Методыка вычыткі электронных тэкстаў вялікага памеру пры дапамозе сэрвісаў платформы www.corpus.by

Ніжэй прыведзена методыка вычыткі электроннага тэксту праз праграмнае забеспячэнне, распрацаванае супрацоўнікамі Лабараторыі распазнавання і сінтэзу маўлення Аб'яднанага інстытута праблем інфарматыкі Нацыянальнай акадэміі навук Беларусі. Выкананне дадзенай методыкі дазваляе атрымаць вычытаны, арфаграфічна правільны тэкст на беларускай мове.

Прапанаваным праграмным забеспячэннем з'яўляюцца сэрвісы апрацоўкі электроннай тэкставай інфармацыі, якія размешчаны на Інтэрнэт-платформе для апрацоўкі тэксту і маўлення www.corpus.by. Дадзенае праграмнае забеспячэнне працуе ў рэжыме анлайн і не патрабуе ўсталявання на камп'ютар.

Сутнасць методыкі вычыткі палягае ў апрацоўцы тэксту наступнымі анлайн-сэрвісамі:

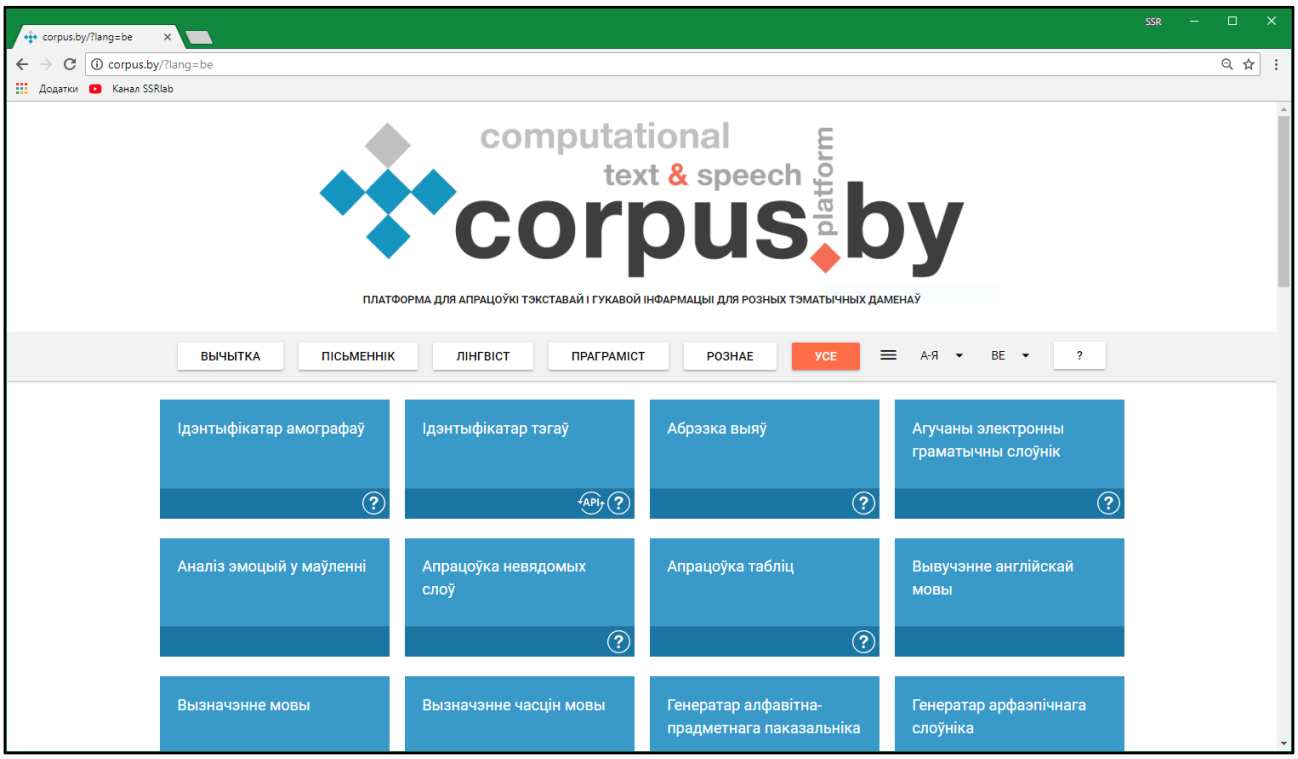
- «Падлік частотнасці сімвалаў»;
- «Праверка правапісу»;
- «Праверка правапісу "Ў"»;
- «Ідэнтыфікатар амографаў».

Па выніках апрацоўкі тэксту кожным сэрвісам карыстальнік можа праглядаць атрыманыя вынікі і па жаданні ўносіць у тэкст адпаведныя праўкі. Таксама вынікі можна скапіраваць, уставіць у файл *.doc і захаваць.

Дадзеная методыка вычыткі ахоплівае арфаграфічны раздзел правапісу, але не ахоплівае пунктуацыйны і сінтаксічны раздзелы: ***правільнасць дапасавання слоў і расстаноўкі знакаў прыпынку знаходзіцца па-за кампетэнцыяй дадзенай методыкі***. Таксама методыка апрабаваная ў шматлікіх праектах Лабараторыі распазнавання і сінтэзу маўлення АІПІ НАН Беларусі, таму рэкамендавана ажыццяўляць апрацоўку сэрвісамі ў прыведзеным ніжэй парадку, які змяшчае 6 паслядоўных этапаў.

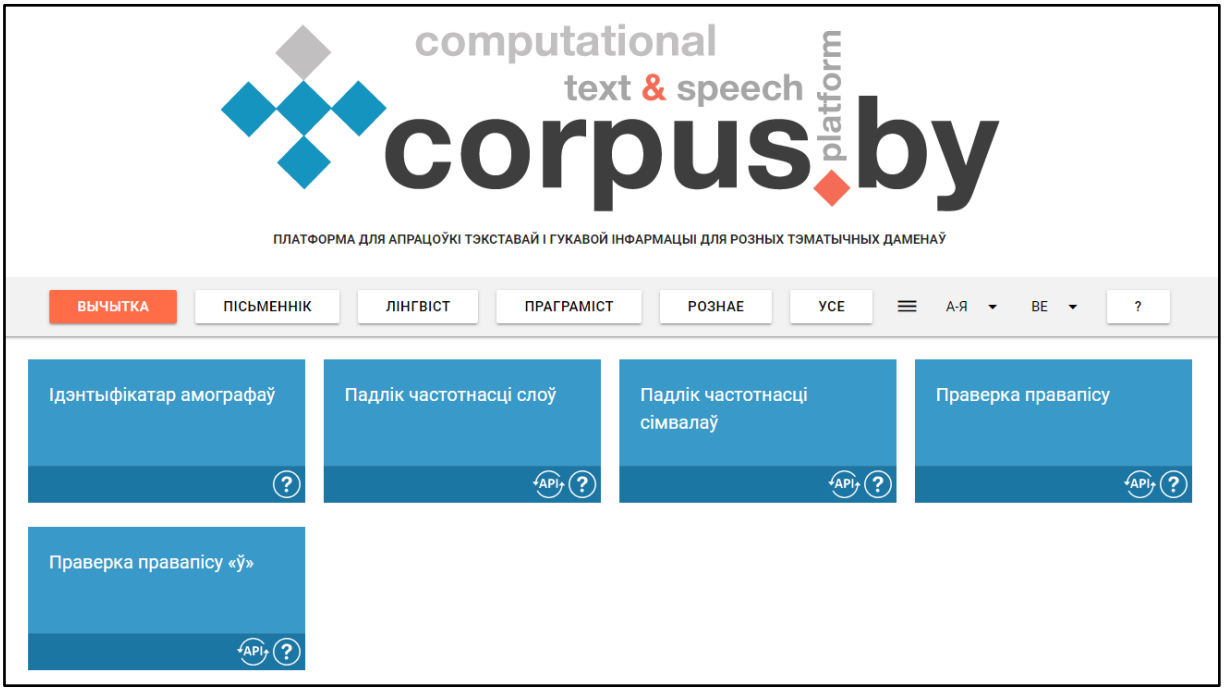
Працэс вычыткі

Для атрымання арфаграфічна правільнага тэксту неабходна прайсці 4 этапы вычыткі праз сэрвісы платформы www.corpus.by (малюнак 1). Перад пачаткам унясення правак рэкамендавана зрабіць копію правяраемага тэксту.



Малюнак 1. Галоўная старонка камп'ютарна-лінгвістычнай платформы www.corpus.by

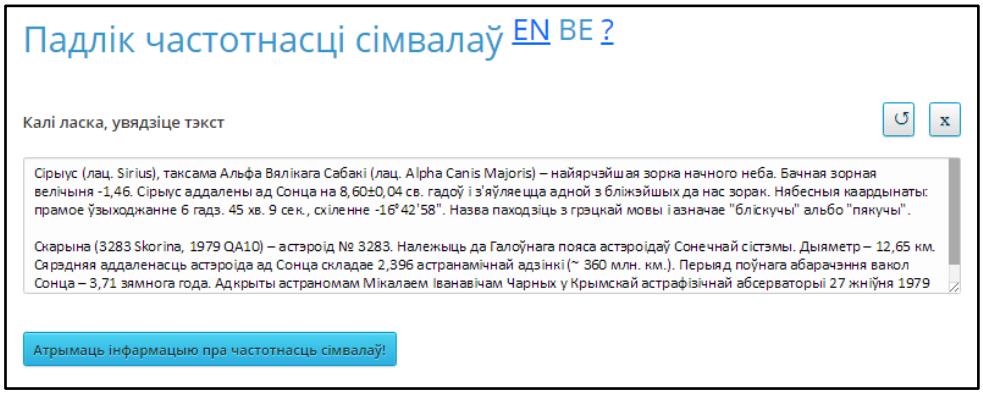
На галоўнай старонцы платформы для большай зручнасці працы трэба націснуць кнопку «Вычытка» (малюнак 2). Застануцца толькі сэрвісы, задзейнічаныя ў вычытцы тэксту.



Малюнак 2. Сэрвісы вычыткі на платформе www.corpus.by

Этап 1. Вычытка праз сэрвіс «Падлік частотнасці сімвалаў»

Сэрвіс «Падлік частотнасці сімвалаў» (малюнак 3) прадэманструе спіс усіх сімвалаў, выкарыстаных у тэксце, і дазволіць выявіць і выправіць іх памылковае выкарыстанне.



Малюнак 3. Інтэрфейс сэрвіса[1] «Падлік частотнасці сімвалаў»

Для атрымання вынікаў неабходна скапіраваць і ўставіць тэкст у поле ўводу, пасля чаго націснуць кнопку «Атрымаць інфармацыю пра частотнасць сімвалаў!». Адлюструюцца вынікі (малюнак 4).

ІНФАРМАЦЫЯ ПРА ЎСЕ ЗНОЙДЗЕНЫЯ СІМВАЛЫ:

АГУЛЬНАЯ КОЛЬКАСЦЬ СІМВАЛАЎ У ТЭКСЦЕ: **836**
КОЛЬКАСЦЬ УНІКАЛЬНЫХ СІМВАЛАЎ У ТЭКСЦЕ: **86**

↓С.	Код	Назва	Частата		Кантэкст
	U+000A	ПЕРАВОД РАДКА	2	0.24%	ліскучы" альбо "пякучы".
	U+000D	ВЯРТАННЕ КАРЭТКІ	2	0.24%	ліскучы" альбо "пякучы".
	U+0020	ПРАБЕЛ	118	14.11%	Сірыус (лац. Sirius), таксам
"	U+0022	ЗНАК ДВУКОССЯ	4	0.48%	з грэцкай мовы і азначае "б
'	U+0027	АПОСТРАФ	4	0.48%	на 8,60±0,04 св. гадоў і з'я
(	U+0028	ЛЕВАЯ КРУГЛАЯ ДУЖКА	4	0.48%	Сірыус (лац. Sirius), таксам
)	U+0029	ПРАВАЯ КРУГЛАЯ ДУЖКА	4	0.48%	Сірыус (лац. Sirius), таксам
,	U+002C	КОСКА	9	1.08%	Сірыус (лац. Sirius), таксам
-	U+002D	ЗЛУЧОК-МІНУС	2	0.24%	Бачная зорная велічыня -1,4
.	U+002E	КРОПКА	20	2.39%	Сірыус (лац. Sirius), таксам
0	U+0030	ЛІЧБА НУЛЬ	5	0.6%	на 8,60±0,04 св. гадоў і з'я
1	U+0031	ЛІЧБА АДЗІН	7	0.84%	Бачная зорная велічыня -1,4
2	U+0032	ЛІЧБА ДВА	6	0.72%	-16°42'58". Назва паходзіць
3	U+0033	ЛІЧБА ТРЫ	7	0.84%	Скарына (3283 Skorina, 1979

Малюнак 4. Вынікі працы сэрвіса «Падлік частотнасці сімвалаў»

[1] Тут і далей адмыслова выкарыстана форма роднага склону з канчаткам «-а», нягледзячы на тое, што нарматыўнай з'яўляецца форма «сэрвіс**у**». У афіцыйнай крыніцы пад словам «сэрвіс» маецца на ўвазе абстрактнае значэнне — «ажыццяўленне паслуг», у той час як у дадзеным выпадку «сэрвіс» набывае канкрэтнае злічальнае значэнне — «алгарытм апрацоўкі ўваходных дадзеных».
Такім чынам, паводле правілаў напісання канчаткаў назоўнікаў роднага склону адзіночнага ліку мужчынскага роду гэтае значэнне пападае пад катэгорыю злічальных прадметаў і мусіць мець канчатак «-а».

Неабходна прагледзець выніковы спіс і праверыць у ім наступныя моманты:

– ці аднолькавая колькасць дужак **( )**, **[]**;
– ці аднолькавая колькасць падвоеных двукоссяў **“ ”** і **« »**;
– ці прысутнічаюць у тэксце адзіночныя двукоссі **"**, якія не павінны прысутнічаць, калі ўжо выкарыстоўваюцца падвоеныя двукоссі **“ ”** і **« »**;
– ці правільна выкарыстоўваецца злучок /-/, кароткі /–/, доўгі /—/ працяжнікі;
– ці прысутнічаюць лацінскія літары ў кірылічным тэксце.

Так, напрыклад, калі колькасць левых і правых дужак не супадае, то, хутчэй за ўсё, у тэксце ёсць пунктуацыйныя памылкі.

Па знойдзеных памылковых ужываннях сімвалаў неабходна ўнесці праўкі ў тэкст, пераправерыць сэрвісам яшчэ раз скарэктаваны тэкст, і перайсці да наступнага этапу вычыткі.

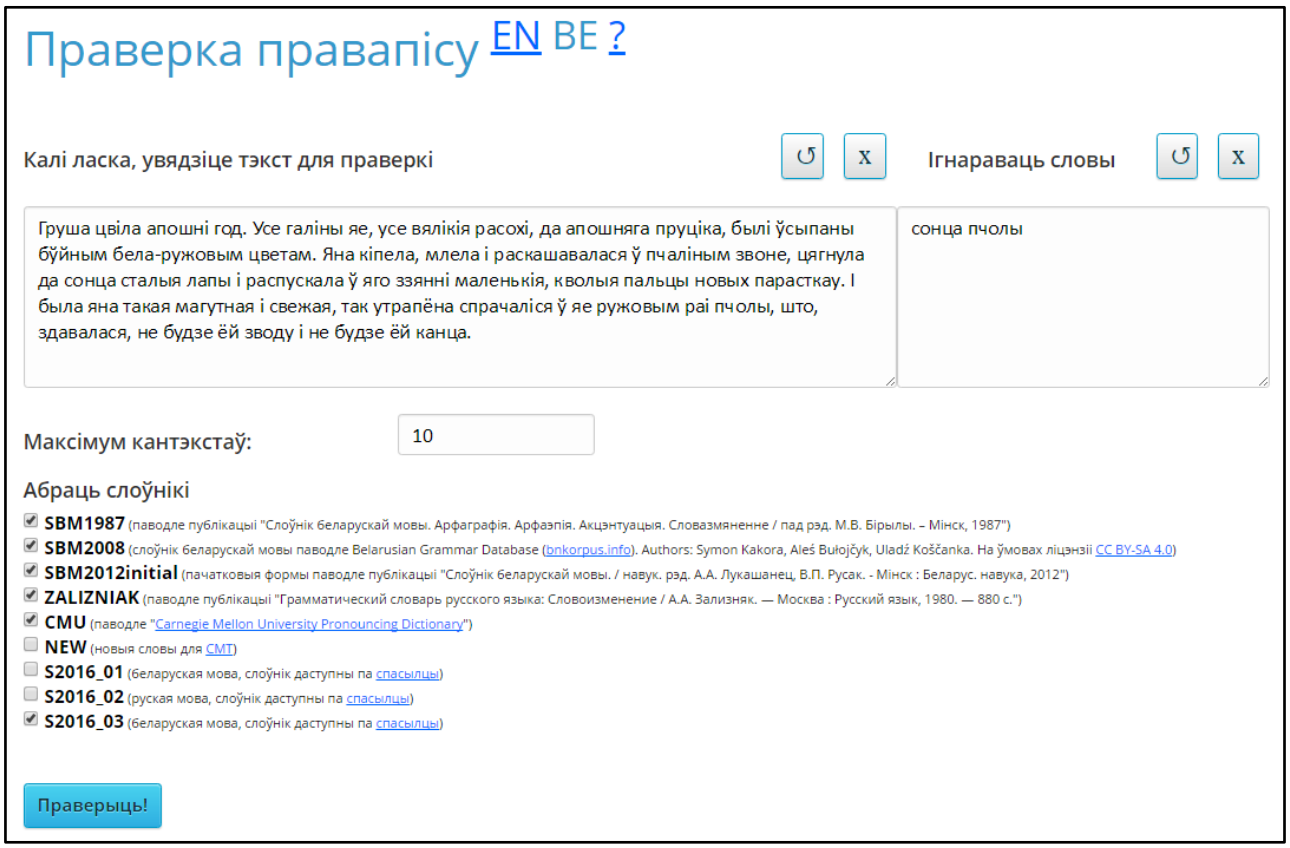
Сэрвіс «Падлік частотнасці сімвалаў» даступны па спасылцы:
http://corpus.by/CharacterFrequencyCounter/?lang=be
Падрабязная інструкцыя па карыстанні сэрвісам: http://ssrlab.by/3323

Этап 2. Вычытка праз сэрвіс «Праверка правапісу»

Сэрвіс «Праверка правапісу» (малюнак 5) выяўляе:

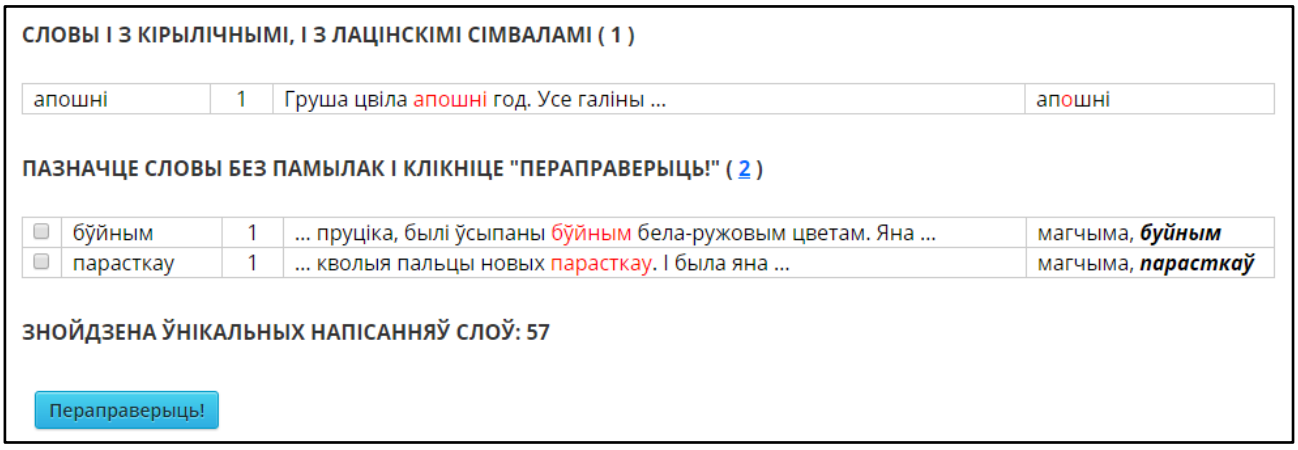
1. Словы, у якіх ужытыя лацінскія сімвалы: калі ў словах кірылічнага напісання ёсць візуальна аднолькавыя лацінскія сімвалы (/a/, /i/, /c/ і г.д.), то тэкст будзе ў далейшым паўсюль апрацоўвацца недакладна, таму такія сімвалы трэба выявіць і замяніць кірылічнымі.

2. Словы з памылкамі: слова, напісанае з памылкай, адсутнічае ў слоўніках і будзе пазначана сэрвісам як невядомае. У спіс невядомых слоў таксама трапляюць словы, якія не змяшчаюць памылку, але адсутнічаюць у слоўніку. Таксама сэрвіс дае магчымасць ігнараваць пэўныя словы. Гэтая магчымасць можа спатрэбіцца пры вычытцы карыстальнікам вузкаспецыяльнага тэксту, каб выключыць пападанне загадзя невядомага сэрвісу слова ў спіс невядомых і паскорыць прагляд гэтага спісу.



Малюнак 5. Інтэрфейс сэрвіса «Праверка правапісу»

Каб атрымаць спіс слоў, у якіх, верагодна, знаходзяцца памылкі ці лацінскія сімвалы, трэба ўставіць тэкст у поле ўводу і націснуць кнопку «Праверыць!». Невядомыя словы і словы з лацінскімі сімваламі выводзяцца ў выглядзе спісу з магчымасцю пабачыць кантэкст іх выкарыстання (малюнак 6).



Малюнак 6. Вынікі працы сэрвіса «Праверка правапісу»

Неабходна прагледзець спіс невядомых сэрвісу слоў і знайсці сярод іх словы, напісаныя з памылкамі. Можна, але неабавязкова, гачыкам злева адзначыць правільна напісаныя словы і націснуць «Пераправерыць!» — спіс гэтых слоў можна будзе скапіраваць і ўставіць у поле выключэнняў. Гэтая магчымасць карысная пры працы з вузкаспецыяльным тэкстам вялікага памеру, дзе шмат тэрмінаў, імёнаў уласных ці інш. Па знойдзеных памылках неабходна

ўнесці праўкі ў тэкст, пераправерыць сэрвісам яшчэ раз скарэктаваны тэкст, і перайсці да наступнага этапу вычыткі.

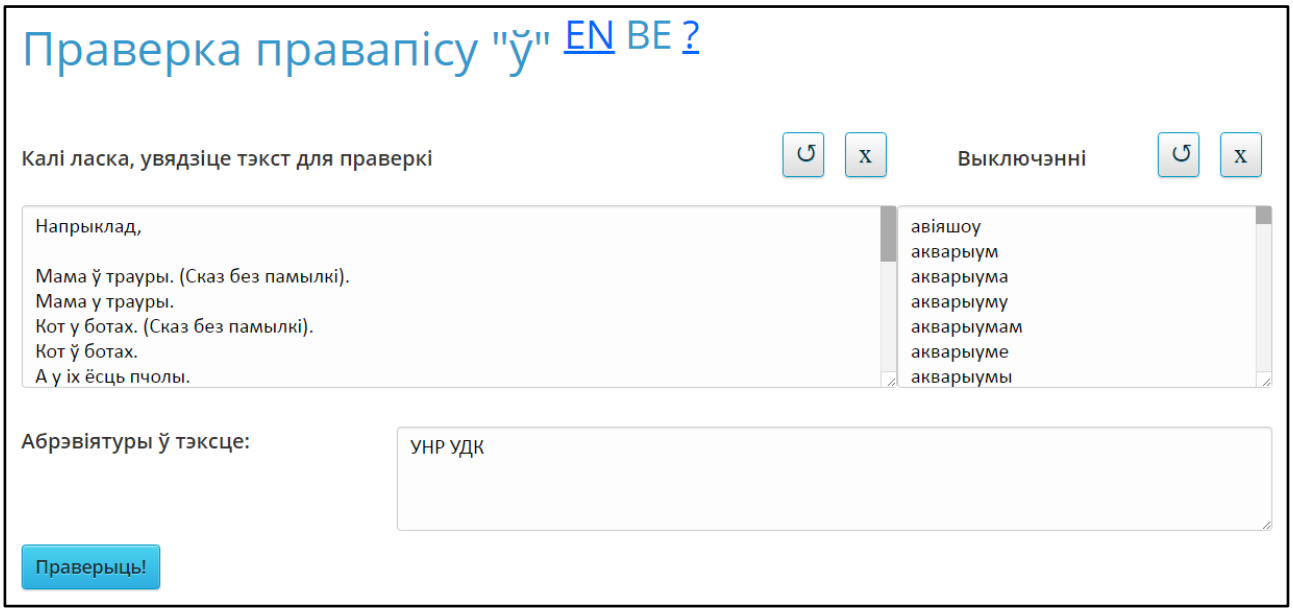
Сэрвіс «Праверка правапісу» даступны па спасылцы:
http://corpus.by/SpellChecker/?lang=be
Больш падрабязная інструкцыя па карыстанні сэрвісам:
http://ssrlab.by/3334

Этап 3. Вычытка праз сэрвіс «Праверка правапісу “Ў”»

Сэрвіс «Праверка правапісу “Ў”» (малюнак 7) правярае правільнасць ужывання ў тэксце літар «у» і «ў»

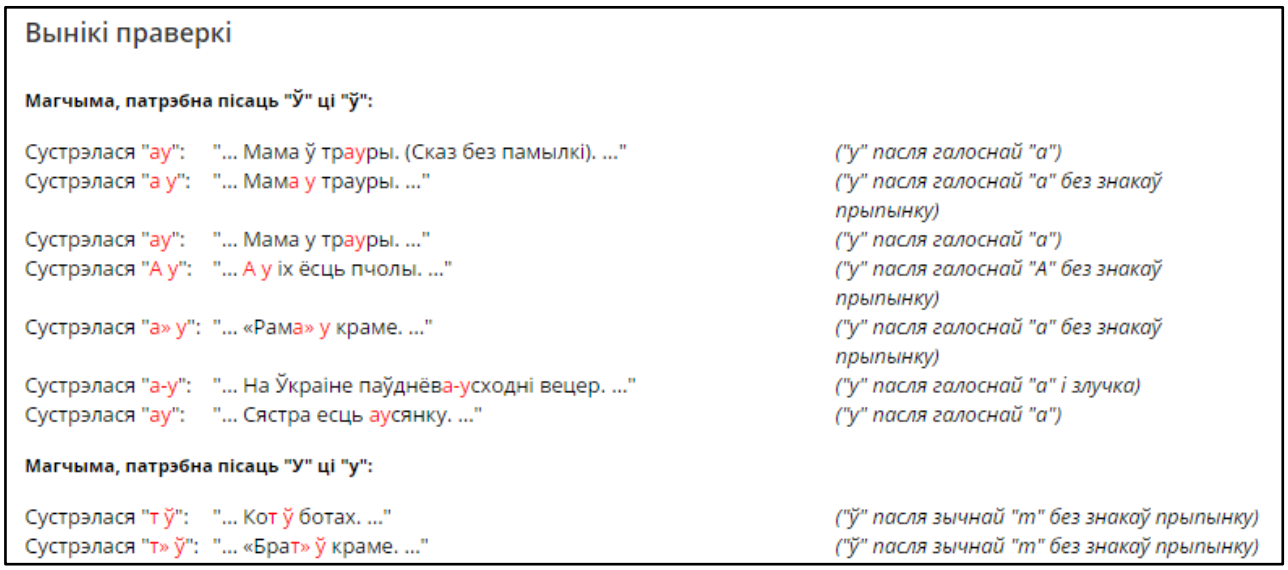
Алгарытм сэрвіса шукае сімвалы /у/, /У/, /ў/, /Ў/ і глядзіць на папярэдні сімвал, такім чынам правяраючы правільнасць ужывання.

Сэрвіс мае поле «Выключэнні», дзе змяшчаюцца актуальныя на гэты момант выключэнні з правілаў правапісу літары «ў». Поле можна рэдагаваць: выдаляць уведзеныя па змаўчанні і дадаваць неабходныя карыстальніку выключэнні.



Малюнак 7. Інтэрфейс сэрвіса «Праверка правапісу “Ў”»

Для атрымання вынікаў неабходна скапіраваць і ўставіць тэкст у поле ўводу, пасля чаго націснуць кнопку «Праверыць!». Адлюструюцца вынікі (малюнак 8).



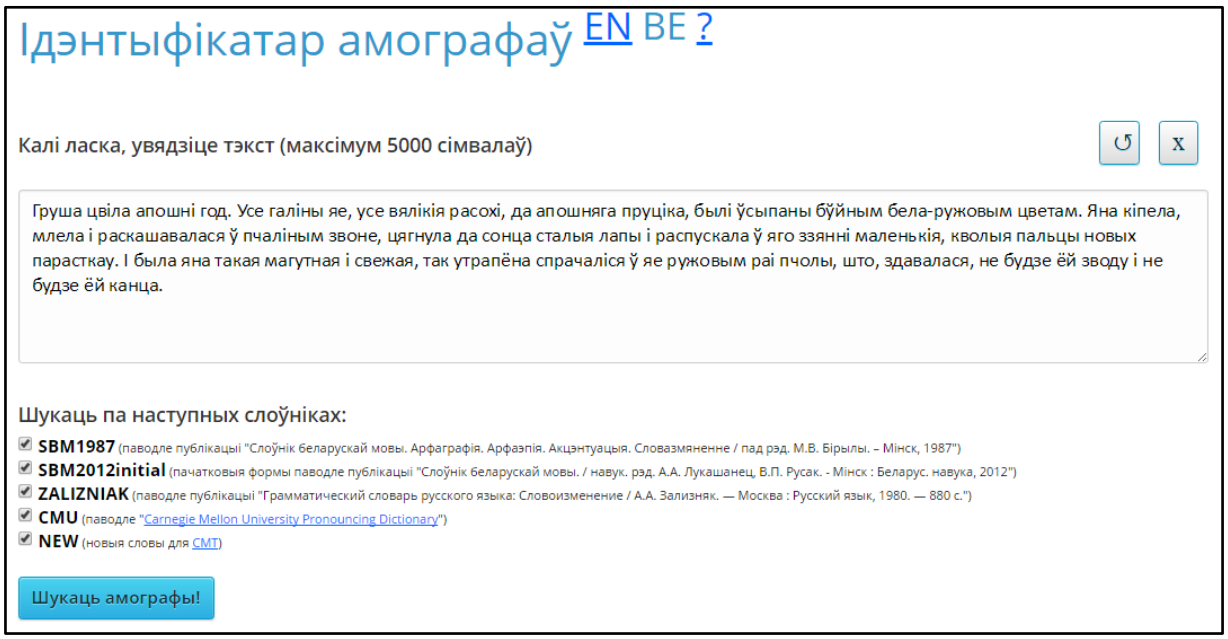
Малюнак 8. Вынікі працы сэрвіса «Праверка правапісу Ў»

Неабходна прагледзець спіс месцаў, дзе сэрвіс знайшоў выпадкі няправільнага ўжывання «у» і «ў». Па знойдзеных памылках трэба ўнесці праўкі ў тэкст, пераправерыць сэрвісам яшчэ раз скарэктаваны тэкст, і перайсці да наступнага этапу вычыткі.

Сэрвіс «Праверка правапісу “Ў”» даступны па спасылцы:
http://corpus.by/ShortUSpellChecker/?lang=be
Падрабязная інструкцыя па карыстанні сэрвісам:
http://ssrlab.by/1404

Этап 4. Вычытка праз сэрвіс «Ідэнтыфікатар амографаў»

Сэрвіс «Ідэнтыфікатар амографаў» (малюнак 9) паказвае карыстальніку амографы – словы, якія маюць аднолькавае напісанне і рознае вымаўленне, напрыклад, *му́зыка* і *музы́ка*.



Малюнак 9. Інтэрфейс сэрвіса «Ідэнтыфікатар амографаў»

Каб атрымаць спіс выкарыстаных у тэксце амографаў, трэба ўставіць тэкст у поле ўводу і націснуць кнопку «Шукаць амографы!». За адзін раз сэрвіс можа апрацаваць тэкст аб’ёмам каля 20-30 старонак. Для зручнасці таксама пададзены кантэксты, у якіх знаходзяцца амографы (малюнак 10).

Знойдзена амографаў: 3 (паглядзець спіс амографаў можна тут)

Амограф	Варыянты націску	Тып амографа	↓ Кольк.	Кантэксты	Слоўнік
галіны	галі́ны галіны́	адна парадыгма	1	... апошні год. Усе галіны яе, усе вялікія ...	SBM1987
раі	ра́і раі́	адна часціна мовы	1	... ў яе ружовым раі пчолы, што, здавалася, ...	SBM1987
былі	бы́лі былі́	розныя часціны мовы	1	... да апошняга пруціка, былі ўсыпаны бўйным бела-ружовым ...	SBM1987

Малюнак 10. Вынікі працы сэрвіса «Ідэнтыфікатар амографаў»

Неабходна ўважліва прагледзець спіс знойдзеных амографаў і прыняць рашэнне, у якіх словах-амографах і на якіх складах у тэксце трэба пазначыць націскі, і ўнесці гэтыя націскі ў тэкст.

Націск у праграме Microsoft Office Word дадаецца наступным чынам: трэба паставіць курсор пасля літары, на якой неабходны націск, заціснуць клавішу Alt і набраць на лічбавай клавіятуры справа паслядоўнасць лічбаў 0769.

Сэрвіс «Ідэнтыфікатар амографаў» даступны па спасылцы:
http://corpus.by/HomographIdentifier/?lang=be
Больш падрабязная інструкцыя па карыстанні сэрвісам:
http://ssrlab.by/4218

Выкананне ўсіх этапаў дадзенай методыкі дазваляе атрымаць вычытаны, арфаграфічна правільны тэкст на беларускай мове.

Пералічаныя ў гэтай метадычцы сэрвісы і апісаны алгарытм вычыткі знаходзяцца ў стане пастаяннай дапрацоўкі і ўдасканалення. Лабараторыя распазнавання і сінтэзу маўлення вітае ўсе заўвагі і прапановы па паляпшэнні працы сэрвісаў і дадзенай методыкі.

Кантактныя дадзеныя для зваротнай сувязі:

Лабараторыя распазнавання і сінтэзу маўлення

Адрас: вул. Сурганава, 6, пакоі 422, 430 і 432
220012, г. Мінск, Беларусь

Тэл.: +375 (17) 284-27-73 (пакой 422)
Факс: +375 17 284-21-75 (прыёмная Інстытута)
E-mail: yuras.hetsevich@newman.bas-net.by, ssrlab221@gmail.com

Нацыянальная акадэмія навук Беларусі
Аб’яднаны інстытут праблем інфарматыкі