УДК 004.934.5:811.161.3

Андрэй Бакуновіч, Яўгенія Зяноўка,
Анастасія Драгун, Юрый Гецэвіч

Аб'яднаны інстытут праблем інфарматыкі НАН Беларусі,
г. Мінск, Беларусь

КОМПЛЕКС КАМП'ЮТАРНЫХ СРОДКАЎ ДЛЯ АПРАЦОЎКІ ФАНЕТЫЧНЫХ З'ЯЎ БЕЛАРУСКАЙ ЛІТАРАТУРНАЙ МОВЫ

В данной статье рассматривается проблема использования современных компьютерных технологий в лингвистических исследованиях, в частности, в прикладной фонетике белорусского литературного языка. Описана платформа обработки текстовой и звуковой информации для различных тематических доменов, представляющая собой набор инструментов для решения лингвистических задач. Подробно рассмотрены сервисы платформы обработки фонетических явлений белорусского языка.

Ключевые слова: *компьютерные технологии, автоматическая обработка, фонетические явления, прикладные исследования.*

Авалоданне камп'ютарнымі тэхналогіямі ў сучасным інфармацыйным асяроддзі з'яўляецца неад'емнай часткай развіцця індывідыума. Свабодная арыентацыя ў велізарнай інфармацыйнай прасторы, наяўнасць неабходных ведаў і навыкаў, у тым ліку пошуку, апрацоўкі і захоўвання інфармацыі з выкарыстаннем сучасных інфармацыйных тэхналогій, камп'ютэрных сістэм і сетак характарызуе яго як адукаваную асобу. Складана ўявіць сабе асобную сферу дзейнасці, і асабліва навуку, у якой не прымяняюцца разнастайныя камп'ютарныя і мабільныя прыкладанні, праграмнае забеспячэнне ці сродкі і сістэмы аўтаматычнай апрацоўкі інфармацыі. Прыкладныя даследаванні за-

кранаюць і фаналагічную сістэму беларускай мовы, у працэсе якіх адбываецца апрацоўка фанетычных з'яў камп'ютарнымі сродкамі, адсочваюцца змены ў функцыянаванні мовы і замацоўваюцца новыя фіксаваныя тэндэнцыі ў навуковых крыніцах [9].

Адным з найбольш актуальных напрамкаў прымянення камп'ютарных тэхналогій з'яўляецца аўтаматызаваная апрацоўка тэкстаў вялікіх памераў, якая атрымала асаблівую актуальнасць. Скарачэнне выдаткаванага часу і чалавечых рэсурсаў садзейнічае ўдасканальванню спосабаў апрацоўкі, пошуку і распрацоўцы метадаў і алгарытмаў для вырашэння прыкладных задач асабліва для беларускай мовы. Супрацоўнікі лабараторыі распазнавання і сінтэзу маўлення АІПІ НАН Беларусі [2] стварылі інтэрнэт-платформу для апрацоўкі тэкставай і гукавой інфармацыі для розных тэматычных даменаў corpus.by [6]. Платформа ўяўляе сабой набор розных сэрвісаў (даклад-ная колькасць – 74) для праграмістаў, лінгвістаў, філолагаў, студэнтаў, выкладчыкаў і г. д. Просты і перманентны доступ да сродкаў і інструментаў апрацоўкі электроннага тэксту забяспечвае рэалізацыю такіх функцый, як аналіз, даследаванне або аб'яднанне набораў даных на беларускай, рускай і англійскай мовах. Прынцып функцыянавання corpus.by заключаецца ў адпаведнасці «ўваходныя даныя – выніковыя даныя», дзе карыстальнік уводзіць тэкставую інфармацыю і на выхадзе атрымлівае апрацаваныя вынікі. Падыход да распрацоўкі сэрвісаў заключаецца ў тым, каб карыстальнік мог па ўведзеных тэставых даных запусціць сэрвіс адной кнопкай і азнаёміцца з вынікамі яго працы. Далей карыстальніку прапаноўваецца самастойна выкарыстаць сэрвіс з уведзенымі ўласнымі данымі і выстаўленымі ўласнымі настройкамі.

На платформе прадстаўлены разнастайныя сэрвісы па апрацоўцы фанетыкі беларускай мовы. А менавіта *Генератар арфаэпічнага слоўніка, Генератар транскрыпцый, Графічнае адлюстраванне алафонаў і алафонных фраз, падлік частот-*

насці алафонаў, Ідэнтыфікатар амографаў, Спецыялізаваны фанетычны слоўнік, Дыялекталагічныя карты, Фанетычны мінімізатар, Пошук фанетычных з'яў, Падзел на склады, Лематызатар і інш. Так, напрыклад, сэрвіс «*Фанетычны мінімізатар*» дазваляе карыстальніку на аснове корпуса тэкстаў на беларускай мове сфарміраваць мінімізаванае мноства сказаў, якія пакрываюць усе фанетычныя адзінкі, наяўныя ў зыходным корпусе [11]. На ўваход сэрвісу падаецца ўведзены карыстальнікам база тэкстаў. Могуць быць вызначаны два параметры мінімізацыі: базавая адзінка, паводле якой адбываецца мінімізацыя (алафон, дыфон, трыфон, склад), і мяжа пошуку, да якой адбываецца пошук па кожнай унікальнай фанетычнай адзінцы. На выхадзе атрымліваюцца тры тэкставыя файлы: файл з мінімізаваным корпусам сказаў, файл са спісам унікальных фанетычных адзінак і файл са спісам рэдкіх фанетычных адзінак. Алгарытм «Фанетычнага мінімізатара» дае магчымасць сфарміраваць мінімізаваны корпус тэкстаў, якія пакрываюць усе прысутныя ў зыходным корпусе гукавыя адзінкі [4].

Для большага разумення разгледзім вынікі працы сэрвісу, якія адлюстраваны ніжэй у выглядзе арфаграфічнага тэксту і адпаведнага яму алафоннага тэксту. Такі тэкст уяўляе сабой паслядоўнасць абазначэнняў алафонаў, паўз, словападзелаў і складападзелаў.

Груша цвіла апошні год. Усе галіны яе, усе вялікія расохі, да апошняга пруціка, былі ўсыпаны бурным бела-ружовым цветам.

GH004,R022,U022,>,SH002,A323,/,>,C'002,V'002,I241,>,L002,A012,/,>,A221,>,P001,O012,>,SH002,N'004,I242,/,>,GH001,O032,T000,/,>,#P4,>,U203,>,S'001,E042,/,>,GH004,A233,>,L'002,I042,>,N004,Y323,/,>,J'012,A243,>,J'011,E040,/,>,#C3,>,U203,>,S'001,E043,/,>,V'012,A243,>,L'002,I043,>,K'002,I343,>,J'012,A342,/,>,R002,A222,>,S001,O023,>,H'002,I340,/,>,#C3,>,D004,A322,>,A221,>,P001,O012,>,SH002,N'004,A342,

>,GH004,A231,/,>,P002,R012,U023,>,C'002,I342,>,K004,A330,/,>,#C3,>,B002,Y013,>,L'004,I241,/,>,W013,S001,Y021,>,P002,A312,>,N004,Y221,/,>,B002,U012,R001,>,N004,Y221,M001,/,>,B'002,E141,>,L004,A312,>,R002,U222,>,ZH002,O021,>,V012,Y211,M003,/,>,C'002,V'001,E042,>,T002,A321,M000,/,>,#P4

Коды алафонаў у гэтым запісе складаюцца з літарнай назвы фанемы (напрыклад, GH), знака мяккасці «'» (пры яе наяўнасці) і лічбавага кода, які ўказвае на тыя ці іншыя асаблівасці фанемы. У шэрагу эксперыментаў могуць выкарыстоўвацца не толькі поўныя (напрыклад, ZH002), але і скарочаныя (напрыклад, ZH0) запісы алафонаў, у якіх адкідаюцца дзве апошнія лічбы, што ўказваюць на кантэкст алафона ў слове. Таксама ў дадзеным запісе можна назіраць знакі словападзелу «/» і складападзелу «>». Акрамя прадстаўленага фрагмента мінімізаванага корпуса тэксту, карыстальнік атрымлівае спіс унікальных і рэдкіх фанетычных адзінак.

Сэрвіс можа быць прыменены пры распрацоўцы сістэмы сінтэзу беларускага маўлення, заснаванай на моўнай мадэлі. Значна зменшаны аб'ём корпуса робіць стварэнне такіх сістэм даступным шырокаму колу распрацоўшчыкаў і даследчыкаў. Акрамя таго, аўтаматызацыя адбору мінімізаванага фанетычна поўнага мноства тэкстаў на беларускай мове актуальна ў шэрагу разнастайных навуковых сфер, напрыклад, у лінгвістычных даследаваннях ці пры стварэнні адмысловых дапаможнікаў па вывучэнні беларускай фанетыкі [4].

Сэрвіс «Падзел на склады» выдае апрацаваны тэкст у алафонным выглядзе з падзелам на склады [5]. На ўваход праграме падаецца адвольны тэкст на беларускай мове. Пасля яго апрацоўкі сістэмай карыстальнік атрымлівае ўваходны тэкст у алафонным выглядзе, а таксама тэкст у алафонным запісе з падзелам на склады. Мэтай дадзенага сэрвісу з'яўляецца аўтаматызаваны падзел на склады слоў, што можа спатрэбіцца ў працы мовазнаўцаў, а таксама тых, хто вывучае асаблівасці мовы, вымаўлення.

Адным з даволі значных для фанетыстаў сэрвісаў з'яўляецца *«Пошук фанетычных з'яў»*. Ён прызначаны для ідэнтыфікацыі той ці іншай фанетычнай з'явы ва ўведзеным тэксце. Прынцып функцыянавання сэрвісу падобны да «Фанетычнага мінімізатара». На ўваход падаецца адвольны тэкст. Карыстальніку прапаноўваецца вызначыць такія фанетычныя пары, як:

пары «свісцячы – свісцячы», «свісцячы – шыпячы», «шыпячы – свісцячы», «шыпячы – шыпячы» на сутыку двух слоў;

пары «шумны – шумны» на сутыку двух слоў;

пары «санорны – санорны» на сутыку двух слоў;

пары «зычны – зычны» на сутыку двух слоў;

пары «ётавы галосны ці «й» – ётавы галосны ці «й» на сутыку двух слоў;

пары аднолькавых зычных гукаў на сутыку двух слоў.

Можна выбраць увесь спіс пералічаных з'яў ці асобныя пары згодна з адзначанымі опцыямі. Пасля націскання кнопкі *«Пошук»* сэрвіс выдае колькасць знойдзеных пар з абазначэннем колеру для асобнай пары. Напрыклад, блакітным пазначана пара «санорны – санорны» на стыку слоў.

Так, у невялікім тэкставым фрагменце І. Мележа («Агні над руінамі») знойдзена 4 пары «зычны – зычны» на сутыку двух слоў; 2 пары «санорны – санорны» на сутыку двух слоў; 2 пары «шумны – шумны» на сутыку двух слоў; адна пара аднолькавых зычных гукаў на сутыку двух слоў:

Страшны малюнак убачылі салдаты. Вёска была так разбурана, што амаль немагчыма было вызначыць тое месца, дзе некалі стаялі хаты. У вёсцы не ўцалела ні адной пабудовы, ні аднаго дрэўца. Ніхто не выйшаў насустрач байцам, у вёсцы не было ні аднаго жыхара – гітлераўцы часткай пагналі іх у няволю, часткай знішчылі. Здавалася, вёска сцёрта з зямлі назаўсёды.

Практычная вартасць сэрвісу заключаецца ў магчымасці апрацоўкі тэкстаў на беларускай мове вялікага памеру і хуткага атрымання іх колькасных паказчыкаў з пералікам фанетычных з'яў. Гэта значна спрашчае працу фанетыстаў і простых карыстальнікаў, якія зацікаўлены ў атрыманні практычных вынікаў без звяртання да арфаэпічнага слоўніка беларускай літаратурнай мовы [1].

Сэрвіс «Лематызатар» прымяняецца для вызначэння пачатковых формаў слоў (лем) [3]. Лемы выкарыстоўваюцца ў слоўніках у якасці загаловачных слоў, пасля якіх могуць перайчвацца іншыя формы лексемы. У рускай і беларускай мовах першапачатковымі лічацца наступныя марфалагічныя формы, якія і выдае праграма:

для назоўнікаў – назоўны склон, адзіночны лік;

для прыметнікаў – назоўны склон, адзіночны лік, мужчынскі род;

для дзеясловаў, дзеепрыметнікаў, дзеепрыслоўяў – дзеяслоў у інфінітыве незакончанага трывання.

У камп'ютарнай лінгвістыцы лематызацыя часта вызначаецца як метад марфалагічнага аналізу, у працэсе якога ад лексемы павінны быць адкінуты ўсе флектыўныя элементы, якія не адпавядаюць пачатковай форме слова. Для атрымання дапаможных даных, у прыватнасці, для вызначэння стандартнай структуры пачатковай формы слоў пэўнай часціны мовы, сістэма лематызацыі можа выкарыстоўваць пошук па слоўніку. На дадзены момант метад пошуку па слоўніках з'яўляецца адзіным рашэннем, якое прымяняе сэрвіс. На сённяшні дзень сэрвісам апрацоўваюцца 7 слоўнікаў для беларускай мовы і адзін для рускай (на старонцы сэрвісу прадстаўлены пералік слоўнікаў). Тым не менш ступень паўнаты слоўнікаў і спецыфіка размяшчэння інфармацыі ў іх электронных версіях параджаюць некаторыя недакладнасці пры апрацоўцы тэксту. Пры падрабязным аналізе працы сэрвісу былі выяўлены наступныя праблемы:

Не лематызуюцца ўласныя імёны.

Не лематызуюцца многія запазычанні (*«рэлаксацыя»*, *«такенізацыя»*, *«шугарынг»*), асабліва ў выпадках, калі такія

запазычанні маюць прэфіксы або суфіксы. Тым не менш некаторыя запазычанні (напрыклад, *«ідэнтыфікацыя»*, *«дэпрывацыя»*) могуць быць лематызаваныя, калі ў іх адсутнічаюць падобныя марфемы.

Не лематызуецца большасць слоў з суплетыўнымі асновамі. Гэта тычыцца некаторых ступеней параўнання прыметнікаў *(«добры – лепшы»*, *«хороший – лучший»)*, некаторых назоўнікаў *(«чалавек – людзі»)*, дзеясловаў *(«класціся – легчы»)*. Займеннікі *(«мы – нас»*, *«я – мяне»)* у большасці выпадкаў (хаця і не ва ўсіх) будуць лематызаваныя.

Не лематызуюцца многія словы, якія маюць два і больш карані *(«чорна-зялёны»)*, а таксама лексемы з дадаткамі *(«чалавек-амфібія»)*, хаця частка падобных слоў усё ж можа быць лематызавана.

Словы з памяншальна-ласкальнымі і павелічальна-знявагальнымі суфіксамі *(«сильненький»*, *«городишко»)* у многіх выпадках таксама не будуць лематызаваныя. Асабліва гэта тычыцца рускай мовы. Для беларускай мовы частка падобных словаформаў *(«цёпленькі»*, *«гарадочак»)* будзе апрацоўвацца як самастойныя лексемы.

У многіх выпадках цяжкасці будуць выклікаць назвы маладых істот *(«качаня»*, *«жарабя»)*.

Сінтэтычныя формы ступеней параўнання *(«мацнейшы»*, *«найпрыгажэйшы»)* у агульным выпадку будуць прыводзіцца не да зыходнай формы прыметніка, а да формы назоўнага склону адзіночнага ліку мужчынскага роду, як у выпадку, калі б формы ступеней параўнання былі самастойнымі прыметнікамі. Для рускай мовы сінтэтычныя формы ступеней параўнання *(«сильнейший»*, *«наимощнейший»)* не будуць лематызаваныя.

Усе словаформы, утвораныя аналітычным шляхам *(«самы прыгожы»*, *«зрабіў бы»)*, разглядаюцца сэрвісам як набор асобных слоў і апрацоўваюцца адпаведным чынам.

Многія з прыведзеных вышэй праблем вырашаюцца шляхам укаранення марфалагічнага аналізу, які ідзе пасля слоўнікавага аналізу. Таму распрацоўка і ўкараненне правіл марфалагічнага аналізу з'яўляецца прыярытэтнай задачай развіцця сэрвісу.

Пры ўводзе тэксту для апрацоўкі сістэмай карыстальнік мае магчымасць выбраць асобныя опцыі. А менавіта ўвесці не-знаёмыя для сябе словы, выбраць фармат вынікаў працы сэрвісу (змяшчае такія варыянты, як *«Паказваць вынік у зыходным фармаце»*, *«Паказваць увесь вынік у адзін радок»*, *«Паказваць вынікі радкамі»*, *«Паказваць вынікі ў слупок»*), абраць слоўнікі, якія будуць задзейнічаны ў апрацоўцы. Пасля націскання кнопкі *«Паказаць спіс слоў з часцінамі мовы!»* адбываецца апрацоўка тэксту і выдаюцца вынікі. У акне *«Словы з часцінамі мовы»* выводзяцца словаформы ў выбраным карыстальнікам фармаце, іх лемы і назвы слоўнікаў, у якіх лемы былі знойдзены (калі карыстальнік актываваў адпаведную опцыю). А ў акне *«Невядомыя словы»* змяшчаюцца словы, якія сэрвісу не ўдалося апрацаваць.

Метад лематызацыі прымяняецца ў пошукавых алгарытмах у працэсе схематызацыі вэб-дакуметаў, а таксама пры іх індэксіраванні. Нягледзячы на высокі тэхналагічны ўзровень сучасных пошукавых сістэм, падобная апрацоўка не заўсёды бывае дакладнай, паколькі пошукавы робат часцей улічвае толькі адну з магчымых лем словаформы, прыведзенай у тэксце дакумента. Таму далейшае развіццё метадаў лематызацыі з'яўляецца прыярытэтнай тэхналагічнай задачай. Выкарыстанне лематызацыі значна паляпшае якасць аналізу сайтаў і дакументаў. Сістэмы лематызацыі рускіх тэкстаў распрацаваны на сённяшні дзень дастаткова добра, у той час як для беларускай мовы сітуацыя выглядае інакш. Многія працы беларускіх вучоных даступныя чытачам на беларускай мове. Правільная лематызацыя як асноўных тэкстаў, так і дапаможных даных (назвы артыкулаў, звесткі пра аўтараў, спісы літаратуры) можа быць паспяхова прыменена ў дзейнасці бібліятэк. Прыватнымі выпадкамі прымянення лематызацыі могуць быць крыміналістычная лінгвістычная экспертыза,

аналіз тэкстаў на прадмет плагіяту, аналіз мовы тэкстаў пісьменніка, аналіз электронных вучэбных тэкстаў і электронных тэкстаў, складзеных навучэнцамі, у сістэмах адаптыўнага навучання.

Такім чынам, прадстаўленая платформа corpus.by накіравана на аўтаматызацыю працэсу апрацоўкі тэкставай і гукавой інфармацыі на розных узроўнях. Апісаныя інструменты для пошуку і апрацоўкі фанетычных з'яў прыдатныя для выкарыстання ў шэрагу разнастайных навуковых сфер, напрыклад, у лінгвістычных даследаваннях ці пры стварэнні адмысловых дапаможнікаў па вывучэнні беларускай фанетыкі. Яны спрашчаюць працу фанетыстаў, філолагаў, лінгвістаў і простых карыстальнікаў, што садзейнічае эканоміі часу і змяншэнню колькасці памылак падчас ручной апрацоўкі інфармацыі. Таксама мэтазгоднасць распрацоўкі і выкарыстання дадзеных рэсурсаў абумоўлена іх убудаваннем у беларускамоўныя сістэмы сінтэзу маўлення [10], якія агучваюць тэксты на рускай, беларускай і англійскай мовах. Прадстаўленыя ў артыкуле сэрвісы закліканы аўтаматызаваць працэс далейшага зніжэння аб'ёму даных для навучання сістэм сінтэзу маўлення і паляпшэння іх якасці. Гэта абумоўлена запатрабаванасцю дадзеных тэхналогій, а менавіта тым, што наяўнасць якасных мадэлей сінтэзу маўлення адкрывае для беларускай мовы перспектывы далейшага развіцця больш складаных моўных тэхналогій: галасавы ўвод тэксту, галасавыя дапаможнікі, аўтаматызаванае навучанне беларускай мове, галасавыя чат-боты і інш.

Спіс выкарыстаных крыніц

1. Арфаэпічны слоўнік беларускай мовы / Нац. акад. навук Беларусі, Інстытут мовазнаўства імя Якуба Коласа, Аб'яднаны інстытут праблем інфарматыкі ; уклад.: В. П. Русак [і інш.] ; рэдкал.: В. П. Русак, Ю. С. Гецэвіч. – Мінск : Беларус. навука, 2017. – 757 с.

2. Лабараторыя распазнавання і сінтэзу маўлення [Электронны рэсурс]. – 2022. Рэжым доступу: http://ssrlab.by/. – Дата доступу: 18.02.2020.

3. Лематызатар // Платформа для апрацоўкі тэкставай і гукавой інфармацыі для розных тэматычных даменаў corpus.by [Электронны рэсурс]. – 2022. Рэжым доступу: https://corpus.by/Lemmatizer/?lang=be. – Дата доступу: 21.02.2022.

4. Лысы, С. І. Фанетычная мінімізацыя корпуса тэкстаў на беларускай мове для навучання сістэмы сінтэзу маўлення / С. І. Лысы // Информатика. – 2019. – Т. 16, № 1. – С. 75–85.

5. Падзел на склады // Платформа для апрацоўкі тэкставай і гукавой інфармацыі для розных тэматычных даменаў corpus.by [Электронны рэсурс]. – 2022. – Рэжым доступу: https://corpus.by/Syllabifier/?lang=be. – Дата доступу: 11.04.2022.

6. Платформа для апрацоўкі тэкставай і гукавой інфармацыі для розных тэматычных даменаў corpus.by [Электронны рэсурс]. – 2019. – Рэжым доступу: http://corpus.by/. – Дата доступу: 12.07.2021.

7. Пошук фанетычных з'яў // Платформа для апрацоўкі тэкставай і гукавой інфармацыі для розных тэматычных даменаў corpus.by [Электронны рэсурс]. – 2022. – Рэжым доступу: https://corpus.by/PhoneticPhenomenaSearcher/?lang=be. – Дата доступу: 04.02.2022.

8. Праблемы нормы, культура мовы і генератар маўлення / В. П. Русак [і інш.] // Зборнік дакладаў і тэзісаў VIII Міжнароднай навукова-практычнай канферэнцыі «Традыцыі і сучасны стан культуры і мастацтваў» (Мінск, Беларусь, 7–8 верасня 2017 года) / Цэнтр даследаванняў беларускай культуры, мовы і літаратуры НАН Беларусі ; гал. рэд. А. І. Лакотка. – Мінск : Права і эканоміка, 2018. – С. 748–752.

9. Роля сучасных камп'ютарна-лінгвістычных рэсурсаў у фарміраванні культуры вуснай і пісьмовай мовы / В. П. Русак [і інш.] // Першы міжнародны навуковы кангрэс беларускай культуры : зб. матэрыялаў (Мінск, Беларусь, 5–6 мая 2016 г.) / Цэнтр даследаванняў беларускай культуры, мовы і літаратуры НАН Беларусі ; гал. рэд. А. І. Лакотка. – Мінск : Права і эканоміка, 2016. – С. 364–366.

10. Сінтэзатар маўлення па тэксце [Электронны рэсурс]. – 2022. – Рэжым доступу: http://corpus.by/TextToSpeechSynthesizer/?lang=be. – Дата доступу: 18.03.2022.

11. Фанетычны мінімізатар // Платформа для апрацоўкі тэкставай і гукавой інфармацыі для розных тэматычных даменаў corpus.by [Электронны рэсурс]. – 2022. – Рэжым доступу: https://corpus.by/PhoneticMinimizer/?lang=be. – Дата доступу: 17.03.2022.