КОНТРАСТИВНЫЕ ИССЛЕДОВАНИЯ и ПРИКЛАДНАЯ ЛИНГВИСТИКА

Материалы
Международной научной конференции

г. Минск, 29–30 октября 2014 г.

В двух частях
Часть вторая

СОДЕРЖАНИЕ

ТЕХНОЛОГИИ ИЗВЛЕЧЕНИЯ ИНФОРМАЦИИ ИЗ ИНТЕРНЕТА И ЕЕ ОБРАБОТКА

АВТОМАТИЧЕСКИЙ АНАЛИЗ И СИНТЕЗ ТЕКСТОВ

КОРПУСНЫЕ ИССЛЕДОВАНИЯ

МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
Минский государственный лингвистический университет

КОНТРАСТИВНЫЕ ИССЛЕДОВАНИЯ И ПРИКЛАДНАЯ ЛИНГВИСТИКА

МАТЕРИАЛЫ
Международной научной конференции

г. Минск, 29–30 октября 2014 г.

В двух частях
Часть вторая

Минск 2015

УДК 81'42 + 81'33
ББК 81.04 + 81.1
К65

Рекомендованы Редакционным советом Минского государственного лингвистического университета. Протокол № 3 (19) от 23.12.2014 г.

Рецензенты: доктор филологических наук, профессор Д. Г. Богушевич (МГЛУ); кандидат филологических наук, доцент А. И. Головня (БГУ)

К65 **Контрастивные** исследования и прикладная лингвистика : материалы Междунар. науч. конф., Минск, 29–30 окт. 2014 г. В 2 ч. Ч. 2 / отв. ред.: А. В. Зубов, Т. П. Карпилович. – Минск : МГЛУ, 2015. – 227 с.

ISBN 978-985-460-669-9 (Ч. 2).
ISBN 978-985-460-658-3.

В издание, состоящее из двух частей, включены материалы, раскрывающие возможности контрастивного изучения самых различных языков: белорусского, русского, английского, немецкого, башкирского, туркменского, итальянского, китайского, грузинского, казахского, польского. Показано как результаты такого изучения используются при обучении, автоматической обработке текстов, их переводе.

Данная вторая часть книги содержит тексты выступлений участников конференции в секциях «Технологии извлечения информации из Интернета и ее обработка», «Автоматический анализ и синтез текстов», «Корпусные исследования», «Контрастивная лингвистика и вопросы преподавания языков» и «Особенности речевого поведения в разных культурах».

Для научных работников, аспирантов и магистрантов.

УДК 81'42 + 81'33
ББК 81.04 + 81.1

ISBN 978-985-460-669-9 (Ч. 2)
ISBN 978-985-460-658-3

4. What is BLARK? [Electronic resource] // Рэжым звароту : http://www.blark.org/. – Дата звароту : 04.09.2014.
5. Global WordNet Association [Electronic resource] // Рэжым звароту : http://www.globalwordnet.org. – Дата звароту : 04.09.2014.
6. EuroWordNet [Electronic resource] // Рэжым звароту : http://www.illc.uva.nl/EuroWordNet/. – Дата звароту : 04.09.2014.
7. *Vossen, P.* Wordnet-LMF : a standard representation for multilingual wordnets, in LMF Lexical Markup Framework, ed. by Gil Francopoulo ISTE / P. Vossen, C. Soria, M. Monachini. – Wiley, 2013.
8. *Ponzetto, S. P.* Knowledge-rich Word Sense Disambiguation rivaling supervised systems / S. P. Ponzetto, R. Navigli // Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL), 2010. – P. 1 522–1 531.
9. *Кошчанка, У.* Актуальны стан і перспектывы развіцця корпуснай лінгвістыкі і камп'ютэрнай лексікаграфіі ў Інстытуце мовы і літаратуры НАН Беларусі / У. Кошчанка // Беларуская мова ў культурнай і моўнай прасторы Славіі : матэрыялы Міжнар. навук. канф., Мінск, 24–25 лістап. 2009 г. – Мінск : Права і эканоміка, 2009.

С. І. Лысы, Ю. С. Гецэвіч (Мінск)

ЧАСТОТНЫ АНАЛІЗ ЭЛЕКТРОННАГА ТЭКСТУ НА ПРАДМЕТ ВЫКАРЫСТАННЯ СЛОЎ І ІНШЫХ СІМВАЛЬНЫХ ПАСЛЯДОЎНАСЦЯЎ З ДАПАМОГАЙ WWW.CORPUS.BY

Задачы, звязаныя са статыстычным аналізам электронных тэкстаў, з'яўляюцца актуальнымі, а праграмныя сродкі для іх вырашэння – запатрабаванымі многімі галінамі навукі і ў прыватнасці інгвістыкай. У сувязі з гэтым быў распрацаваны даступны ўсім інтэрнэт-сэрвіс «Частотнасць слоў» (www.corpus.by/wordFrequency) [1]. Ён спрыяе атрыманню статыстыкі вакарыстанні электронным тэксце слоў альбо канкрэтных сімвальных паслядоўнасцяў, згодна з наладкамі карыстальніка.

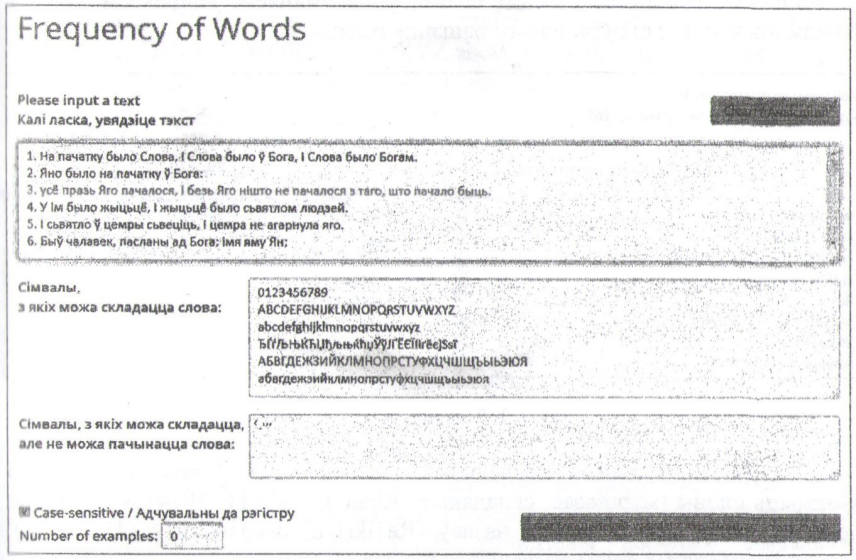
Для гэтага неабходна увесці любы тэкст і задаць (скарэктаваць) два мноства сімвалаў. Першае – сімвалы, з якіх можа складацца слова. Сюды можна змясціць алфавіт альбо сімвалы, якія будуць выкарыстоўвацца для распазнавання «слоў» у тэксце. Па замоўчванні ў гэтым вакне змешчаны ўсе літарныя сімвалы кадзіравання «Windows-1251» [2] і лічбаў:

0123456789
ABCDEFGHIJKLMNOPQRSTUVWXYZ
abcdefghijklmnopqrstuvwxyz
ЂЃґЉЊЌЋЏѓѐєїІіґёєjSsї
АБВГДЕЖЗИЙКЛМНОПРСТУФХЦЧШЩЪЫЬЭЮЯ
Абвгдежзийклмнопрстуфхцчшщъыьэюя.

Другое мноства – сімвалы, якія могуць складаць, але ніколі не пачынаюць слова. Па замоўчванні гэта – асноўны і частковы націск, апостраф

і злучок. Акрамя гэтага, можна пазначыць, ці трэба пры падліку статыстыкі ўлічваць рэгістр сімвалаў, а таксама вызначыць колькасць кантэкстаў знойдзеных «слоў», якія будуць выведзеныя побач з імі. Знешні інтэрфейс сэрвісу прадстаўлены на мал. 1.

Націснуўшы «Get frequency of Words!» ('Атрымаць частату слоў'), карыстальнік атрымлівае спіс слоў (альбо паслядоўнасці сімвалаў), упарадкаваны па частаце альбо паводле алфавіта, з пазначэннем колькасці адзінак ва ўваходным тэксце:



Мал. 1. Знешні інтэрфейс інтэрнэт-сэрвісу «Частотнасць слоў»

Такім чынам, гэты сэрвіс можа выкарыстоўвацца як непасрэдна для атрымання статыстычных дадзеных пра словы ў тэксце, так і для вырашэння іншых больш спецыфічных задач (падліку частотнасці розных лічбавых значэнняў у тэксце, а таксама знакаў прыпынку і іх спалучэнняў і г.д.).

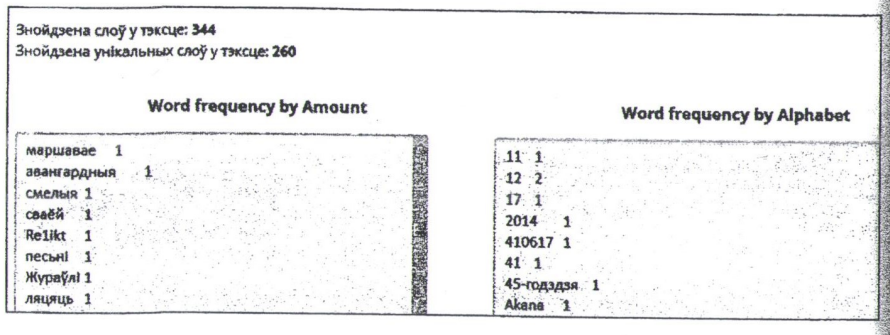
Разгледзім канкрэтныя прыклады выкарыстання сэрвісу. На ўваход пададзена кніга «Новы Запавет» (у перакладзе Васіля Сёмухі) з наладкамі па замоўчванні [3]. У выніку маем наступную табліцу:

Спіс па колькасці (рэгістранезалежны)	Спіс па алфавіту (рэгістранезалежны)	Спіс па колькасці (рэгістразалежны)	Спіс па алфавіту (рэгістразалежны)
і 9155	а 2033	і 7899	а 1284
ў 2536	аарон 1	ў 2536	Аарон 1
што 2099	ааронава 1	што 2066	Ааронава 1
а 2033	ааронавага 1	ня 1593	Ааронавага 1
…	…	…	…

Праца з тэкстам, дзе сустракаюцца розныя алфавіты, лічбы і іх камбінацыі, выглядае так. Адкрываем тэкст па спасылцы http://news.tut.by/culture/410617.html. Адразу ж бачым, што ў ім выкарыстоўваюцца як беларускія (дыек, рэалізаваць, праект), рускія (комментариев), англійскія (Detroit, Naka, Tonqixod) словы, так і лікі (11, 17), змешаныя адзінкі (Re1ikt, 17:41) і іншае (Re:1 Песняры). Прыклады настроек сэрвісу, каб атрымаць рашэнне некаторых камп'ютарна-лінгвістычных задач, могуць выглядаць так.

Задача 1. Знайсці ўсе словы тэксту і падлічыць іх частату выкарыстання.

Рашэнне. Вокны са спісамі сімвалаў не змяняем. Націскаем кнопку «Атрымаць частату слоў!». Бачым рашэнне на мал. 2.



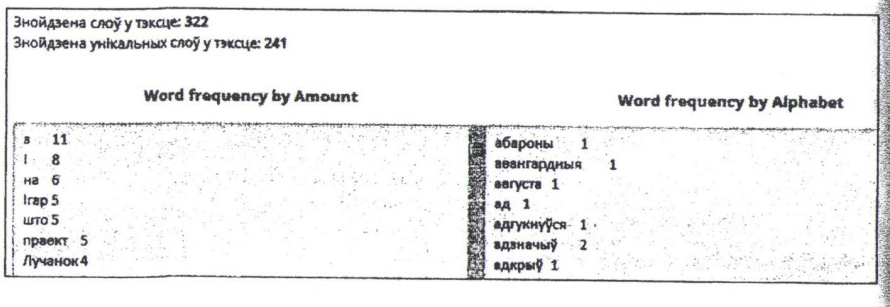
Мал. 2. Рашэнне задачы 1

У атрыманых спісах сімвальных паслядоўнасцяў даследчык можа заўважыць словы (маршавае, складаных, Akana), лікі (11, 2014), камбінацыі словы і лічбаў для стварэння назваў (Re1ikt) ці скарачэнняў (45-годдзя). Прааналізуем іншыя спісы сімвальных паслядоўнасцяў.

Задача 2. Знайсці ўсе словы тэксту, запісаныя кірыліцай, падлічыць частату іх ужывання.

Рашэнне. У вакне з алфавітам пакідаем толькі кірылічныя літары:

АБВГДЕЖЗИЙКЛМНОПРСТУФХЦЧШЩЪЫЬЭЮЯ
АбвгдежзийклмнопрстуфхцчшщъыьэюяЎўЁёІі.



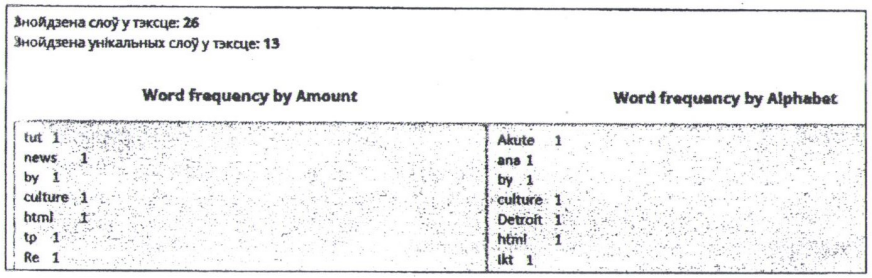
Мал. 3. Рашэнне задачы 2

Сэрвіс адбірае ўсе словы, якія складаюцца цалкам з кірылічных літар (што, Ігар). Але ў выпадку камбінацыі літар і лічбаў апошнія не ўлічваюцца (напрыклад, у слове *45-годдзя* знойдзена сімвальная паслядоўнасць *годдзя*). Таму патрэбна уважліва інтэрпрэтаваць рэзультаты. Аўтары распрацоўкі дапускаюць, што даследчыку спатрэбіцца праглядаць вынікі работы сэрвісу і некаторыя выразы, якія не адпавядаюць пастаўленай задачы будуць выкрэслены.

Задача 3. Знайсці ўсе словы тэксту, запісаныя лацінскімі літарамі, падлічыць іх частату ўжывання.

Рашэнне. У вакне з алфавітам пакідаем толькі лацінскія літары: ABCDEFGHIJKLMNOPQRSTUVWXYZ abcdefghijklmnopqrstuvwxyz.

Актывізіруем сэрвіс. У выніку маем (мал. 4):



Мал. 4. Рашэнне задачы 3

У выніку знойдзены ўсе словы, што цалкам складаюцца з лацінскіх літар, але ў выпадку камбінаванага напісання адзінка «разрываецца» на часткі (напрыклад, у назве *Re1ikt* знойдзены асобныя сімвальныя паслядоўнасці *Re* і *ikt*). Таму такія словы трэба апрацоўваць асобна.

У бліжэйшай будучыні аўтарамі сэрвісу ставіцца задача дадаць розных мностваў сімвалаў па замоўчванні для шэрагу моў, павялічыць аб'ём магчымай апрацоўкі тэксту. Таксама абмяркоўваюцца пытанні па дапрацоўцы сэрвісу www.corpus.by/wordFrequency з мэтай выкарыстання яго для падліку частотнасці слоў канкрэтнай мовы. Напрыклад, карыстальнік зможа аналізаваць тэксты на грэчаскай мове, змясціўшы яе алфавіт у першае мноства, а спіс дыякрытычных сімвалаў – у другое.

ЛІТАРАТУРА

1. Frequency of Words [Electronic resource] // www.Corpus.by web-site. – 2014. – Mode of access: http://www.corpus.by/wordFrequency/. – Date of access: 28.08.2014.
2. Go Global Developer Center [Electronic resource] // Windows 1251. – 2014. – Mode of access : http://msdn.microsoft.com/en-us/goglobal/cc305144.aspx. – Date of access : 28.08.2014.
3. Біблія [Электронны рэсурс] // Беларуская Палічка. – 2014. – Рэжым звароту : http://knihi.com/none/Biblija.html. – Дата звароту : 28.08.201