9. Авдонина М. Ю., Жабо Н. И. Эмоциональная составляющая как предмет обучения при подготовке переводчиков в сфере профессиональной коммуникации // Вестник УДН. Серия: Экология и безопасность жизнедеятельности. 2013. № 4. С. 129–138.

ПЛАТФОРМА ДЛЯ АПРАЦОЎКІ ТЭКСТАВАЙ І ГУКАВОЙ ІНФАРМАЦЫІ ДЛЯ РОЗНЫХ ТЭМАТЫЧНЫХ ДАМЕНАЎ БЕЛАРУСКАЙ МОВЫ

Д. А. Дзенісюк, Я. С. Зяноўка, А. Е. Драгун, С. С. Маеўскі, А. А. Бакуновіч, Ю. С. Гецэвіч
АІПІ НАН Беларусі
Мінск, Беларусь
e-mails: d.denissyuk@gmail.com; evgeniakacan@gmail.com; ndrahun@gmail.com; maevskiiss@gmail.com; bakunovich.andrei@gmail.com; mix1122@gmail.com

Дадзены артыкул апісвае інтэрнэт-платформу па апрацоўцы тэкставай і гукавой інфармацыі corpus.by, якая прапаноўвае набор сэрвісаў па аўтаматычнай апрацоўцы электронных і гукавых тэкстаў. Падрабязна разгледжаныя мэтанакіраванасць выкарыстання платформы, яе структурныя часткі і прадстаўлены шэраг задач, якія паўстаюць перад распрацоўшчыкамі анлайн-прыкладання для яго паляпшэння і аптымізацыі.

Ключавыя словы: камп'ютарныя тэхналогіі; платформа для апрацоўкі тэкставай і гукавой інфармацыі; аўтаматычная апрацоўка; інтэрнэт-сэрвіс; сінтэзатар маўлення па тэксце.

THE PLATFORM FOR TEXT AND SPEECH PROCESSING FOR DIFFERENT THEMATIC DOMAINS OF THE BELARUSIAN LANGUAGE

D. A. Dzienisiuk, Ja. S. Zianoŭka, A. Je. Drahun, S. S. Majeŭski, A. A. Bakunovič, Ju. S. Hiecevič
UIIP of NASB
Minsk, Belarus
e-mails: d.denissyuk@gmail.com; evgeniakacan@gmail.com; ndrahun@gmail.com; maevskiiss@gmail.com; bakunovich.andrei@gmail.com; mix1122@gmail.com

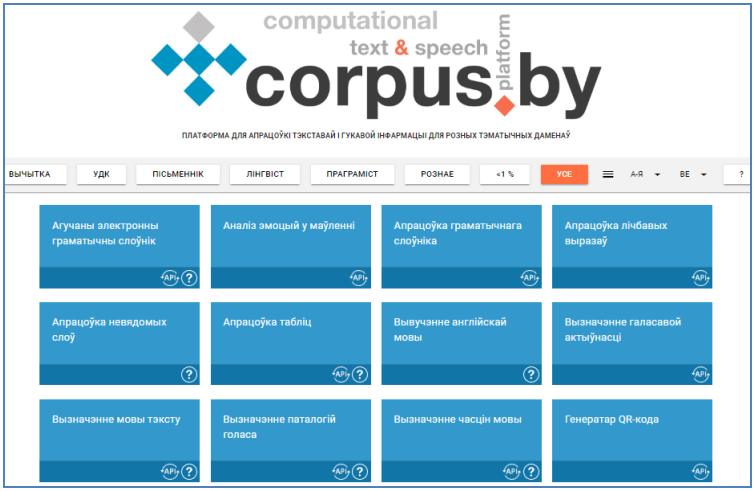
This article describes an online platform corpus.by for processing text and speech information, which offers a set of services for automatic processing of electronic texts and audio files. The purpose of the platform, its structural parts, and a number of tasks that are presented to developers of an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

Key words: computer technology; the platform for processing text and speech information; automatic processing; Internet service; text-to-speech synthesizer.

За мінулыя паўстагоддзя ў галіне камп'ютарнай лінгвістыкі былі атрыманы значныя навуковыя і практычныя вынікі. Адным з асноўных пытанняў, якое зараз паўстае перад навукоўцамі, з'яўляецца праблема аўтаматызаванай апрацоўкі тэксту, якая атрымала асаблівую актуальнасць. Высокая хуткасць росту колькасці даступнай інфармацыі змушае ўдасканальваць спосабы яе апрацоўкі, рэалізоўваць частковую ці поўную аўтаматызацыю гэтых працэсаў. Лёгка заўважыць, што асноўным і найбольш запатрабаваным спосабам прадстаўлення інфармацыі з'яўляецца тэкст на натуральнай мове. Таму адным з важных накірункаў прымянення камп'ютарных тэхналогій з'яўляецца распрацоўка сістэм, здольных аўтаматычна апрацоўваць электронныя тэксты. Супрацоўнікамі лабараторыі распазнавання і сінтэзу маўлення Аб'яднанага інстытута праблем інфарматыкі Нацыянальнай акадэміі навук Беларусі [1] быў распрацаваны інтэрнэт-рэсурс www.corpus.by [2], які дапамагае вырашаць мноства задач, звязаных з апрацоўкай электронных тэкстаў і маўленчых сігналаў. На працягу апошніх 60 гадоў асноўным накірункам дзейнасці лабараторыі з'яўляецца распрацоўка сістэм сінтэзу маўлення. Таму большая частка распрацаваных рэсурсаў убудаваныя ў БСМТ і ахопліваюць шэраг задач, якія павінны быць вырашаны сінтэзатарам.

Платформа corpus.by уяўляе сабой набор розных сэрвісаў, якія накіраваны на мэтавую аўдыторыю (праграмісты, лінгвісты, філолагі, студэнты, выкладчыкі і г.д.). Сэрвісы забяспечваюць просты і ўстойлівы доступ да сродкаў і інструментаў апрацоўкі электроннага тэксту для аналізу, выяўлення, даследавання або аб'яднання такіх набораў даных на беларускай, рускай і англійскай мовах. Прынцып функцыянавання corpus.by заключаецца ў адпаведнасці "уваходныя даныя – выніковыя даныя": карыстальнік уводзіць тэкставую інфармацыю і на выхадзе мае апрацаваныя вынікі. Платформа прадстаўляе інструменты для такенізацыі, марфалагічнага аналізу, агучвання электроннага граматычнага слоўніка, пошуку амонімаў, падліку частотнасці сімвалаў і слоў, праверкі арфаграфіі, сінтэзатар маўлення, сэрвісы па распазнаванні маўлення і эмоцый і многае іншае. Агульная колькасць праграмных рэалізацый, прадстаўленых у платформе – 67 (малюнак 1).

Кожны з сэрвісаў вырашае свае ўласныя задачы, у той жа час удасканальваючы функцыянальнасць працы беларускамоўнага сінтэзатара маўлення па тэксце, які таксама ўваходзіць у спіс сэрвісаў. Падыход да распрацоўкі палягае ў тым, каб карыстальнік мог увесці тэставыя даныя, запусціць сэрвіс націсканнем адной кнопкі і азнаёміцца з вынікамі працы.



Малюнак 1. Інтэрфейс платформы corpus.by

Далей карыстальніку прапаноўваецца самастойна выкарыстаць сэрвіс з уведзенымі ўласнымі данымі і выстаўленымі ўласнымі ж настойкамі. Аднак папярэдне рэкамендуецца азнаёміцца з апісаннем сэрвіса, націснуўшы на «?». Апісанне мае на мэце прадставіць карыстальніку задачу, якую вырашае сэрвіс, а таксама падказаць, як можна выкарыстоўваць сэрвіс для сваіх мэт (малюнак 2).

Апрацоўка тэксту, маўлення і іншых дадзеных на Інтэрнэт-платформе www.Corpus.by

11.06.2020

Платформа www.corpus.by распрацавана і пастаянна дадаткова распрацоўваецца, каб прадставіць карыстальніку набор інструментальных сродкаў (сэрвісаў) па апрацоўцы тэксту, маўлення і іншых дадзеных. Распрацаваныя сэрвісы групуюцца ў тэматычныя дамены для больш зручнага выкарыстання ў канкрэтных практычных сферах.

Падыход да распрацоўкі кожнага сэрвіса палягае ў тым, каб карыстальнік мог па ўведзеных тэставых дадзеных запусціць сэрвіс адной кнопкай і азнаёміцца з яго вынікамі працы. Далей карыстальніку прапануецца самастойна выкарыстоўваць сэрвіс на свой страх і рызыку з уведзенымі ўласнымі дадзенымі і выстаўленымі ўласнымі ж настройкамі. Аднак папярэдне рэкамендуецца азнаёміцца з апісаннем сэрвіса (?).

Малюнак 2. Апісанне інтэрнэт-платформы на сайце лабараторыі https://ssrlab.by/

Апісанне сэрвісаў з'яўляюцца па меры іх дапрацоўкі і ўкаранення ў практыку. Дадатковыя каментары і водгукі збіраюцца праз кантакты платформы, каб больш аб'ектыўна складаць тэхнічныя заданні па аптымізацыі сэрвіса для яго мадэрнізацыі [3].

Выстаўленыя сэрвісы адпавядаюць наступным патрабаванням:

1) прастата, зручнасць і інтуітыўная зразумеласць інтэрфейсу;

2) захаванне ўсіх даных, што пададзены на ўваход;

3) захаванне ўсіх выніковых даных;

4) аўтаапавяшчэнне распакоўшчыкаў аб памылках у працы сэрвісаў.

Распрацаваныя сэрвісы групіруюцца ў тэматычныя дамены для больш зручнага выкарыстання ў канкрэтных практычных сферах (*Вычытка, УДК, Пісьменнік, Лінгвіст, Праграміст, Рознае)* для задавальнення патрэб мэтавых карыстальнікаў. Так, напрыклад, тэматычны дамен “Лінгвіст” прапаноўвае сэрвісы па апрацоўцы тэкставай і гукавой інфармацыі, фанетычных з’яў мовы, інструментальныя сродкі вызначэння, аналізу і пошуку разнастайных моўных прыкмет і г.д. З усімі тэматычнымі даменамі можна пазнаёміцца на афіцыйнай старонцы платформы *https://www.corpus.by/?lang=be* (малюнак 3).

Малюнак 3. Спіс сэрвісаў, карысных для лінгвіста

Распрацоўка новых сэрвісаў адбываецца па жаданні ці па замове праз выкарыстанне сэрвіса-шаблона “Дэманстрацыйны сэрвіс”, які можна вольна спампаваць (даступныя варыянты для розных моў праграмавання). Распакоўшчык дае назву новаму сэрвісу X і карэктуе яго пад выкананне пастаўленай задачы. Калі вырашэнне пастаўленай задачы адбылося паспяхова і сэрвіс X апрацоўвае ўваходныя даныя з дакладнасцю больш чым 1%, то яго можна апублікаваць на платформе праз кантакт з распакоўшчыкамі платформы [4].

Тэхналагічны стэк, які выкарыстоўвае платформа: PHP, MySql, Python, JavaScript, HTML, CSS. Больш за 90% сэрвісаў запраграмаваны на мове PHP, працэс далучэння мовы Python пачаты з распрацоўкі і выстаўлення “Дэманстрацыйнага сэрвіса” на гэтай мове. Далейшы перавод усіх сэрвісаў на Python ідзе, для паскарэння дадзенага працэсу ажыццяўляецца пошук дадатковага распакоўшчыка. Іншыя мовы праграмавання могуць быць выкарыстаны для распрацоўкі сэрвісаў праз непасрэдны кантакт з распакоўшчыкамі Для паляпшэння якасці апісанай платформы плануюцца наступныя крокі:

- фарміраванне новых тэматычных даменаў і сэрвісаў для іх;
- стварэнне карыстальніцкіх акаўнтаў для магчымасці захоўваць вынікі сваіх эксперыментаў і дзяліцца імі з іншымі;
- распрацоўка сістэмы выстаўлення рэйтынгаў сэрвісаў;
- распрацоўка сістэмы збору статыстыкі выкарыстання сэрвісаў для паляпшэння працы найбольш папулярных сэрвісаў;
- пашырэнне каманды распрацоўшчыкаў платформы;
- напісанне (пашырэнне) апісанняў сэрвісаў, якія прайшлі ўкараненне, апрабацыю, мадэрнізацыю; пераклад устойлівых апісанняў на англійскую і іншыя мовы;
- распрацоўка новых сэрвісаў для апрацоўкі новых электронных рэсурсаў для розных моў, тэматычных даменаў і задач;
- распрацоўка версіі платформы і сэрвісаў для Android, IPhone, Promobot V.4.

Зваротная сувязь ад мэтавай аўдыторыі, экспертаў і звычайных карыстальнікаў дапамагае распрацоўшчыкам у выяўленні памылак працы сэрвісаў, пошуку рашэнняў па іх дапрацоўцы і прымяненні новых сродкаў і спосабаў павышэння эффектыўнасці працы згодна з карыстальніцкімі запытамі і заўвагамі.

Разгледжаная ў артыкуле інтэрнэт-платформа для апрацоўкі тэкставай і гукавой інфармацыі corpus.by з‘яўляецца карысным і эфектыўным сродкам аўтаматычнай апрацоўкі тэкстаў на беларускай, рускай і англійскай мовах. Паасобку кожны сэрвіс дае магчымасць вырашыць пэўную камп’ютарна-лінгвістычную задачу, а ў сукупнасці яны дазваляюць атрымаць якасны вынік апрацоўкі электроннага тэксту. Укараненне беларускай мовы ў інфармацыйныя тэхналогіі, стварэнне электронных слоўнікаў і новых праграм для апрацоўкі менавіта беларускай мовы на сённяшні дзень з’яўляецца актуальнай задачай і не страціць сваёй актуальнасці дзякуючы пастаяннаму пашырэнню ролі камп’ютарных тэхналогій у жыцці чалавека. Платформа www.corpus.by знаходзіцца ў адкрытым доступе і бясплатная для выкарыстання. Акрамя

таго, яна пастаянна дадаткова распрацоўваецца, каб прадставіць карыстальніку набор інструментальных сродкаў (сэрвісаў) па апрацоўцы тэксту, маўлення і іншых даных.

Бібліяграфічныя спасылкі

1. Лабараторыя распазнавання і сінтэзу маўлення [Электронны рэсурс]. Рэжым доступу: http://ssrlab.by/ (дата звароту: 11.07.2019).
2. Платформа для апрацоўкі тэкставай і гукавой інфармацыі для розных тэматычных даменаў Corpus.by [Электронны рэсурс]. Рэжым доступу: http://www.corpus.by/ (дата звароту: 02.03.2020).
3. Камп'ютарна-лінгвістычныя сэрвісы www.corpus.by для аўтаматычнай апрацоўкі тэкстаў / Ю. С. Гецэвіч [і інш.] // Нацыянальна-культурны кампанент у літаратурнай і дыялектнай мове : зб. навук. арт. Брэст : БрДУ, 2016. С. 93–104.
4. Станіславенка Г. Р., Лысы С. І., Гецэвіч Ю. С. Рэдагаванне электронных масіваў тэкстаў на беларускай мове з выкарыстаннем камп'ютарна-лінгвістычных сэрвісаў платформы www.corpus.by // Карповские научные чтения / БГУ ; под ред. А. И. Головня [и др.]. Минск : ИВЦ Минфина, 2016. С. 262–267.

О СЕМАНТИКЕ ДЕЙКТИЧЕСКИХ НАРЕЧИЙ С ОБЩЕЙ ПРЕДЛОЖНОЙ БАЗОЙ (НА МАТЕРИАЛЕ НЕМЕЦКОГО ЯЗЫКА)

Т. С. Котик
Минский государственный лингвистический университет
Минск, Беларусь
e-mail: tanjakotik@gmail.com

В данной статье рассматриваются дейктические наречия, образованные с помощью дейктических формантов *hin-\her- по схеме hin-\her-* + предлог/наречие. Всего в языке зафиксировано 10 дейктических наречий с формантом *hin-* и 12 дейктических наречий с формантом *her-*. 8 пар производных наречий имеют общую предложную базу.

Ключевые слова: немецкий язык; дейксис; локальный дейксис; дейктические наречия; словообразование; дейктические форманты.

SEMANTICS OF DEACTIC ADVERBS WITH THE SAME PREPOSITIONAL BASIS (BASED ON THE MATERIAL OF THE GERMAN LANGUAGE)

T. S. Kotik
Minsk State Linguistic University
Minsk, Belarus
e-mail: tanjakotik@gmail.com

The article considers deictic adverbs formed with the *hin- / her-* deictic formants according to *the hin- / her- + preposition / adverb* scheme. In total, the language has 10