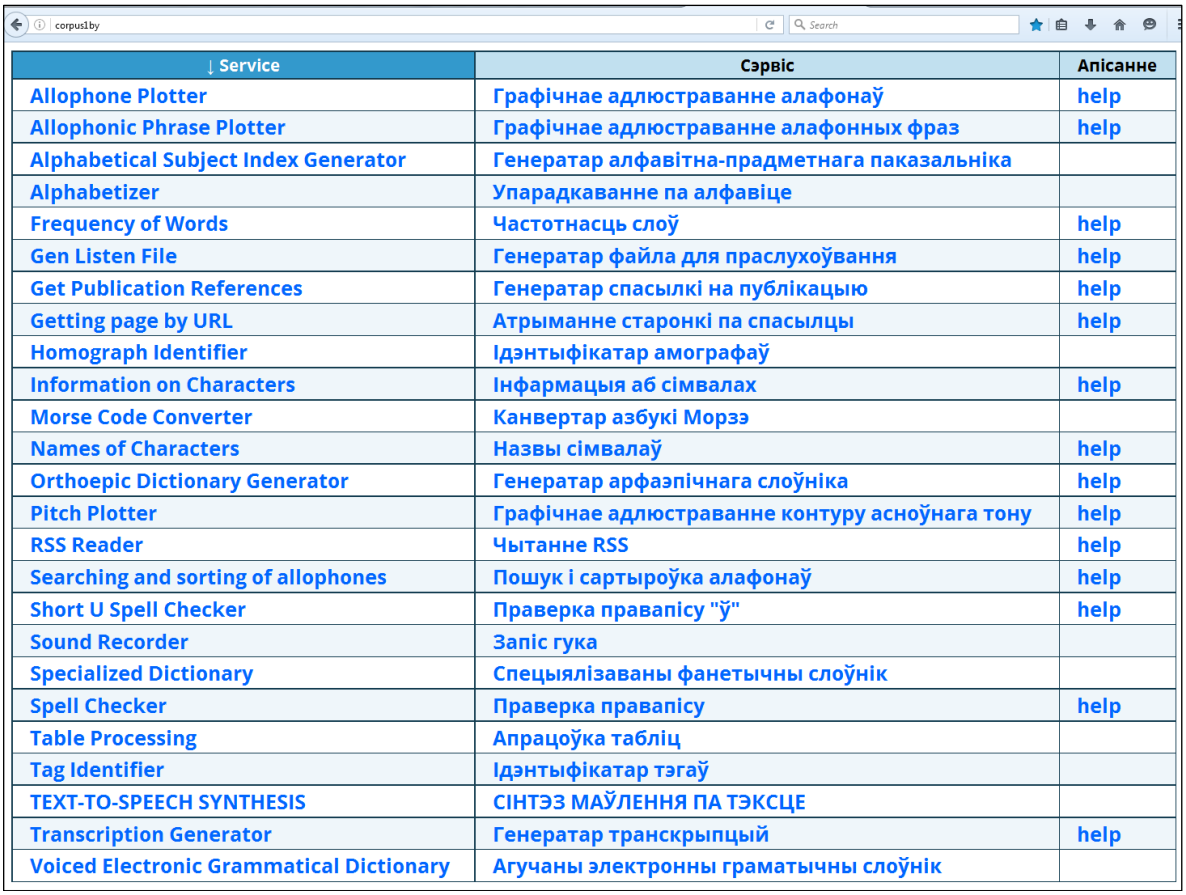
Г.Р. Станіславенка, С.І. Лысы, Ю.С. Гецэвіч (Мінск, НАН РБ)7

РЭДАГАВАННЕ ЭЛЕКТРОННЫХ МАСІВАЎ ТЭКСТАЎ НА БЕЛАРУСКАЙ МОВЕ З ВЫКАРЫСТАННЕМ КАМП'ЮТАРНА-ЛІНГВІСТЫЧНЫХ СЭРВІСАЎ ПЛАТФОРМЫ WWW.CORPUS. BY

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. Дадзены артыкул прысвечаны апісанню методыкі рэдагавання электронных масіваў тэкстаў на беларускай мове з выкарыстаннем сэрвісаў, распрацаваных лабараторыяй распазнавання і сінтэзу маўлення Аб'яднанага інстытута праблем інфарматыкі Нацыянальнай акадэміі навук Беларусі. Гэтыя сэрвісы размешчаны на платформе www.Corpus.by [1] (малюнак 1) і маюць мэту, вырашаць падзадачы сістэм сінтэзу і распазнавання маўлення. Але ў той жа час дадзеныя сэрвісы могуць выкарыстоўвацца для вырашэння праблем лексікаграфічнага і граматычнага характару ў працы з тэкстамі.



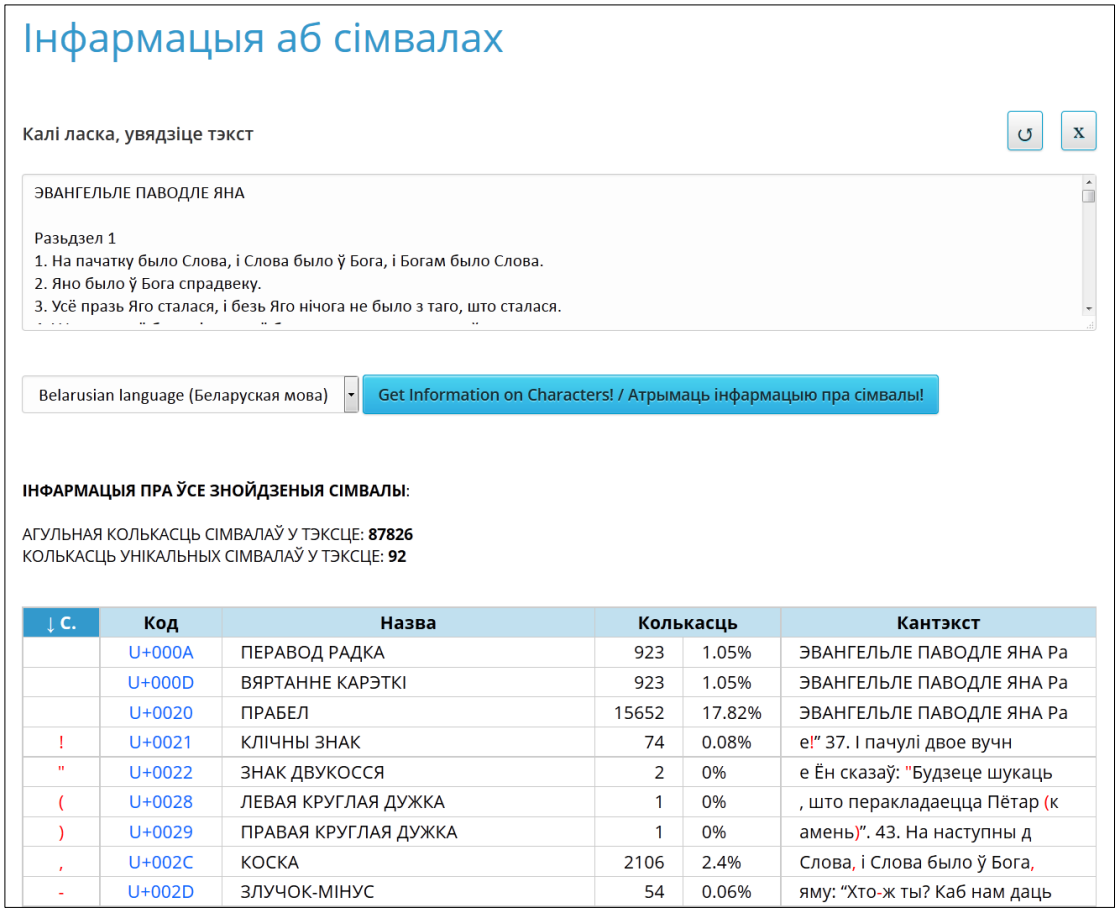
↓ Service	Сэрвіс	Апісанне
Allophone Plotter	Графічнае адлюстраванне алафонаў	help
Allophonic Phrase Plotter	Графічнае адлюстраванне алафонных фраз	help
Alphabetical Subject Index Generator	Генератар алфавітна-прадметнага паказальніка	
Alphabetizer	Упарадкаванне па алфавіце	
Frequency of Words	Частотнасць слоў	help
Gen Listen File	Генератар файла для праслухоўвання	help
Get Publication References	Генератар спасылкі на публікацыю	help
Getting page by URL	Атрыманне старонкі па спасылцы	help
Homograph Identifier	Ідэнтыфікатар амографаў	
Information on Characters	Інфармацыя аб сімвалах	help
Morse Code Converter	Канвертар азбукі Морзэ	
Names of Characters	Назвы сімвалаў	help
Orthoepic Dictionary Generator	Генератар арфаэпічнага слоўніка	help
Pitch Plotter	Графічнае адлюстраванне контуру асноўнага тону	help
RSS Reader	Чытанне RSS	help
Searching and sorting of allophones	Пошук і сартыроўка алафонаў	help
Short U Spell Checker	Праверка правапісу "ў"	help
Sound Recorder	Запіс гука	
Specialized Dictionary	Спецыялізаваны фанетычны слоўнік	
Spell Checker	Праверка правапісу	help
Table Processing	Апрацоўка табліц	
Tag Identifier	Ідэнтыфікатар тэгаў	
TEXT-TO-SPEECH SYNTHESIS	СІНТЭЗ МАЎЛЕННЯ ПА ТЭКСЦЕ	
Transcription Generator	Генератар транскрыпцый	help
Voiced Electronic Grammatical Dictionary	Агучаны электронны граматычны слоўнік	

Мал. 1. Галоўная старонка інтэрнэт-рэсурсу www.Corpus.by

Дадзеная методыка складаецца з 4 этапаў, але аўтары лічаць вартым адзначыць, што алгарытм можа дапаўняцца іншымі этапамі ў залежнасці ад пастаўленых задач.

Першым этапам вычыткі з'яўляецца сімвальная апрацоўка тэксту праз сэрвіс «Інфармацыя аб сімвалах» [2] (малюнак 2), які паказвае статыстыку ўжывання сімвалаў, што выкарыстаны ў тэксце. Гэта дазваляе выявіць і выправіць памылковае ўжыванне тых ці іншых сімвалаў, што паляпшае далейшую працу з тэкстам і робіць больш карэктнай яго далейшую апрацоўку іншымі сэрвісамі, якія працуюць толькі з абмежаваным наборам сімвалаў. Варта адзначыць важнасць гэтага сэрвісу для праверкі пунктуацыйных памылак, так як ён паказвае колькасць парных сімвалаў (напрыклад, дужак, двукоссяў): несупадзенне колькасці адпаведных левых і правых сімвалаў паказвае на пунктуацыйную памылку. Таксама праз сэрвіс можна праверыць выкарыстанне кароткіх і доўгіх працяжнікаў і дэфісаў, слэшаў, розных відаў двукоссяў і інш.

Для рэдагавання тэкстаў на беларускай мове *другім этапам* з'яўляецца праверка правапісу «ў». Для гэтай мэты распрацаваны адмысловы сэрвіс «Праверка правапісу "ў"» [3]. Алгарытм сэрвісу працуе наступным чынам: адбываецца пошук усіх літар «у» і «ў» іаналізуецца папярэдні сімвал, асобна разглядаюцца вялікія літары «Ў», словы, што сканчваюцца на «-ум» і «-ус» і некаторыя іншыя выпадкі. Такім чынам, сэрвіс аўтаматычна аналізуе тэкст на падставе некалькіх простых правіл без выкарыстання слоўнікаў і спісаў выключэнняў. Пры такім алгарытме праверкі правапісу сэрвіс выдае толькі рэкамендацыі, а канчатковае рашэнне павінна заставацца за карыстальнікам сэрвісу. Прыклад працы сэрвісу прадстаўлены на малюнку 3.



Інфармацыя аб сімвалах

Калі ласка, увядзіце тэкст

ЭВАНГЕЛЬЛЕ ПАВОДЛЕ ЯНА

Разьдзел 1
1. На пачатку было Слова, і Слова было ў Бога, і Богам было Слова.
2. Яно было ў Бога спрадвеку.
3. Усё празь Яго сталася, і безь Яго нічога не было з таго, што сталася.

Belarusian language (Беларуская мова) | Get Information on Characters! / Атрымаць інфармацыю пра сімвалы!

ІНФАРМАЦЫЯ ПРА ЎСЕ ЗНОЙДЗЕНЫЯ СІМВАЛЫ:

АГУЛЬНАЯ КОЛЬКАСЦЬ СІМВАЛАЎ У ТЭКСЦЕ: **87826**
КОЛЬКАСЦЬ УНІКАЛЬНЫХ СІМВАЛАЎ У ТЭКСЦЕ: **92**

↓ С.	Код	Назва	Колькасць		Кантэкст
	U+000A	ПЕРАВОД РАДКА	923	1.05%	ЭВАНГЕЛЬЛЕ ПАВОДЛЕ ЯНА Ра
	U+000D	ВЯРТАННЕ КАРЭТКІ	923	1.05%	ЭВАНГЕЛЬЛЕ ПАВОДЛЕ ЯНА Ра
	U+0020	ПРАБЕЛ	15652	17.82%	ЭВАНГЕЛЬЛЕ ПАВОДЛЕ ЯНА Ра
!	U+0021	КЛІЧНЫ ЗНАК	74	0.08%	е!” 37. І пачулі двое вучн
"	U+0022	ЗНАК ДВУКОССЯ	2	0%	е Ён сказаў: "Будзеце шукаць
(	U+0028	ЛЕВАЯ КРУГЛАЯ ДУЖКА	1	0%	, што перакладаецца Пётар (к
)	U+0029	ПРАВАЯ КРУГЛАЯ ДУЖКА	1	0%	амень)”. 43. На наступны д
,	U+002C	КОСКА	2106	2.4%	Слова, і Слова было ў Бога,
-	U+002D	ЗЛУЧОК-МІНУС	54	0.06%	яму: “Хто-ж ты? Каб нам даць

Мал. 2. Вынікі працы сэрвісу «Інфармацыя аб сімвалах»

Вынікі праверкі

Магчыма, патрэбна пісаць "Ў" ці "ў":

Сустрэлася "я у":	"... анёлаў Божых, якія узыходзяць і зыходзяць да ..."	*("у" пасля галоснай "я" без знакаў прыпынку)*
Сустрэлася "ау":	"... І быў у Капэрнауме нейкі ўрадовец, у ..."	*("у" пасля галоснай "а")*
Сустрэлася "е у":	"... калі другі прыйдзе у сваё імя, таго ..."	*("у" пасля галоснай "е" без знакаў прыпынку)*
Сустрэлася "ы у":	"... шмат травы. Тады узьлеглі мужчыны лікам каля ..."	*("у" пасля галоснай "ы" без знакаў прыпынку)*
Сустрэлася "ау":	"... сынагозе, навучаючы ў Капэрнауме. ..."	*("у" пасля галоснай "а")*
Сустрэлася "е у":	"... але ніхто не узлажыў на Яго рукі. ..."	*("у" пасля галоснай "е" без знакаў прыпынку)*
Сустрэлася "ю у":	"... Я не раблю учынкаў Айца Майго, ня ..."	*("у" пасля галоснай "ю" без знакаў прыпынку)*
Сустрэлася "ы у":	"... вы рассеецеся кожны у свой бок і ..."	*("у" пасля галоснай "ы" без знакаў прыпынку)*
Сустрэлася "і у":	"... да Ісуса, калі убачылі, што Ён ужо ..."	*("у" пасля галоснай "і" без знакаў прыпынку)*
Сустрэлася "е у":	"... цьвікоў і не уткну пальца майго ў ..."	*("у" пасля галоснай "е" без знакаў прыпынку)*

Магчыма, патрэбна пісаць "У" ці "у":

Сустрэлася "ь ў":	"... глыбокі. Дык адкуль ў Цябе вада жывая? ..."	*("ў" пасля зычнай "ь" без знакаў прыпынку)*
Сустрэлася "к ў":	"... бо Ісус зьнік ў натоўпе, які быў ..."	*("ў" пасля зычнай "к" без знакаў прыпынку)*
Сустрэлася "ь ў":	"... Прарок, Які прыходзіць ў сьвет”. ..."	*("ў" пасля зычнай "ь" без знакаў прыпынку)*
Сустрэлася "к ў":	"... вучняў Ягоных, дык ўвайшлі ў чаўны і ..."	*("ў" пасля зычнай "к" без знакаў прыпынку)*
Сустрэлася "с ў":	"... палове сьвята Ісус ўвайшоў у сьвятыню і ..."	*("ў" пасля зычнай "с" без знакаў прыпынку)*
Сустрэлася "ь ў":	"... што Ісус прыходзіць ў Ерусалім, ..."	*("ў" пасля зычнай "ь" без знакаў прыпынку)*
Сустрэлася "й ў":	"... каб мелі супакой ў-ва Мне. У гэтым ..."	*("ў" пасля зычнай "й" без знакаў прыпынку)*
Сустрэлася "ў ў":	"... я не знайшоў ў Ім віны”. ..."	*("ў" пасля зычнай "ў" без знакаў прыпынку)*
Сустрэлася "ь ў":	"... Пілата, каб перабіць ў іх галёнкі і ..."	*("ў" пасля зычнай "ь" без знакаў прыпынку)*

Мал. 3. Вынікі працы сэрвісу «Праверка правапісу “ў”»

Трэцім этапам з’яўляецца паслоўная апрацоўка тэксту. З дапамогай сэрвісу «Праверка правапісу» ў тэксце адбываецца пошук слоў, якія адсутнічаюць у слоўніках [4]. На сённяшні дзень сэрвіс аналізуе тэксты на падставе наступных слоўнікаў:

– «Слоўнік беларускай мовы. Арфаграфія. Арфаэпія. Акцэнтуацыя. Словазмяненне» / пад рэд. М.В. Бірылы. — Мінск, 1987;

– «Слоўнік беларускай мовы» / навук. рэд. А.А. Лукашанец, В.П. Русак. — Мінск, 2012.

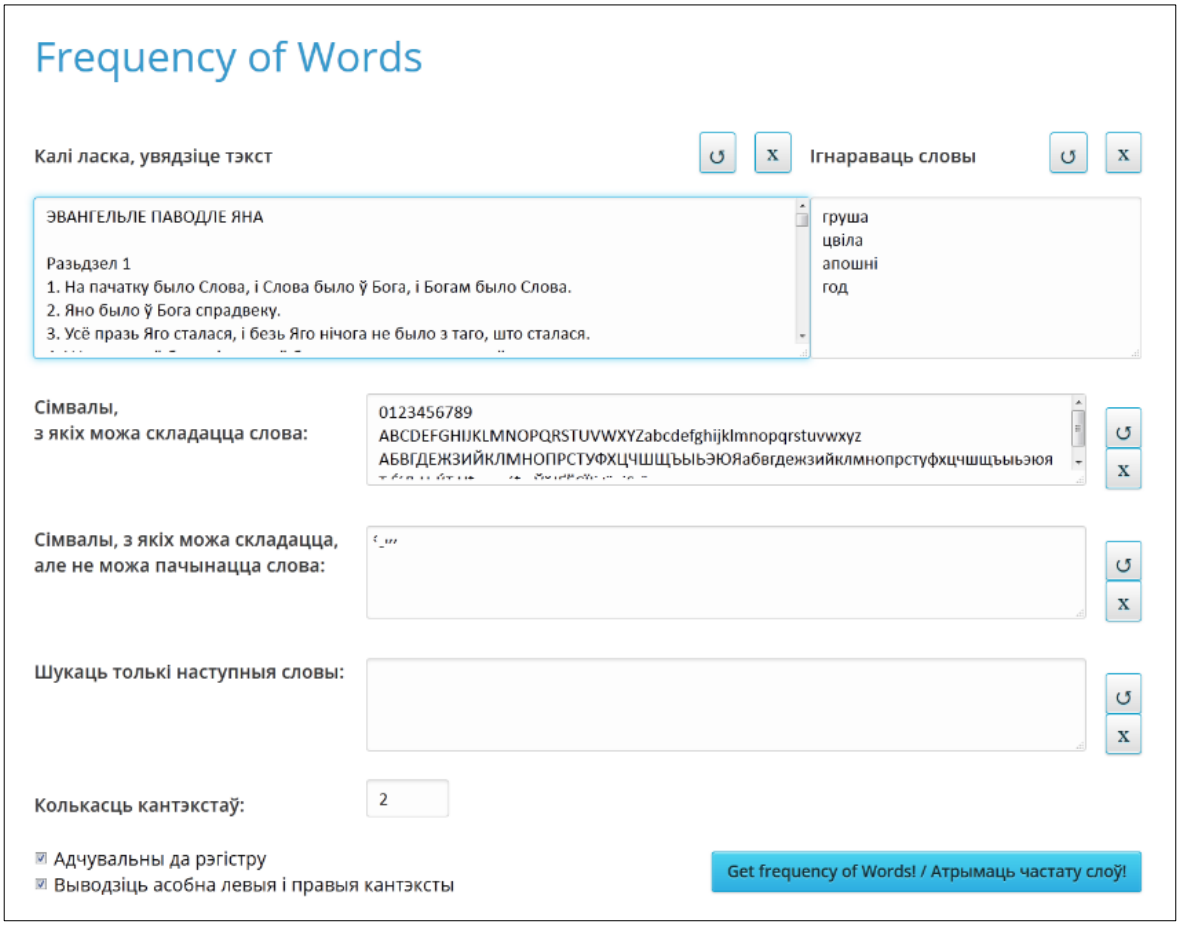
Невядомыя сэрвісу словы выводзяцца ў выглядзе спісу з прыкладамі кантэксту. Калі карыстальнік лічыць, што пэўнае слова з гэтага спісу не мае ў сабе памылкі, то ён можа адзначыць гэта слова, як вядомае. Таксама карыстальнік можа ўвесці ў адмысловае акно спіс слоў, якія не патрэбна правяраць. Гэта могуць быць словы з аўтарскім правапісам, правапісам па старых нормах, імёны ўласныя і г.д. Прыклад працы сэрвісу паказаны на малюнку 4.

ЦІ ВЕДАЕЦЕ ВЫ ГЭТАЕ СЛОВА? (спампаваць спіс)

абвесьціць	Так Не	... калі Ён прыйдзе, абвесьціць нам усё. ...
абвінавальнік	Так Не	... ёсьць на вас абвінавальнік Майсей, на якога ...
абвінавачваньне	Так Не	... і сказаў: “Якое абвінавачваньне вы выстаўляеце адносна ...
абвясьціў	Так Не	... гэты пайшоў і абвясьціў Юдэям, што гэта ...
абнаўленьня	Так Не	... ў Ерусаліме сьвята абнаўленьня, і была зіма. ...
Абрагам	Так Не	... Яму: “Бацька наш – Абрагам”. Кажа ім Ісус: “ ...
Абрагама	Так Не	... адказалі: “Мы – насеньне Абрагама і нікому не ...
абразаньне	Так Не	... Майсей даў вам абразаньне; хоць яно і ...
аддасьць	Так Не	... тую, калі хто аддасьць душу сваю за ...
адзеньне	Так Не	... скідвае з Сябе адзеньне і, узяўшы ручнік, ...
Адпачатны	Так Не	... сказаў ім Ісус: “Адпачатны, як і кажу ...
адпусьціце	Так Не	... 23. Каму адпусьціце грахі, таму будуць ...
адпусьціць	Так Не	... і маю ўладу адпусьціць Цябе?” ...
адпусьціш	Так Не	... закрычалі, кажучы: “Калі адпусьціш Яго, ты ня ...
адыйду	Так Не	... калі Я не адыйду, Абаронца ня прыйдзе ...
адыйсьці	Так Не	... шукаеце, дазвольце ім адыйсьці, ...
адыйшла	Так Не	... гляк свой і адыйшла ў горад, і ...
адыйшлі	Так Не	... Бо вучні Ягоныя адыйшлі ў горад набыць ...
адыйшоўшы	Так Не	... сказаў Ісус і, адыйшоўшы, схаваўся ад іх. ...
адыйшоў	Так Не	... 1. Пасьля гэтага адыйшоў Ісус на другі ...
аздароўленаму	Так Не	... гэтага Юдэі казалі аздароўленаму: “Сёньня субота, ня ...

Мал. 4. Вынікі працы сэрвісу «Праверка правапісу»

Чацвёртым этапам вычыткі тэкстаў пры дапамозе сэрвісаў www.Corpus.by можа быць складанне слоўніка-частотніка па тэксце. Гэту задачу можна вырашыць пры дапамозе сэрвісу «Частотнасць слоў» [5]. Пасля апрацоўкі тэксту карыстальнік можа пабачыць усе ўжытыя ў ім словы і частату іх ужывання (карыстальнік можа вызначыць колькасць кантэкстаў, якую выведзе сэрвіс насупраць кожнага слова). Гэты сэрвіс дапаможа паскорыць стылістычную вычытку, так як карэктар будзе бачыць ці ў правільным значэнні выкарыстоўваецца тое ці іншае слова. Карыстальнік можа таксама скарыстацца полем «Шукаць толькі наступныя словы». Сэрвіс «Частотнасць слоў» накіраваны на вырашэнне многіх лінгвістычных задач і можа працаваць з любой мовай, дакладней з любой сістэмай сімвалаў. На малюнку 5 прадстаўлена пачатковая старонка сэрвісу.



Мал. 5. Знешні інтэрфейс сэрвісу «Частотнасць слоў»

Для зняцця шматзначнасці слоў у тэкстах можа быць выкарыстаны сэрвіс «Ідэнтыфікатар амографаў» [6], які паслоўна праглядае тэкст на наяўнасць некалькіх спосабаў прачытання таго ці іншага слова паводле інфармацыі, змешчанай у слоўніках. Прыклад працы сэрвісу прадстаўлены на малюнку 6.

Такім чынам сэрвісы Інтэрнэт-платформы www.Corpus.by могуць вырашаць задачы, якія патрабуюць пасімвальнай, паслоўнай і слоўнікавай апрацоўкі тэкстаў.

Варта адзначыць, што гэтым інтэрнэт-рэсурс не абмяжоўваецца і таксама вырашае задачы, якія паўстаюць пры распрацоўцы праграм для распазнавання і сінтэзу маўлення:

– пераўтварэнне арфаграфічнага тэксту ў фанетычную транскрыпцыю;
– сінтэз маўлення па тэксце;
– запіс маўлення ў гукавыя файлы;
– зручны анлайн доступ да лінгвістычных рэсурсаў.

Амограф	Варыянты націску	Тып амографа	↓ Кольк.	Кантэксты	Слоўнік
крыві	крыві́ кры́ві	адна парадыгма	1	… што не ад крыві, не ад жаданьня …	SBM1987
судзіцца	судзі́цца су́дзіцца	адна парадыгма	1	… у Яго, ня судзіцца, а хто ня …	SBM1987
часам	ча́сам часа́м	адна парадыгма	1	… да Яго. 31. Тым часам вучні прасілі Яго, …	SBM1987
слугі	слугі́ слу́гі	адна парадыгма	1	… дамоў, перанялі яго слугі ягоныя і паведамілі, …	SBM1987
травы	травы́ тра́вы	адна парадыгма	1	… тым месцы шмат травы. Тады узьлеглі мужчыны …	SBM1987
чынам	чы́нам чына́м	адна парадыгма	1	… якія ўзьляглі, падобным чынам і рыбы, колькі …	SBM1987
загубіць	загубі́ць заѓубіць	адна парадыгма	1	… даў, нічога не загубіць, але ўваскрасіць усё …	SBM1987
адлучыць	адлучы́ць адлу́чыць	адна парадыгма	1	… Яго за Хрыста, адлучыць ад сынагогі. 23. Дзеля …	SBM1987
сілу	сі́лу сілу́	адна часціна мовы	1	… дало Яно ім сілу дзецьмі Божымі стацца, …	SBM1987
духам	ду́хам духа́м	адна часціна мовы	1	… Той, Які хрысьціць Духам Сьвятым. 34. І я …	SBM1987
падае	пада́е па́дае	адна часціна мовы	1	… Кожны чалавек перш падае добрае віно, а …	SBM1987
рассыпаў	рассыпа́ў рассы́паў	адна часціна мовы	1	… і грошы мянялам рассыпаў, і сталы іх …	SBM1987
пакланяліся	пакла́няліся пакланя́ліся	адна часціна мовы	1	… прарок. 20. Бацькі нашыя пакланяліся на гэтай гары, …	SBM1987

Мал. 6. Вынікі працы сэрвісу «Ідэнтыфікатар амографаў»

У заключэнне лічым вартым падсумаваць перавагі сэрвісаў Інтэрнэт-платформы www.Corpus.by. Яны зніжаюць верагоднасць з'яўлення звычайных абдруковак у тэксце, значна скарачаюць час працы карэктара/рэдактара, дапамагаюць уніфікаванню выкарыстоўваемых у тэксце сімвалаў, стылістычнаму рэдагаванню лексікі, працуюць з вялікімі аб'ёмамі тэкстаў. Сэрвісы знаходзяцца ў пастаяннай распрацоўцы, і карыстальнікі маюць магчымасць напісаць пра памылку ці недахоп таго ці іншага сэрвісу. Дзякуючы гэтаму, пастаянна ўдасканальваюцца алгарытмы і лінгвістычныя рэсурсы [7].

ЛІТАРАТУРА

1. Corpus.by [Электронны рэсурс]. — 2016. — Рэжым доступу: http://www.Corpus.by/. — Дата доступу: 15.03.2016.

2. Інфармацыя аб сімвалах[Электронны рэсурс]. — 2016. Рэжым доступу: http://www.Corpus.by/ InformationOnCharacters/. — Дата доступу: 15.03.2016.

3. Праверка правапісу «ў» [Электронны рэсурс]. — 2016. Рэжым доступу: http://www.Corpus.by/ShortUSpellChecker/. — Дата доступу: 15.03.2016.

4. Праверка правапісу [Электронны рэсурс]. — 2016. Рэжым доступу: http://www.Corpus.by/spellChecker/. — Дата доступу: 15.03.2016.

5. Частотнасць слоў [Электронны рэсурс]. — 2016. Рэжым доступу: http://www.Corpus.by/wordFrequency/. — Дата доступу: 15.03.2016.

6. Ідэнтыфікатар амографаў [Электронны рэсурс]. — 2016. Рэжым доступу: http://www.Corpus.by/HomographIdentifier/. — Дата доступу: 15.03.2016.

7. Гецэвіч, Ю.С., Лысы, С.І. Праектаванне інтэрнэт-сервісаў для працэсараў сінтэзатара маўлення па тэксце з магчымасцю прадстаўлення бясплатных электронных паслуг насельніцтву // Развитие информатизации и государственной системы научно-технической информации (РИНТИ–2014): доклады XIII Международной конферен–ции. — Минск: ОИПИ НАН Беларуси, 2014. — С. 265–269.