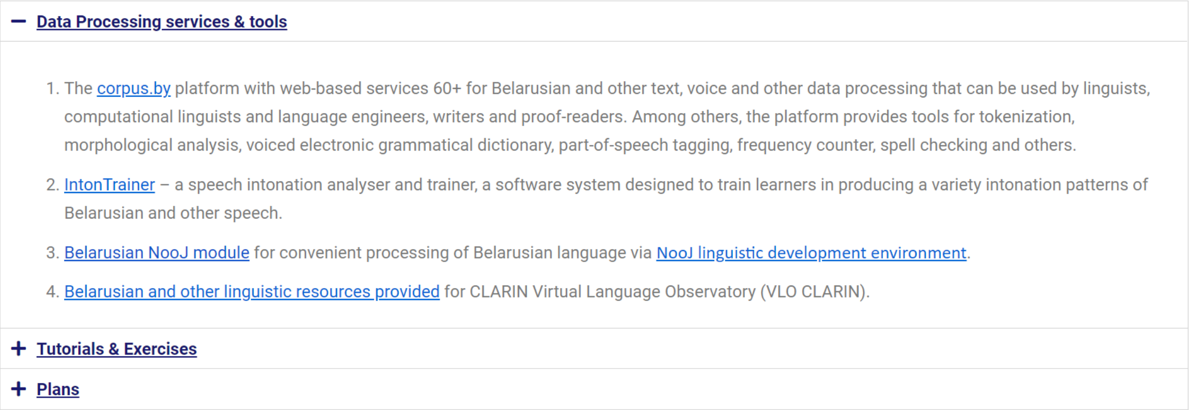
Primary stressed vowel must be marked by '+' or ', a secondary stressed vowel – by '=' or '. To mark two words as one phonetic word use '^' or '‿'.

For example: *Пайно=чна-захо+дні вятры+ска садзьму+ў^бы ўсё= лі+сце на^вы+спе, але^по+тым калі+сьці.* or: *Пайно̀чна-захóдні вятры́ска садзьмýў‿бы ўсѐ лíсце на‿вы́спе, але‿пóтым калíсьці.*

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

Беларуская | AlesiaBel | □ Show log information

Figure 2. Text-to-Speech Synthesizer



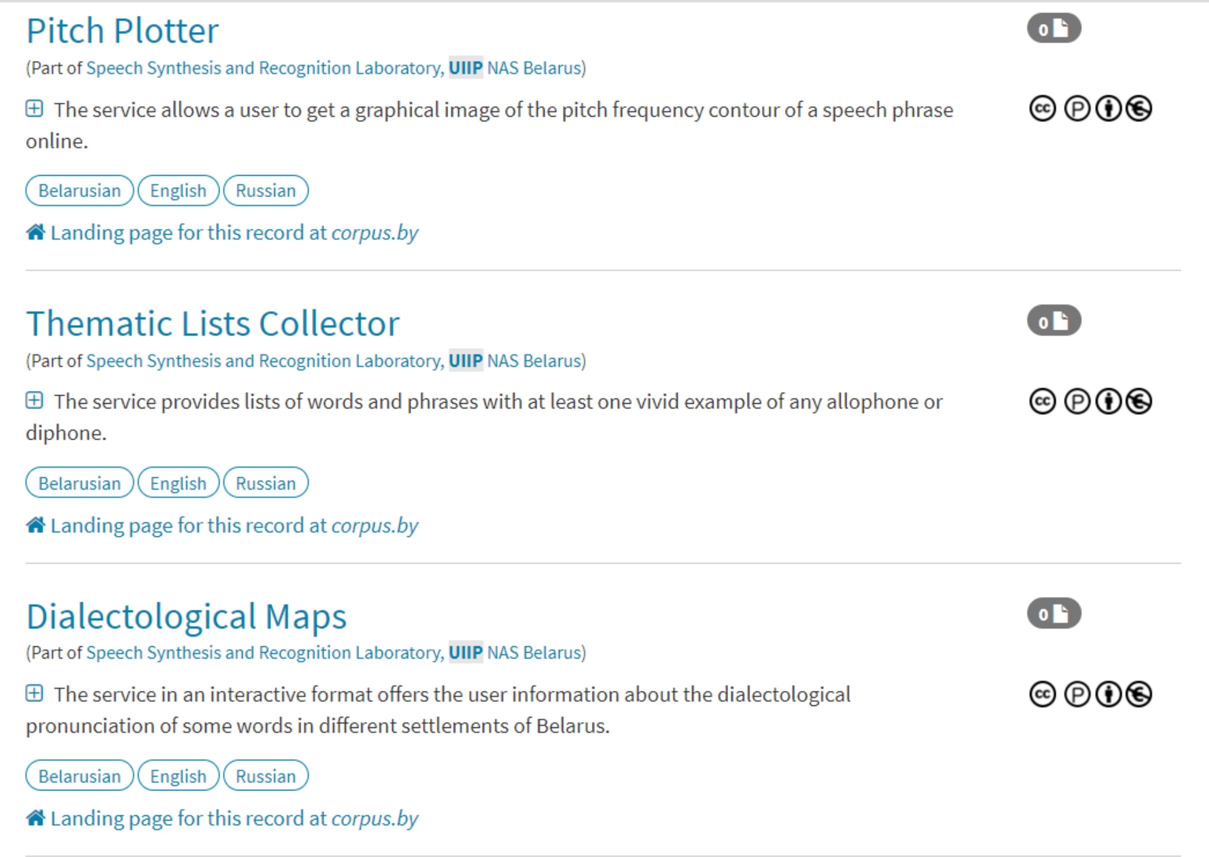
— Data Processing services & tools

1. The corpus.by platform with web-based services 60+ for Belarusian and other text, voice and other data processing that can be used by linguists, computational linguists and language engineers, writers and proof-readers. Among others, the platform provides tools for tokenization, morphological analysis, voiced electronic grammatical dictionary, part-of-speech tagging, frequency counter, spell checking and others.
2. IntonTrainer – a speech intonation analyser and trainer, a software system designed to train learners in producing a variety intonation patterns of Belarusian and other speech.
3. Belarusian NooJ module for convenient processing of Belarusian language via NooJ linguistic development environment.
4. Belarusian and other linguistic resources provided for CLARIN Virtual Language Observatory (VLO CLARIN).

+ Tutorials & Exercises

+ Plans

Figure 3. Actual materials of K-BLP



Pitch Plotter

(Part of Speech Synthesis and Recognition Laboratory, UIIP NAS Belarus)

⊞ The service allows a user to get a graphical image of the pitch frequency contour of a speech phrase online.

Belarusian English Russian

Landing page for this record at *corpus.by*

Thematic Lists Collector

(Part of Speech Synthesis and Recognition Laboratory, UIIP NAS Belarus)

⊞ The service provides lists of words and phrases with at least one vivid example of any allophone or diphone.

Belarusian English Russian

Landing page for this record at *corpus.by*

Dialectological Maps

(Part of Speech Synthesis and Recognition Laboratory, UIIP NAS Belarus)

⊞ The service in an interactive format offers the user information about the dialectological pronunciation of some words in different settlements of Belarus.

Belarusian English Russian

Landing page for this record at *corpus.by*

Figure 4. Tools for text processing in K-BLP

Our staff also uses the approaches to process audio and text forms of speech, which are often found in the development of modern systems that work with the input and output of large-scale 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.

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 considerable 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.

Certainly, the K-BLP Help Desk was organised to provide people from Belarus and foreign countries with information about CLARIN ERIC, about Belarusian language in general and computer tools for text researches. There are two main possibilities to apply for such information: via contact page on the web-site https://clarin-belarus.corpus.by/contacts/ or via contact email which is available on the platform corpus.by. We receive one or two inquires every month.

Besides, several employees who works at SSRLab teaches computer linguistics and other courses ("Problems of AI", "Computer technologies in linguistics", "Speech synthesis and recognition", "Automatic translation", etc.) at Belarusian State University and Minsk State Linguistic University. Therefore, students (about 90 per year) ask a lot of questions during classes and while preparing their home works and projects. The most frequent questions are the following: How to get acquainted with computer linguistics, esp. with Belarusian computer linguistics? Where one can find digital resources on the Belarusian language and literature? Where one can learn about history of Belarus and the Belarusian language, about traditions and culture?

We try to respond by ourselves, there is a special collection with most useful resources on the web-site about Clarin https://clarin-belarus.corpus.by/materials/. A lot of services are presented on the platform www.corpus.by and on the site of our colleagues https://bnkorpus.info/index.en.html. There is a nice cooperation between SSRLab and the Institute of linguistics named after Yakub Kolas of the Belarusian Academy of Sciences http://iml.basnet.by/, so their consultations are possible, too.

## 4 Optimization of Information Pre-Processing for the Corpora Creation

Various types of the text processing are important directions of the Belarusian CLARIN Knowledge Centre activities: speech to text and text to speech conversion, spell checking, transliteration, transcription, etc. One of them is NooJ platform. The NooJ application is a shell for word processing and a convenient tool to compile a corpus of texts. It was developed by Max Silberztein (Silberztein 2016), professor of Université de Franche-Comté, France. (Cf. section 5 below.)

At different stages of the corpus preparation, we had specific tasks. To solve them, as well as for other purposes, the laboratory staff developed a number of useful tools and services. Some resources are currently being improved as part of the 2021–2022 project "Preparation of available K-BLP tools and resources for the metadata ingestion into CLARIN virtual language observatory (VLO) and representation in CLARIN Resource Families". The following discussion will overview some of them.

When data is collected in large quantities, there may be texts or their fragments in different languages. In our case, it was important to select texts in one language, either only in Russian or only in Belarusian. To check that the text is written in the required language, it is useful to use the service (LanguageIdentifier).

The "Language Identifier" service was developed to the identify the language of the text which has been submitted to the input. For now, the service recognizes five languages: Belarusian, Russian, Ukrainian, English and German. The text language is identified by the service using the statistical method and the rule application method. The priority of "statistics over rules" or "rules over statistics" is determined by the position of a special toggle switch variable. The ability to change the position of this toggle switch is currently hidden from the user. However, if necessary, it can always be used by the developer. The sensitivity threshold of the algorithm, the minimum and maximum number of characters of the text to be processed can be just as easily changed. The plans for the improvement of the service include the ability to define several languages of multilingual text and the generation of statistics on the use of each individual language, the expansion of the language palette, using new identification rules. To access the "Language Identifier" service via the API, one needs to send an AJAX request of the POST type to the address https://corpus.by/LanguageIdentifier/api.php.

High quality of the created program largely depends on the source data. It is important to ensure that the collected texts of a compiled corpus, do not contain errors, typos, repetitions, or unnecessary information. One spell checker package for MS Office Word, LibreOffice, OpenOffice, Thunderbird and several web-browsers was developed by one of our colleagues (Praverka pravapisu). It corrects texts fairly well, but it requires special search, download and installation.

In contrast, in our projects the majority of texts are proofread by editors and proofreaders – members of the project team. In addition, all texts are automatically checked by a special free online service which was also developed by the laboratory staff (SpellChecker).

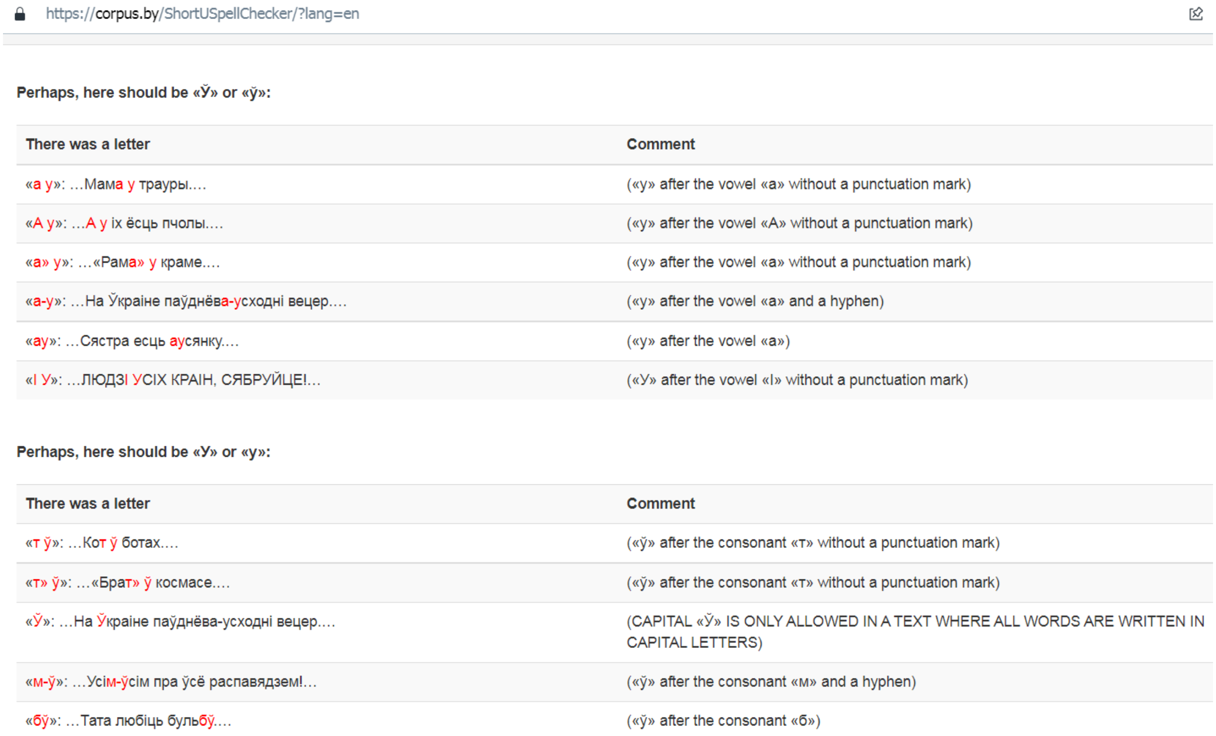
The service receives an electronic document that requires verification. By pressing the “Check it!” button, the service compares text words with words in attached dictionaries. The service qualifies the words of the input text found in at least one dictionary as spelled correctly and discards them. Words that are not found in dictionaries are qualified by the service as misspelled. The service displays them in a list in alphabetical order. Currently, the quality of the text proofreading is an integral requirement for many fields of activity, especially for communication between people and institutions. In addition, spelling-correct electronic text is necessary for proper functioning of computer systems of human-machine communications. The relevance of the service development is also determined by complicated access to processing tools for Belarusian-language texts. The proofreading of an electronic text by machine tools always remains relevant, since manual checking of texts by the user almost definitely means skipping mistakes.

The named service checks Russian and Belarusian documents. To check the Russian language, a well-known dictionary by Andrei Zaliznyak (Zaliznyak 2003) is used. To check the Belarusian language several large modern dictionaries are used, cf. the full list on the web-site (Spell Checker). In addition, the laboratory replenishes its own dictionary, where words that are not included in published editions are indicated because they are recent or used in narrow areas. Some of the mentioned dictionaries are being constantly enlarged.

Among several Belarusian services of spell checking, the “Spell Checker” service was created as one of the stages of preliminary text processing and normalization for a speech synthesizer. It is worth noting that this service covers the orthographic section of the spelling, but not grammar, syntax or punctuation. The correctness of word matching and punctuation is outside the competence of the service and remains for the user or other services that are also involved in the methodology of large electronic texts proofreading using the platform www.corpus.by services. “Spell Checker” service can process both small and large texts. For example, it successfully checked the spelling of legislative codes and literary works with a volume of about 470 000 characters with spaces.

It is important to mention another spell checking tool. There is a specific alteration in Belarusian orthography. The letter ***у*** and the sound [u] are used after consonants and punctuation marks, and after vowels the letter ***ў*** and the sound [w] are used instead (so called “non-syllable w” or “short w”). Besides this, the sound [w] and the letter ***ў*** alternate with the letters ***в***, ***л*** and sounds [v] and [l] depending on the place in the word and its origin. This alternation has certain peculiarities and limitations, so it was decided not to embed the control of this phenomenon into the general spell checking service, but to develop a separate tool – “ShortUSpellChecker” (Figure 5).

While searching for possible errors, the service not only determines whether the vowel or consonant is before “u/w”, but also analyzes characters that are not letters, if the letter “u” is at the beginning of a word. These characters directly influence the writing of a word. Not all words of the Belarusian language adhere to the general rules for writing the letter “Ў”. For this reason, the service provides the opportunity to use an exceptions dictionary (can be attached by a special box) or a user list of exceptions. The service processes a text, considering these sets of words. There are special rules for writing abbreviations with the letter “y” in the Belarusian language. To obtain accurate results (since the service does not distinguish abbreviations from other words automatically), the user is prompted to enter the abbreviations that appear in the text in the corresponding field. The service considers the following characters as punctuation marks: “,”, “.”, “:”, “;”, “!”, “?”, “–”, “—”, “(”, “)”. Symbols “[”, “]”, “{”, “}”, “_”, “%”, “№”, “#”, “^”, “$”, “@” and others are not punctuation marks for the processing algorithm of the service. A hyphen (“-”) is a punctuation mark (identified with a dash) only if it is surrounded by spaces on both sides.



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

**Perhaps, here should be «Ў» or «ў»:**

| 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 Ўкраіне паўднёва-усходні вецер.… | («у» after the vowel «а» and a hyphen) |
| «ау»: …Сястра есць аусянку.… | («у» after the vowel «а») |
| «І У»: …ЛЮДЗІ УСІХ КРАІН, СЯБРУЙЦЕ!… | («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 Ўкраіне паўднёва-усходні вецер.… | (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 бульбў.… | («ў» after the consonant «б») |

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

One more interesting service which is used to prepare data for the corpus is “Grammatical Dictionary Processor” (Grammatical). This service allows the user to receive previously loaded and converted to the required format lexicographic data of the grammar dictionary in the form of the HTML table, and to receive SQL instructions for creating a database that contains the entered information in a structured form.

Many text analysis-oriented systems need extensive and well-structured vocabulary databases – for example, automatic annotation and abstracting systems, systems of market analysis, legal linguistic examination. In addition, the vocabulary base can become the basis of commercial products – such as programs designed to help the user improve the grammar of the text he or she wrote, or popular entertainment applications that offer word games to the user. Filling such vocabulary databases (and especially filling grammatical dictionaries) is a very time-consuming and painstaking process. “Grammatical Dictionary Processor” service is designed to simplify and automate this process in the case of working with Belarusian-language data. Thus, the service devotes to provide additional support to strengthen the position of the Belarusian language in the electronic space.

The results of the “Grammatical Dictionary Processor” service were repeatedly applied in other tools of the Corpus.by platform.

The service processes texts only in Belarusian. It is available via the API too. The details are presented here: https://ssrlab.by/en/8071.

This tool is under development yet in a frame of the new project for 2022 year “Preparation of available K-BLP tools and resources for the metadata ingestion into CLARIN virtual language observatory (VLO) and representation in CLARIN Resource Families”.

It is planned that part Belarusian general corpus will comprise oral speech, recordings of speeches and spontaneous conversations. The “Thematic Speech Recognizer” program https://ssrlab.by/en/4962 was developed to speed up the decryption and transcription of audio files. It allows the user to convert speech to electronic text online. A phonogram no larger than 20 MB is given at the input to the service. It provides a recognized electronic text of the phonogram at the output. The soundtrack can be selected from the provided examples, downloaded to the service from the computer's hard drive in WAV format, and can also be recorded online.

Speech recognition has great scientific perspectives and wide possibilities of application in many “human-machine” systems, which are built on the basis of speech communication. There are other areas that are particularly in need of speech recognition services. For example, journalism, shorthand and many others. In particular, the recognition of Belarusian speech, which becomes possible with the help of this service, will allow the full development of Belarusian technical sciences, including robotics. At the moment, the service is a demonstration that recognizes the Belarusian language of the following thematic domains: clothes, cities, numbers, etc. The list of domains will be continued. The tool is implemented and works according to the instruction on creation of programs on the basis of CMU Sphinx (CMUSphinx). It is available via the API.

When texts are selected, it is often necessary to get their quantitative characteristics. For example, it is usually useful to make a list of wordforms used. The “Tokenizer” tool was created for this and similar tasks. The token is a wordform, e. g. *come, came, coming*, or *cat, cats*, or Belarusian cases of ‘hand’ *рука, руку, рукі, руцэ*, etc. “Tokenizer” is intended to locate tokens in the text that requires tokenization. It is sent to the service input. After its processing, the user receives a list of the extended tokens on the output. Figures and punctuation marks are processed too. The service handles Belarusian, Russian and English. It is available via the API. The details are presented here: https://ssrlab.by/en/5900. The tool is also being developed as part of the “Preparation of available K-BLP tools and resources for the metadata ingestion into CLARIN virtual language observatory (VLO) and representation in CLARIN Resource Families”.

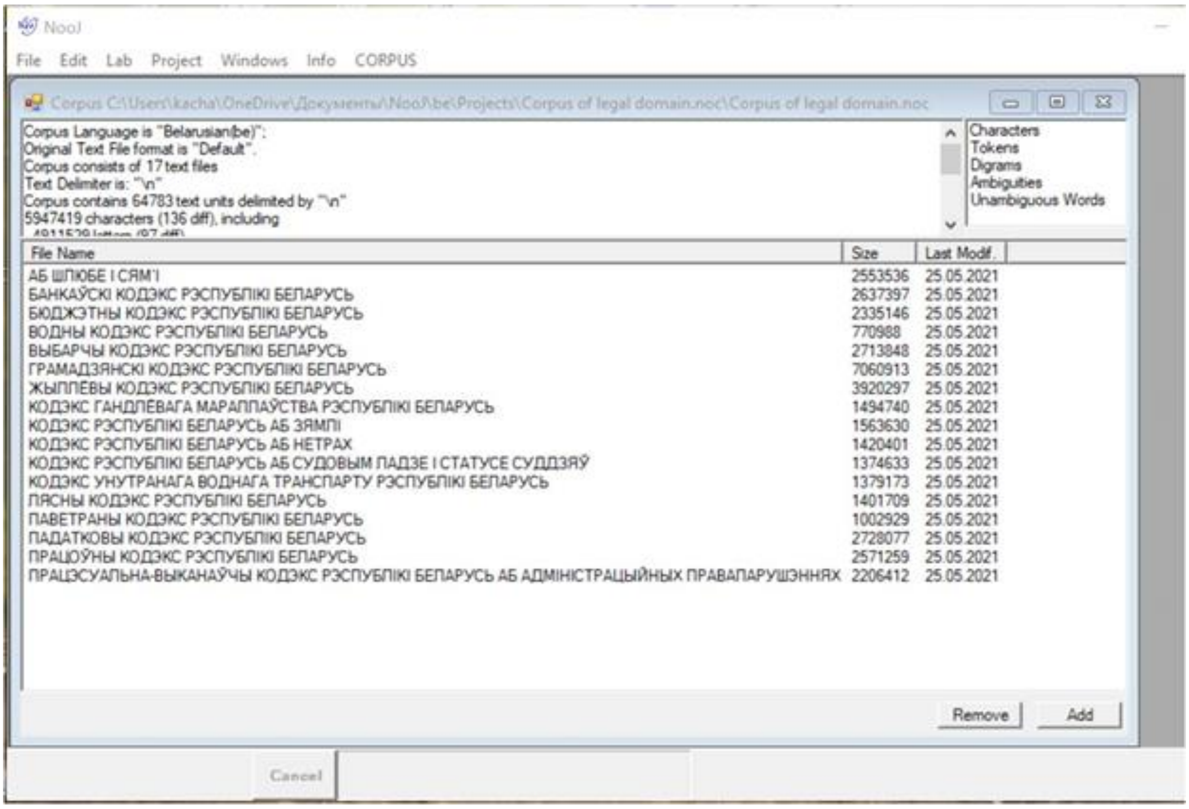
## 5 Work on Legal Texts

The Speech synthesis and recognition laboratory is working to create a parallel body of legal texts (Hetsevich 2021). The creation of a corpus of legal texts will allow solving several important tasks to improve the automatic processing of texts in Belarusian. First, the corpus will allow us to conduct comparative research and identify features of the Belarusian language in comparison with Slavic and other European languages. Second, it is possible to create various dictionaries, both monolingual and multilingual, as well as a linguistic knowledge base for the machine translation system. The corpus can also be a basis for creating a variety of morphological and syntactic grammars that can be used for automatic grammatical analysis of texts in the Belarusian language. Finally, with the help of such a corpus, it is possible to identify the features of the legal style in Belarusian.

For the created corpus the codes of the Republic of Belarus – the most important (together with the Constitution) legislative acts regulating civil legal relations in various spheres of public activity – were taken as a base. It is planned to present the texts of the codes in two official languages of the Republic of Belarus – Belarusian and Russian, and in the future – in other languages. At the beginning of 2022, 18 out of 26 codes were processed and uploaded to the project page, the total number of word forms in the existing building is about 1 million. The results of the work are at https://ssrlab.by/7804.

The corpus of codes was also compiled in the NooJ format (NooJ), and a trilingual Belarusian-Russian-English dictionary of legal terms was created on its basis (Figure 6).
Yuras Hetsevich and Sviatlana Hetsevich from SSRLab laboratory and Yauheniya Yakubovich from Universitat Autònoma de Barcelona created a module to gather Belarusian texts and process them on the basis of Nooj (Hetsevich Y. and Hetsevich S., 2012). The program helps to ‘develop linguistic resources to formalize various linguistic phenomena at the orthographical, lexical, morphological, syntactic and semantic levels, for any natural language’ (NooJ). In addition, with its help everyone can form their own text corpus, select concordances for the analysed words, search in accordance with different language parameters and get the necessary statistical information about certain linguistic facts.

The Budgetary Code, Water Code, Electoral Code, Civil Code, Housing Code, Tax Code, Marriage and Family Code, Forest Code, etc. have been translated till now. When 17 law codes were translated into Belarusian, a group of specialists compiled a unified corpus in NooJ format to create a dictionary of legal terms and expressions. The total number of word tokens is 1,043,018 and 731,584 word forms.


NooJ
File Edit Lab Project Windows Info CORPUS
Corpus C:\Users\kacha\OneDrive\Документы\NooJ\be\Projects\Corpus of legal domain.noc\Corpus of legal domain.noc
Characters
Tokens
Digrams
Ambiguities
Unambiguous Words
Remove
Add
Cancel

```
Corpus Language is "Belarusian(be)";
Original Text File format is "Default".
Corpus consists of 17 text files
Text Delimiter is: "\n"
Corpus contains 64783 text units delimited by "\n"
5947419 characters (136 diff), including
4911529 letters (97 diff)
```

| File Name | Size | Last Modif. |
|---|---|---|
| АБ ШЛЮБЕ І СЯМ'І | 2553536 | 25.05.2021 |
| БАНКАЎСКІ КОДЭКС РЭСПУБЛІКІ БЕЛАРУСЬ | 2637397 | 25.05.2021 |
| БЮДЖЭТНЫ КОДЭКС РЭСПУБЛІКІ БЕЛАРУСЬ | 2335146 | 25.05.2021 |
| ВОДНЫ КОДЭКС РЭСПУБЛІКІ БЕЛАРУСЬ | 770988 | 25.05.2021 |
| ВЫБАРЧЫ КОДЭКС РЭСПУБЛІКІ БЕЛАРУСЬ | 2713848 | 25.05.2021 |
| ГРАМАДЗЯНСКІ КОДЭКС РЭСПУБЛІКІ БЕЛАРУСЬ | 7060913 | 25.05.2021 |
| ЖЫЛЛЁВЫ КОДЭКС РЭСПУБЛІКІ БЕЛАРУСЬ | 3920297 | 25.05.2021 |
| КОДЭКС ГАНДЛЁВАГА МАРАПЛАЎСТВА РЭСПУБЛІКІ БЕЛАРУСЬ | 1494740 | 25.05.2021 |
| КОДЭКС РЭСПУБЛІКІ БЕЛАРУСЬ АБ ЗЯМЛІ | 1563630 | 25.05.2021 |
| КОДЭКС РЭСПУБЛІКІ БЕЛАРУСЬ АБ НЕТРАХ | 1420401 | 25.05.2021 |
| КОДЭКС РЭСПУБЛІКІ БЕЛАРУСЬ АБ СУДОВЫМ ЛАДЗЕ І СТАТУСЕ СУДДЗЯЎ | 1374633 | 25.05.2021 |
| КОДЭКС УНУТРАНАГА ВОДНАГА ТРАНСПАРТУ РЭСПУБЛІКІ БЕЛАРУСЬ | 1379173 | 25.05.2021 |
| ЛЯСНЫ КОДЭКС РЭСПУБЛІКІ БЕЛАРУСЬ | 1401709 | 25.05.2021 |
| ПАВЕТРАНЫ КОДЭКС РЭСПУБЛІКІ БЕЛАРУСЬ | 1002929 | 25.05.2021 |
| ПАДАТКОВЫ КОДЭКС РЭСПУБЛІКІ БЕЛАРУСЬ | 2728077 | 25.05.2021 |
| ПРАЦОЎНЫ КОДЭКС РЭСПУБЛІКІ БЕЛАРУСЬ | 2571259 | 25.05.2021 |
| ПРАЦЭСУАЛЬНА-ВЫКАНАЎЧЫ КОДЭКС РЭСПУБЛІКІ БЕЛАРУСЬ АБ АДМІНІСТРАЦЫЙНЫХ ПРАВАПАРУШЭННЯХ | 2206412 | 25.05.2021 |

Figure 6. Corpus of translations of legal texts prepared in NooJ

As the Belarusian NooJ module already had the general vocabulary "general_be.dic" it was decided to create an additional dictionary for general unknown words (nearly 150 word forms in initial form) and a law dictionary (nearly 200 forms). Every initial form (lemma) from the list of unknown general words was assigned a morphological class. This class shows the flexion features, i. e. how the word changes. Then the law dictionary in the Belarusian was compiled for NooJ format and now is available for further text processing with Belarusian texts of any domains. It is an addition to the main NooJ dictionary for the Belarusian language.

All terms of the Belarusian law dictionary Law_codes_be.dic were correlated with their Russian equivalents from the parallel corpus of a legal domain. The next step was their translation into English.

Next, several grammars were developed that show the possibilities of applying the established legal corpus in solving various problems of machine texts processing.

## 6 Conclusion

Building and running a distributed knowledge centre K-BLP for computational linguistics and natural language processing 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 tools, algorithms and methods 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.

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

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 and represent results within the K-BLP Centre. 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.

## References

Silberztein, M. 2016. *Formalizing Natural Languages*: *The NooJ Approach*. Wiley Eds. Hoboken, NJ, USA. 346 pp.

NooJ: A Corpus Processor. URL: https://www.nooj-association.org/.

Jong de, 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. URL: https://www.aclweb.org/anthology/2020.lrec-1.417.

Dzienisiuk, D. A., Zianouka, Ja. S., Drahun A. Je. [et al.]. 2020. Platforma dlia apracouki tekstavaj i hukavoj infarmacyi dlia roznych tematycnych damienau bielaruskaj movy. *Yazykovaya lichnost' i èffektivnaya kommunikatsiya v sovremennom polikul'turnom mire*: materialy VI Mezhdunar. nauch.-prakt. konf., posvyashch. 100-letiyu Belorus. 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.

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, A. Karpov, R. Potapova. Springer: 280–288.

LanguageIdentifier. URL: https://corpus.by/LanguageIdentifier/?lang=en.

Praverka pravapisu. Праверка правапісу // Беларускі N-корпус. URL: https://bnkorpus.info/spell.html.

Spell Checker. URL: https://corpus.by/SpellChecker/?lang=en.

Zaliznyak, A. A. 2003. Зализняк, А. А. Грамматический словарь русского языка: Словоизменение [Grammaticheskij slovar' russkogo yazyka: Slovoizmenenie]. Москва, Рус. словари [Moscow, Rus. slovari]. 800 pp.

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

GrammaticalDictionaryProcessor. URL: https://corpus.by/GrammaticalDictionaryProcessor/?lang=en.

CMUSphinx. URL: https://cmusphinx.github.io/wiki/about/.

Hetsevich, Yu. 2021. Creation of a legal domain corpus for the Belarusian NooJ module: texts, dictionaries, grammars / Yu. Hetsevich, Ya. Zianouka, V. Varanovich, M. Suprunchuk, Ts. Prakapenka, Dm. Dzenisiuk. *15th International Conference NooJ 2021*: book of abstracts / Virtual conference ; ed. M. Bigey [et al.]. Besançon, France: 36-37.

Hetsevich, Y. and Hetsevich, S. 2012. Overview of Belarusian and Russian dictionaries and their adaptation for NooJ. *Automatic Processing of Various Levels of Linguistic Phenomena*: selected papers from the NooJ 2011 Intern. conf. / eds. K. Vučković, B. Bekavac, M. Silberztein. Newcastle, Cambridge Scholars Publishing: 29–40.