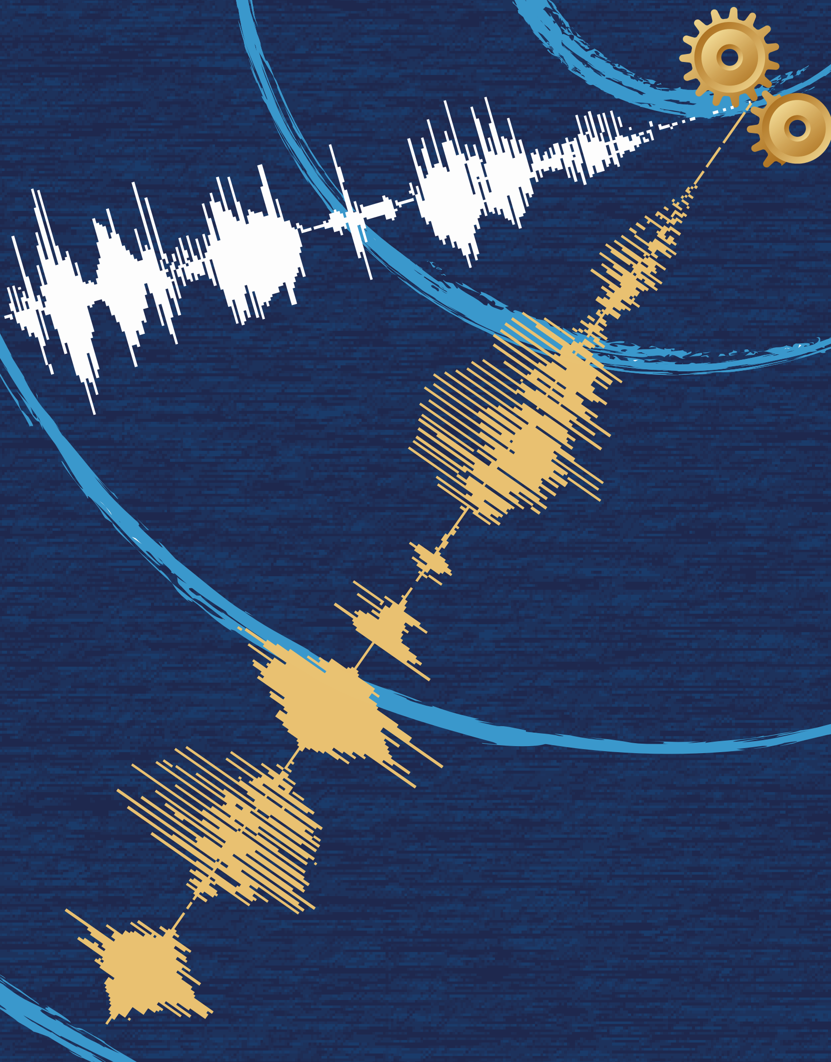
Научно-практический журнал

РЕЧЕВЫЕ ТЕХНОЛОГИИ



Speech technology

1–2' 2021

Содержание

Редакция:
Редактор: *Татьяна Иванова*
Корректор: *Людмила Асанова*
Дизайн: *Анна Ладанюк*
Вёрстка: *Андрей Кинсбурский*

Адрес редакции: 109341, Москва, ул. Люблинская, д. 157, корп. 2.
Тел.: (495) 345 52 00

Подписано в печать 26.11.2021. Формат 60×90⅛. Бумага офсетная. Печать офсетная.
Печ. л. 14,25. Заказ № 21С07. Издательский дом «Народное образование».
Отпечатано в типографии НИИ школьных технологий.
109341 Москва, ул. Люблинская, д. 157, корп. 2 Тел.: (495) 345 52 00/59 00.

Компьютерная платформа для обработки электронного текста и речи на белорусском, русском и английском языках

Герцевич Ю.С., *Минск, Республика Беларусь,*
yuras.hetsevich@gmail.com

Зеновко Я.С., *Минск, Республика Беларусь,*
evgeniakacan@gmail.com

Маевский С.С., *Минск, Республика Беларусь,*
maevskiiss@gmail.com

Денисюк Д.А., *Минск, Республика Беларусь,*
d.denissyuk@gmail.com

Драгун А.Е. *Минск, Республика Беларусь,*
ndrahun@gmail.com

Данная статья описывает онлайн-платформу corpus.by для обработки текстовой и речевой информации, которая предлагает комплекс сервисов и инструментов по автоматической обработке электронных текстов и аудиофайлов на белорусском, русском и английском языках. Подробно рассмотрены назначение платформы, ее структурные части, а также ряд задач, которые ставятся перед разработчиками онлайн-платформы для ее совершенствования и оптимизации.

• компьютерные технологии • платформа для обработки текстовой и речевой информации • автоматическая обработка • интернет-сервис • синтезатор речи по тексту.

Introduction

Over the past half-century, significant scientific and practical results have been obtained in the field of computational linguistics. One of the main issues that scientists face now is the problem of automated text and speech processing, which has become particularly relevant. The high rate of available information forces us to improve the ways of its processing, to implement partial or complete automation of these procedures. It is easy to see that the main and most popular way of presenting information is the text in natural language. Therefore, one of the important areas of computer technologies is the development of systems that can automatically process such kind of information.

Employees of the Speech Synthesis and Recognition Laboratory of Joint Institute of Informatics Problems of the National Academy of Sciences of Belarus [1] have worked out ***Computational Platform for Electronic Text and Speech Processing www.Corpus.by*** [2], which helps to solve many issues related to the processing of electronic texts and speech signals. For the last 60 years, the main activity of the laboratory has been the development of speech synthesis systems. Therefore, most of the developed resources are built into the Belarusian text-to-speech synthesizers and cover a number of tasks that must be solved by the synthesizer. Except this, 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. 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. Our staff uses the approaches to process 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.

Main part

The platform ***Corpus.by*** is a set of different tools that are aimed at the target audience (programmers, linguists, philologists, students, teachers, etc.). The services provide easy and sustainable access to electronic text and speech processing and tools for analyzing, detecting, researching, or combining data sets in Belarusian, Russian, and English. The principle of *corpus.by* is concluded according to "*input data-output data*": the user enters text information and receives results at the output. The platform presents tools for tokenization, morphological analysis, vocalization of an electronic grammar dictionary, search for homonyms, counting the frequency of characters and words, spell checking, speech synthesizer, speech and emotion recognition services, and much more. The total number of software implementations presented in the platform is 69 (Figure 1). All products are made to solve the problems of developing algorithms, resources and methods of Internet input and Internet output data, saving and systematizing large volumes of information. The results can be adapted for wide use in applied and practice-oriented research that requires processing large amounts of data at different levels.

Each of the services solves its tasks while improving the functionality of the Belarusian-language text-to-speech synthesizer (BTTS), which is also included in the list of services. The development approach allows the user to enter test data, launch the service with a single click, and view the results.

The services are grouped into thematic domains for more convenient use in specific practical areas (proofreading, UDC, writer, linguist, programmer, other) to meet the needs of target people. All thematic domains can be found on the official page of the platform (Figure 2). For example, the

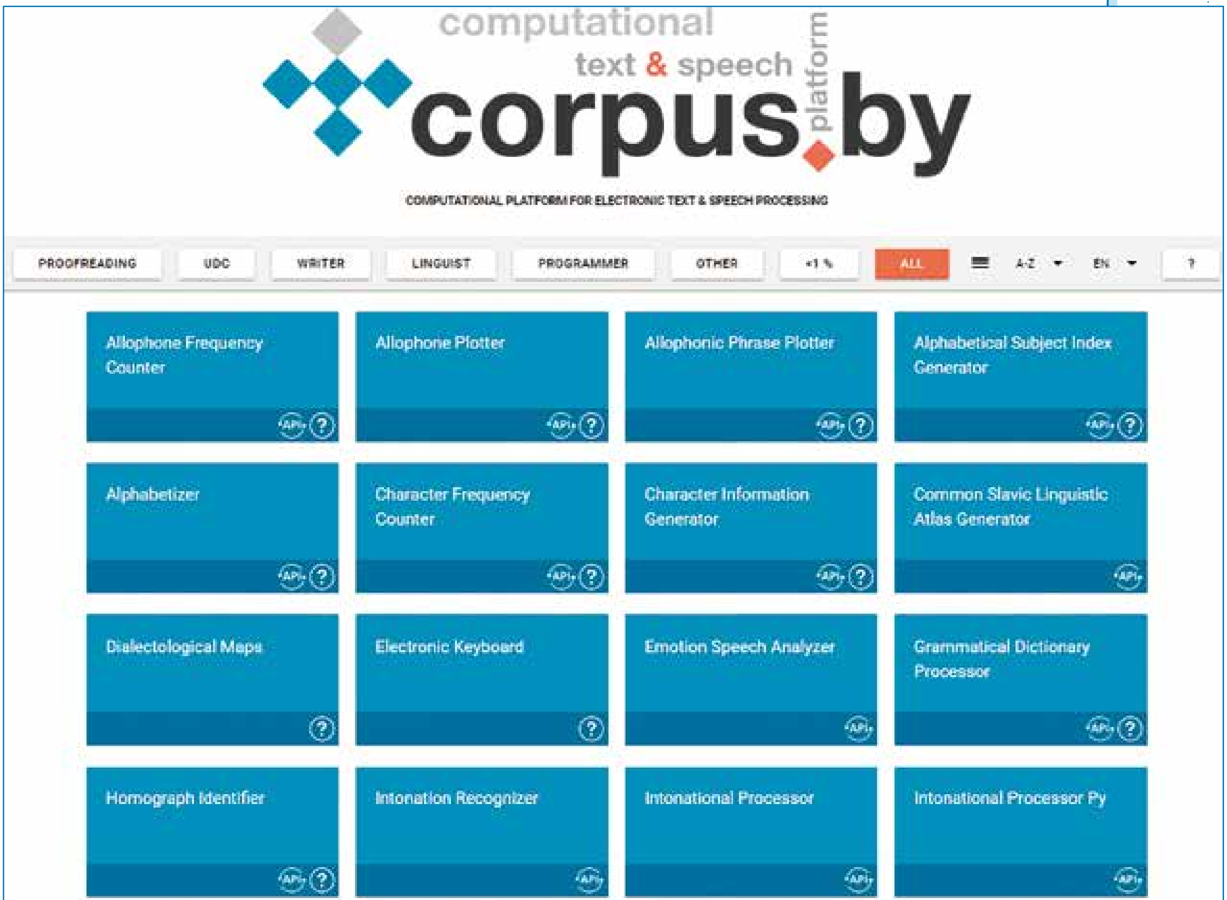


Figure 1. Computational Platform for Electronic Text and Speech Processing www.Corpus.by

Figure 2. The list of services that are useful for a linguist

thematic domain "Linguist" offers services for processing text and speech information, phonetic phenomena of the language, tools for determining, analyzing and searching for various language features, etc.

The user can apply each service independently and modify the settings. However, we recommend reading the instruction for a tool by clicking "?" on the page of the service. The description is intended to provide the user with the task that the service solves, as well as to suggest how the service can be applied for his purposes (Figure 3). As a rule, each instruction contains basic terms and concepts connected with a tool, its features, practical value, description of the user interface, scenarios for working with it. Additional comments and reviews are collected through the platform's contacts to more objectively compile technical tasks for its modernization.

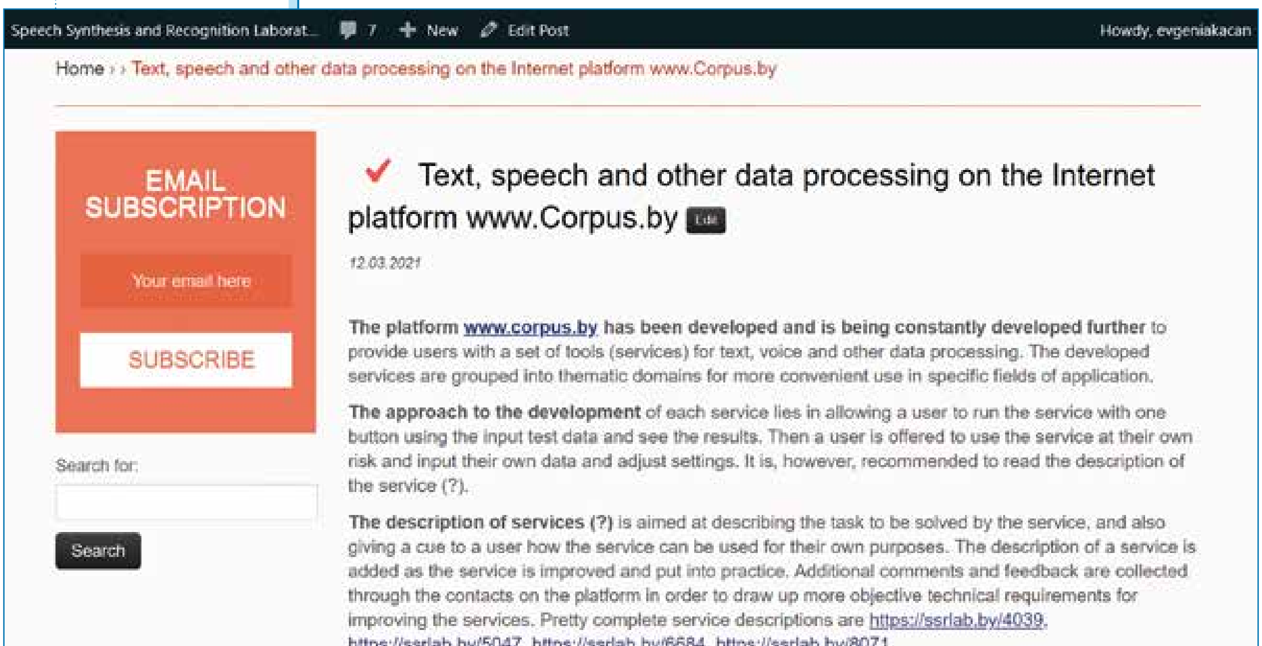


Figure 3. Description of the Internet platform on the laboratory's website https://ssrlab.by/

All tools provided in the platform meet the following requirements:

1. Simplicity, convenience and intuitive interface;
2. Saving all the data provided for input;
3. Saving all output data;
4. Auto-notification about errors in the operation of services.

All mentioned points above can be illustrated on the example of "***Dialectological maps***" service of the thematic domain "Linguist" [3]. This tool in an interactive format offers the user information about the dialectological pronunciation of certain words in various localities of Belarus. It is a software product that represents a set of interactive linguistic maps on Belarusian dialect phonetics. The software prototype was implemented in the programming languages *php* and *javascript*. When developing the prototype, the API of the Google Maps platform was used. In the context of a significant increase in the role of online resources in everyday life, as well as in the educational and scientific processes, the final software product was developed in the form of an Internet service. This makes it accessible to a wide range of people interested in this scientific field, and also expands the possibilities of using the results of research in the educational process. The product may arouse interest among linguists, teachers and students of higher and secondary educational institutions, as well as among programmers and technical specialists when creating new software products for information systems related to machine processing of linguistic material.

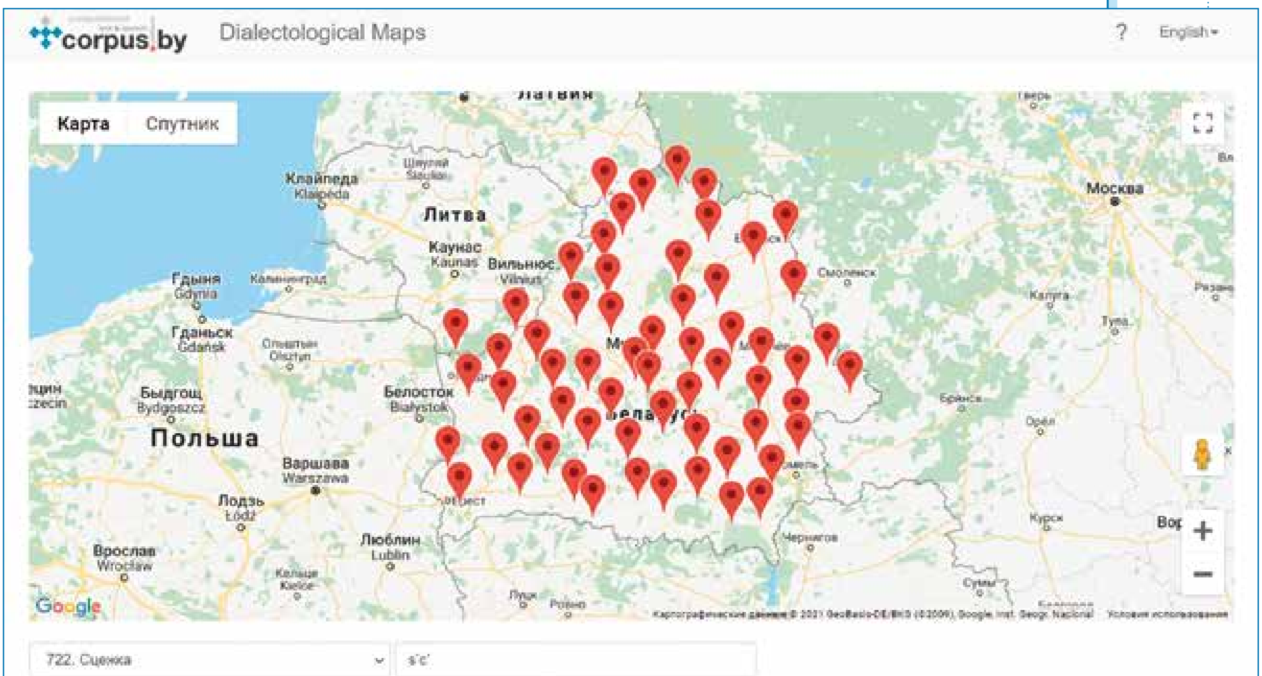


Figure 4. The interface of "Dialectological maps" service

The main screen of the service is a Google map, on which special markers are applied to the settlements of the Republic of Belarus, where the research was conducted (figure 4). The user is given the opportunity to choose from the list, which is located on the left under the map, one of the proposed words. After selecting a word, the map is automatically reloaded, and the corresponding data for the query becomes available. To get information about the pronunciation of the selected word in a particular point of Belarus, the user must click on the marker on the map corresponding to the point of interest. The following information will be displayed in the pop-up window:

- the number of the locality within this study;
- the name of the locality with a link to information about it;
- the locality relative to the administrative-territorial division of the country;
- pronunciation variant (s) of a selected word in the form of a transcription (figure 5).

Another functional feature of the prototype is the search for individual sounds or their combinations in all the studied points of Belarus by one selected word. This functionality is located just below the map. The result is displayed in a special form. For example, by selecting the word "Sciezka" from the list and entering the combination of sounds "žk" in the search field, the user will receive an answer in the following form represented in figure 6:

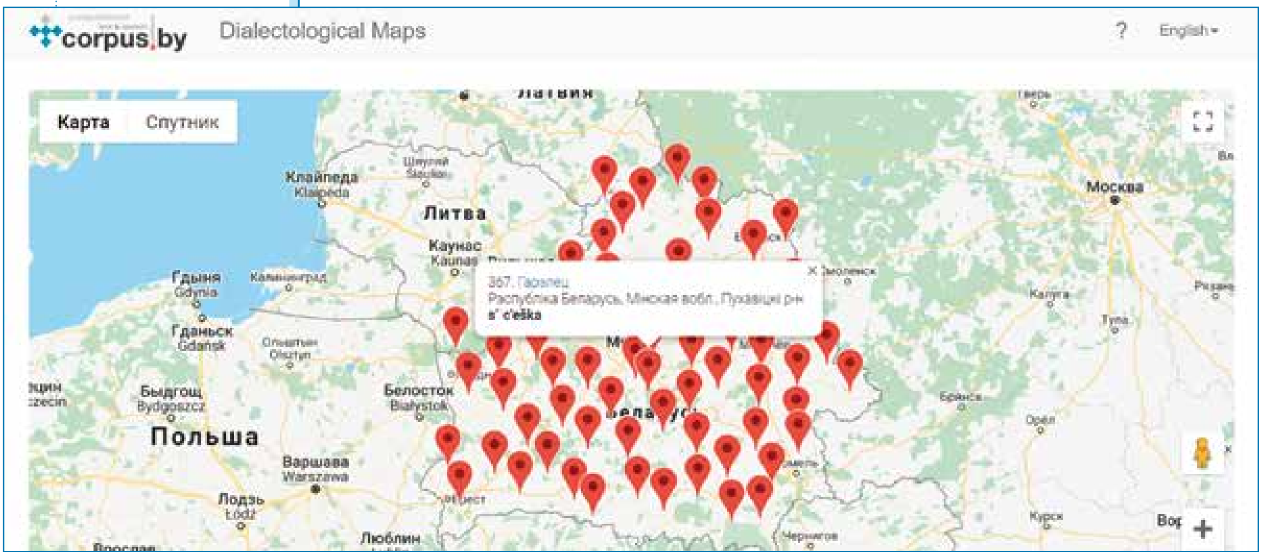


Figure 5. The searching results of the word "Sciezka" in the locality of Hareliec

Another service of "***Other***" domain is "***Thematic speech recognition***". It allows the user to convert speech into electronic text online [4]. A phonogram of thematic words with a size of no more than 20 MB is given at the input. The service gives the recognized electronic text of the phonogram at the output. It can be selected from the given examples, uploaded to the service from the computer's hard disk in the format .wav, and can also be recorded via the audio recording capabilities of the service.

At the moment, the service has a demo version and recognizes the Belarusian speech of the following thematic domains: clothes, cities, numbers, spontaneous speech. The list of domains will be updated. It works according to the instructions for creating programs based on CMU Sphinx.

The graphical interface of the service is shown in Figure 7. It has the following two areas: the audio file input area (on the left) and the output area of the recognized electronic text (on the right). The user can get acquainted with speech recognition thanks to the built-in examples, download a file for speech recognition from the hard disk or record his own audio.

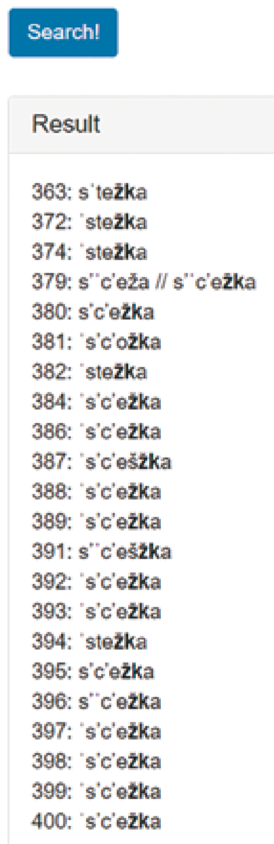


Figure 6. The results of searching "žk" combination in the word "Sciezka"

Speech recognition has great scientific prospects and wide application possibilities in many human-machine systems that are built on the basis of speech communication. In particular, the recognition of Belarusian speech, which becomes possible with the help of this tool, will allow the full development of Belarusian technical sciences, including robotics. There are also other areas of activity that require Belarusian speech recognition. For example, journalism, shorthand, and many others.

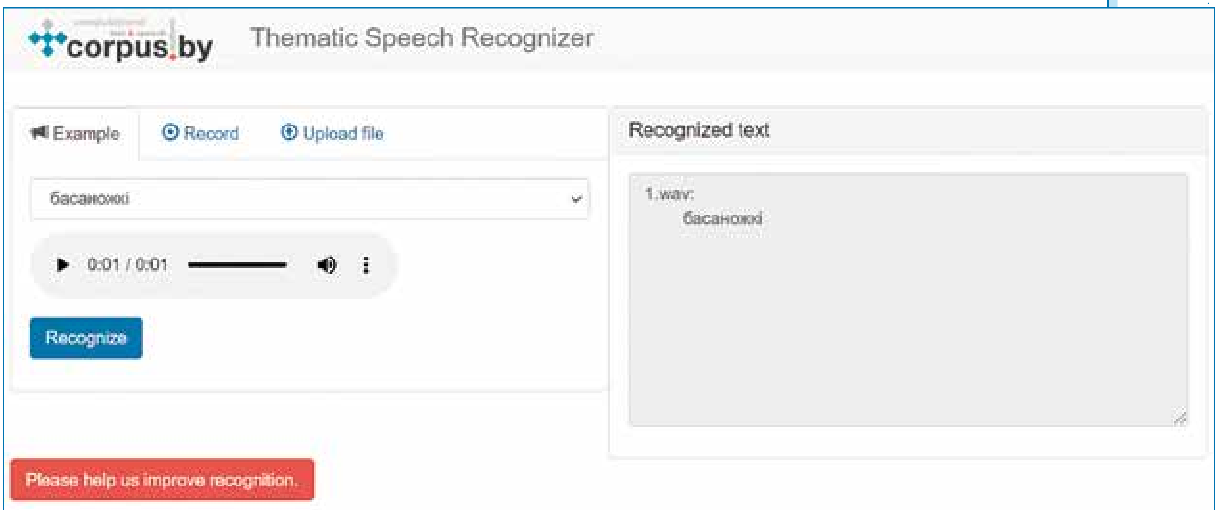


Figure 7. The interface of the service "Thematic speech recognition"

While using these services an individual can estimate simple and intuitive interface as well as the ability to choose the most convenient way of searching necessary information after reading a detailed instruction on the services. The same approach is implemented in all tools of *corpus.by*.

The platform has a technology stack with PHP, MySQL, Python, JavaScript, HTML, CSS. More than 90% of the services are programmed in PHP, the process of joining the Python language has begun with the development and deployment of a *Demonstrative service* in this language. Further transfer of all services to Python is underway. To speed up this process, the search for an additional developer is carried out. Other programming languages can be used to implement services through direct contact with developers.

To improve the quality of *corpus.by*, the following steps are planned:

- to develop new thematic domains and services for them;
- to create user accounts so that users can save the results of their experiments and share them with others;
- to form a rating system for the services;
- to produce a system for statistics collection to improve the performance of the most popular services;
- to expand the platform development team;
- to write (expand) the descriptions of the services that have been put into practice, tested, upgraded; translate fixed descriptions into English and other languages;
- to develop new services for processing of new electronic resources for different languages, thematic domains and tasks;

- to continue using the following approaches in the development of services: "everything has already been installed", "ready to use", "1 click on start-1 instant result", "everything has been saved";
- to create versions of the platform and services for Android, iPhone, Promobot v4.

Feedback from the target audience, experts, and regular users helps developers identify service errors, find solutions to implement them, and use new tools and ways to improve the effectiveness of the work according to single-user requests and comments. Our main 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 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 customer's demands.

Conclusion. The article describes the Internet platform for processing text and audio information *corpus.by* which is a useful and effective tool for automatic text processing in Belarusian, Russian and English. Individually, each service makes it possible to solve a specific computer-linguistic problem, and together they allow users to get a high-quality result of processing electronic text and speech. The information technologies, the creation of electronic dictionaries and new programs for processing languages, in particular Belarusian, is an urgent task today and will not lose its relevance due to the constant expansion of the role of computer technologies in human life. The platform *corpus.by* is publicly available and free to use. In addition, it is constantly developed to present the user with a set of tools for processing text, speech, and other data.

We are intended on creating and maintaining an 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. We try to widen the access to Belarusian developments in computational linguistics and popularize our tools within the Republic of Belarus and abroad. It is very important to support available tools and promote them out to improve and facilitate access for researches in humanities and social sciences that contribute to wide-ranging user support, guidelines and instructions for each service. In near future, we are planning to create and maintain new tools for electronic text and speech processing in the Belarusian language as well as popularize them among the users who are interested in computational linguistics..

References

1. A platform for processing text and audio information for various thematic domains of Corpus.by [Electronic resource]. — 2020. — access mode: http: //www.Corpus.by/. — Access date: 02.03.2020.

2. Laboratory of speech recognition and synthesis [Electronic resource]. — 2020. access mode: http://ssrlab.by/. — Access date: 11.07.2020.
3. Dialectological Maps // Computational Platform for Electronic text & Speech Processing Corpus.by [Electronic resource]. — 2020. Mode of access : https://corpus.by/DialectologicalMaps/?lang=en. — Date of access : 23.05.2020.
4. Thematic Speech Recognizer // Computational Platform for Electronic text & Speech Processing Corpus.by [Electronic resource]. — 2020. Mode of access : https://corpus.by/ThematicSpeechRecognizer/?lang=en. — Date of access : 31.08.2020.
5. *Getsevich, Yu. S.* Computer-linguistic services www.corpus.by for automatic text processing / E. S. Kachan, S. I. Lysy, Yu. S. Getsevich, G. R. Stanislavenko, A.V. Gunter / / national-cultural component in literary and dialect language: collection of sciences. art. / Brest State University named after A. S. Pushkin; redkal.: S. F. But-Gusaim [et al.]. — Brest: BrSU, 2016. — pp. 93–104.
6. *Stanislavenko, G. R.* editing electronic arrays of texts in the Belarusian language using the computer-linguistic services of the platform www.corpus.by / G. R. Stanislavenko, S. I. Lysy, Yu. S. Getsevich / / Karpovsky scientific readings / BSU; edited by I. P. A. I. Golovnya [et al.]. — Minsk: IVC of the Ministry of Finance, 2016. — pp. 262–267.
7. *Дзенісюк, Д. А.* Платформа для апрацоўкі тэкставай і гукавой інфармацыі для розных тэматычных даменаў беларускай мовы / Д. А. Дзенісюк, Я. С. Зяноўка, А. Е. Драгун [і інш.] // Языковая личность и эффективная коммуникация в современном поликультурном мире : материалы VI Междунар. науч.-практ. конф., посвящ. 100-летию Белорус. гос. ун-та, Минск, 29–30 окт. 2020 г. / Белорус. гос. ун-т ; редкол.: С. В. Воробьева (гл. ред.) [и др.]. — Минск : БГУ, 2020. — С. 69–74..

COMPUTATIONAL PLATFORM FOR ELECTRONIC TEXT AND SPEECH PROCESSING IN BELARUSIAN, RUSSIAN AND ENGLISH

Hetsevich Ju.S., *United Institute of Informatics Problems of NASB, Minsk, the Republic of Belarus, yuras.hetsevich@gmail.com*

Zianouka Ja.S., *United Institute of Informatics Problems of NASB, Minsk, the Republic of Belarus, evgeniakacan@gmail.com*

Mayeuski S.S., *United Institute of Informatics Problems of NASB, Minsk, the Republic of Belarus, maevskiiss@gmail.com*

Dzienisiuk D.A., *United Institute of Informatics Problems of NASB, Minsk, the Republic of Belarus, d.denissyuk@gmail.com*

Drahun A.Ja., *United Institute of Informatics Problems of NASB, Minsk, the Republic of Belarus, ndrahun@gmail.com*

This article describes an online platform corpus.by for automatic processing text and speech information, which offers a set of services for working with electronic texts and audio files. The purpose of the platform, its structural parts, and a number of tasks that are presented to developers of an online application for its improvement and optimization are considered in detail.

• computer technology • the platform for processing text and speech information • automatic processing • Internet service • text-to-speech synthesizer.