УНИВЕРЗИТЕТ „ГОЦЕ ДЕЛЧЕВ” – ШТИП
ФАКУЛТЕТ ЗА ИНФОРМАТИКА

Катедра по информациски технологии

Стојанче Спасов

ВЕБ СЕРВИС ЗА ПОВЕЌЕЗНАЧНА ТРАНСЛИТЕРАЦИЈА НА ЦЕЛИ РЕЧЕНИЦИ ОД ЛАТИНИЦА ВО КИРИЛИЦА

- МАГИСТЕРСКИ ТРУД -

Штип, септември 2014

Комисија за оценка и одбрана

Ментор: **доц. д-р Зоран Здравев**
Факултет за информатика
Универзитет „Гоце Делчев“ – Штип

Член: **доц. д-р Игор Стојановиќ**
Факултет за информатика
Универзитет „Гоце Делчев“ – Штип

Член: **доц. д-р Александра Милева**
Факултет за информатика
Универзитет „Гоце Делчев“ – Штип

Членови на комисија за оценка и одбрана

Претседател: **доц. д-р Игор Стојановиќ**
Факултет за информатика
Универзитет „Гоце Делчев“ – Штип

Член: **доц. д-р Зоран Здравев**
Факултет за информатика
Универзитет „Гоце Делчев“ – Штип

Член: **доц. д-р Александра Милева**
Факултет за информатика
Универзитет „Гоце Делчев“ – Штип

Научно поле: Информатика
Научна област: Информациони системи и технологии

Датум на одбрана: _18.09.2014_

Листа на рецензирани и објавени трудови произлезени од истражувањето:

1. Spasov, S., & Zdravev, Z. (2013). Web service for ambiguous transliteration of full sentences from Latin to Cyrillic alphabet. Yearbook of the Faculty of Computer Science, 1(1), 252-263. http://eprints.ugd.edu.mk/8135/

Листа на останати трудови:

2. Stefanova, S., Spasov, S., & Zdravev, Z. (2012). Data Warehouse Design for Climate Change Prediction in Republic of Macedonia. http://eprints.ugd.edu.mk/681/

3. Stefanova, S., Spasov, S., & Zdravev, Z. (2014). Scorm vs Common Cartridge – Case study at University Goce Delcev.

Краток извадок

Постојат многу веб базирани апликации и социјални мрежи каде што се објавени македонски содржини напишани на латиница и покрај можноста за користење на кирилско писмо, што доведува до појава на двосмисленост на некои зборови во самите реченици. Тоа се должи на интернет корисниците кои од навика или тешкотија користат латинско писмо за пишување на содржини (на пр. оглас, коментари итн.), или пак користат електронски уреди кои немаат опција за комуникација преку поддршка за кирилско писмо. За надминување на наведениот проблем се користи процесот на транслитерација.

Низ овој процес се трансформираат сите зборови од латиница во кирилица, и како резултат само некои од нив може да добијат повеќе решенија т.е. повеќе значења на зборот. На пример од „postar" се добива „постар", „поштар" или пак, од „teza" се добива „тежа", „теза".

Во магистерскиот труд се дефинирани и презентирани основните концепти на повеќезначна транслитерација која претставува клучен чекор во решавање на значењето и смислата на зборовите во целите реченициte. Воедно се опишани трите типа на речници: отворен речник или ОР (преземен од OpenOffice Dictionaries), толковен речник или ТР (преземен од дигитален речник на македонскиот јазик), комбиниран речник или КР (комбинација на ОР и ТР), од кои се направени повеќе анализи во текот на истражувањата. Од аспект на решавање на повеќезначната транслитерација, дефиниран е алгоритам кој користи условна веројатност и Бајесови формули. Решението е имплементирано како нов алгоритам за транслитерација на различни типови на содржини. Во текот на истражувањата е применето автоматско тестирање на 300 одбрани реченици поделени на три групи.

За практична примена на развиениот алгоритам, како посебен дел во овој магистерски труд е дизајниран веб сервис кој преку кориснички интерфејс и модул за опслужување може да транслитерира: електронски пораки напишани на латиница, веб публикувања, постари текстови итн.

Во последниот дел од трудот е даден подетален опис на веб сервисот со примена на архитектурата (REST) и нејзина компарација со PHP и MySQL технологиите.

Клучни зборови: условна веројатност, Бајесови формули, алгоритам, кориснички интерфејс, модул за опслужување, архитектурата (REST).

Abstract

There are many web-based applications and social networks with published Macedonian contents written in Latin alphabet, despite the opportunity for using Cyrillic alphabet, which questions the existence of ambiguity of certain words in the sentences. This is due to the Internet users who use Latin alphabet, as a custom or difficulty, for writing contents (for example advertisements, comments etc.) or use electronic devices without option for Cyrillic alphabet as an input language. For overcoming this issue, the process of transliteration is used.

This process transforms all words from Latin to Cyrillic alphabet and as a result, only a few of them will have more solutions i.e. different meanings of the word. For example, from „postar" we obtain „постар", „пoштар" or from teza we obtain „тежа", „теза".

This master paper defines and presents the basic concepts of ambiguous transliteration, which is the key step to solving the meaning and purpose of the words in the full sentences. In the same time, it describes the three types of dictionaries: open dictionary or OD (downloaded from OpenOffice Dictionaries), expository dictionary or ED (downloaded from digital dictionary of Macedonian language), combined dictionary or CD (combination of OD and ED), used to make many analyses during the research. From the point of view of ambiguous transliteration solution, an algorithm is defined, which uses conditional probability and Bayes' formulas. The decision is implemented as a new transliteration algorithm in different types of contents. During the research, an automatic testing of 300 selected sentences, divided to three groups, is performed.

For practical application of the developed algorithm, as a separate part of this master paper, a web service is designed, which can make transliteration of electronic messages written on Latin alphabet, web publications, older texts etc. through user interface and service module. In the last part of the paper, a more detailed description of the web service by application of architecture (REST) and its comparison to PHP and MySQL technologies is given.

Key words: conditional probability, Bayes' formulas, algorithm, user interface, service module, architecture (REST).

СОДРЖИНА

Табели

Слики

1 ВОВЕД

Современите информатички технологии имаат примена во различни области. Во денешно време, Интернетот сè повеќе нуди можност за пристап до голем број на информации за широка примена во различни деловни сегменти. Сите информации мора да бидат складирани, најчесто физички, но како инфраструктурата на глобалната мрежа секојдневно се проширува, податоците сè почесто се чуваат во електронска форма со помош на бази на податоци. Поголем дел од податоците се презентираат преку многубројните веб апликации и социјални мрежи, во кои се наоѓаат форми за внесување на различни информации (на пр. оглас, коментари итн.) или пак, преку мобилен телефон за испраќање на електронски пораки. Тие податоци во нашиот случај се однесуваат на македонски содржини кои наместо со кирилско писмо се напишани на латиница.

Латинската презентација на зборовите понекогаш доведува до појава на двосмисленост во самите реченици. Ова е битен фактор кој го ограничува и го намалува разбирањето на повеќезначните поими во цели реченици, но и создава појава на нов феномен транслитерација за презентирање на дадени писма.

Примарната цел е да се идентификуваат случаите со повеќе значења и да се најде решение за нив. Секундарна цел е креирање на веб базиран сервис т.е. дефинирање на интелигентен алгоритам кој ќе овозможи транслитерација на цели реченици од латиница во кирилица и кој ќе може да врши транслитерација на барање од други апликации.

Овој сервис се очекува да има големо значење за македонското писмо, особено за поимите напишани на латиница. Исто така, сервисот може да се применува во различни области, каде сетовите на податоци се составени од македонски содржини напишани на латиница.

Магистерскиот труд е опишан во повеќе делови и тоа: Цел на истражување, Транслитерација, Методи на истражување, Истражување со примена на

транслитерација, Резултати од истражувањата, Функционални особини на алгоритамот за транслитерација, Веб сервиси.

Во поглавјето (Цел на истражување) се дефинирани три главни цели кои треба да се постигнат во истражувањето, а за секое од нив е даден краток опис за реализација. Насловите на главните цели се: Креирање на база на податоци преку аквизиција и жнеење на зборови од интернет извори; Дефинирање на алгоритам за повеќезначно транслитерирање на цели реченици од латиница во кирилица; Дефинирање на веб сервис за користење на транслитерација на цели реченици.

Во поглавјето (Транслитерација) е даден опис на проблемот кој се појавува т.е. потребата од транслитерација и веб сервисот. Потоа е дадена детална дефиниција на процесот транслитерација и еден мал пример за појаснување. Исто така следува и опис на стандарди на транслитерацијата со помош на две табели во одредени услови.

Во поглавјето (Методи на истражување**)** се опишани детално неколкуте методи кои се користат при креирањето на веб сервисот за транслитерација. Овде спаѓаат: пресметување на коефициент; пребарување на зборови во база на податоци; користење на тест скриптите; коригирање на грешка за настанатиот проблем; користење на кориснички интерфејс. Овде е дадена и шема на која се претставени сите методи.

Во поглавјето (Истражување со примена на транслитерација) се наведени сите истражувања кои досега се направени. Во нив спаѓаат: пребарување, споредување и комбинирање на ОР и ТР. За секое истражување е добиен соодветен резултат, или тоа се однесува на резултат на зборови кои добиваат повеќе од едно значење, при транслитерација.

Потоа е опишано автоматското читање и статистичката обработка на целите реченици. Добиените резултати овде се однесуваат на: вкупно двозначни зборови; вкупно реченици прочитани во сите PDF документи; вкупно зборови прочитани во сите PDF документи; вкупно реченици со повеќезначни зборови; вкупно поими (лево, десно) од двозначните зборови; вкупен број на повторување на сите поими (лево, десно) од двозначните зборови.

На крајот од нас се дефинирани и одредени формули за добивање на коефициентот за решавање на повеќезначна транслитерација.

Поглавјето (Резултати од истражувањата) е најинтересно, бидејќи се тестираат три типа на реченици со помош на три типа на формули. Резултатите кои се добиени се со позитивни и негативни вредности. Анализата која е направена во овој дел се однесува на споредување на негативните реченици од претходно добиените негативни резултати. Исто така е даден опис на алгоритамот кој е тестиран за одреден број на содржини во даден временски интервал. Резултатот од тестираниот алгоритам е прикажан со помош на дијаграми.

Во поглавјето (Функционални особини на алгоритамот за транслитерација) е даден детален опис на алгоритамот за транслитерација. Делот за комбинирани букви од македонската азбука е опишан со максимално комбинирање на букви за одреден двозначен поим напишан на латиница.

Потоа е даден детален опис на делот за добивање на транслитерирани цели реченици од латиница во кирилица во кои се одредува значењето на зборовите. Тестиран е и алгоритамот за транслитерација на неколку реченици, со позиционирање на зборови преку база на податоци.

Функција за ограничување на зборовите е делот каде се опишуваат делови од програмскиот код на алгоритамот. На крај е претставен ER дијаграмот со детален опис за секоја табела и атрибут во базата на податоци.

Веб сервиси е последното поглавје од магистерскиот труд и е најпрактично за крајните корисници на веб сервисот за транслитерација. На почеток е дадена кратка дефиниција за веб сервиси. Потоа е дефинирана архитектурата REST која претставува начин на комуникација помеѓу клиентите и серверот на нашиот веб сервис со помош на HTTP протокол.

Корисничкиот интерфејс и модулот за транслитерација на веб сервисот се исто така дел од ова поглавје, каде се дава детален опис со помош на текст и прикази од онлајн поставениот сервис.

Приказите се однесуваат на: формата за внесување на реченици за транслитерација, внесена реченица како барање за транслитерација, краен резултат на транслитерирана реченица, програмски код на модулот.

Одовде произлегуваат истражувачките прашања кои треба да се одговорат:

1. За кои типови на податоци се однесува транслитерацијата?
2. Каков алгоритам треба да се креира и која ќе биде неговата функционалност?
3. На што ќе се базираат крајните резултати?

Резултатите на истражувањата и одговорите на овие прашања се организирани во наредните поглавја со примери од анализа и тестирање, прикажани во табели, дијаграми и алгоритамски кодови.

2 ЦЕЛ НА ИСТРАЖУВАЊЕ

Главните цели кои треба да се постигнат во истражувањето се:

1. Креирање на база на податоци преку аквизиција и жнеење на зборови од интернет извори (PDF, TXT, HTML, PHP, итн.);

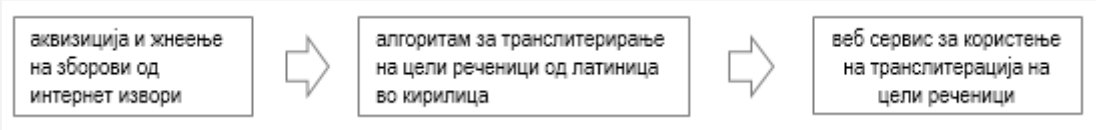
    - Ова се реализира мануелно кога се преземаат документи во различен формат, а автоматски кога се внесуваат зборови во база на податоци.

2. Дефинирање на алгоритам за повеќезначно транслитерирање на цели реченици од латиница во кирилица (PHP скрипти);

    - За да се дефинира алгоритамот потребно е претходна анализа на типови на податоци, како и користење на формули за добивање на резултати, при транслитерација.

3. Дефинирање на веб сервис за користење на транслитерација на цели реченици (PHP, MySQL, jQuery скрипти).

    - Поради различниот обем на содржини за транслитерирање, може да се користи веб сервисот преку кориснички интерфејс (за помал број на зборови) и преку вграден модул за транслитерација во други апликации (за поголем број на зборови).



Слика 1. Цели на истражување (содржини, алгоритам, веб сервис)
Figure 1. Research goals (contents, algorithms, web service)

3 ТРАНСЛИТЕРАЦИЈА

Пребарувачите преку Веб можат да прикажат многу информации, вклучувајќи и македонски информации напишани на латиница. Таквите информации се појавуваат, бидејќи корисниците кога пишуваат некоја содржина од тастатура користат латинско писмо, наместо кирилско. Тоа се должи најчесто од навика или тешкотија, со замисла дека на тастатурата не може брзо да се препознаат копчињата со кои наместо буквите (ш ѓ ж ч ќ љ џ s њ) треба да се употребат ознаките ([ ] \ ; ’ q x y w). Исто така се прифаќа фактот дека некои странски уреди при користење на различни технологии немаат опција за комуникација преку кирилска поддршка и затоа мора да се пишува или комуницира на латиница. Меѓутоа, за да бидат содржините подобро разбрани од страна на локалните читатели (Hsu & Chen, 2010), треба често да се мапираат од латиница во кирилица. Како начин за решавање на овој проблем се користи терминот транслитерација.

3.1 Дефиниција на транслитерација

Транслитерација е процес кој се користи за мапирање на зборови од едно писмо во друго (Chinnakotla, Damani, & Satoskar, 2010). Тоа е постапка која има многу широка примена за македонското писмо кое користи кирилски букви, но во одредени услови се појавува потреба буквите да се пишуваат на латиница. Со помош на современите техники и технологии, се овозможува процесот на транслитерација да се реализира полесно и побрзо во стандардни услови. Но во нестандардни услови, кои се многу чести за македонски услови, заедно со граматичките правила, настанува неразбирливост, особено во делот на смислата и повеќезначноста на зборовите. На пример од „kuka“ се добива „кука“ или „куќа“, од „sok“ се добива „сок“ или „шок“ и многу други.

3.2 Стандарди на транслитерација

Зборовите во било кој јазик понекогаш треба да се напишат во друго писмо. Често тоа се случува и во македонски услови[1] кога кирилски зборови се пишуваат на латиница. Тоа е така бидејќи уредите при комуникација немаат кирилска поддршка. Типичен пример е кога се праќа порака од мобилен телефон

[1] Начин на пишување на содржини од страна на корисниците

која е напишана на латиница. Транслитерацијата може да биде реверзибилна и да конвертира поими од едно писмо во друго. Транслитерацијата не е секогаш едноставен процес за реализација, поради тоа што не се прави разлика при пишувањето на македонско писмо со латински букви (Spasov & Zdravev, 2013). Затоа постојат правила според некои стандарди. Македонската транслитерација е стандардизирана со ISO R9:1968. Овој систем во 1970 година е адаптиран и усвоен од страна на Македонската академија на науките и уметностите и се смета за официјално прифатен во Република Македонија. Досега се направени алгоритами за транслитерација кои перфектно работат во стандардни услови. Под стандардни услови подразбираме кога пишувањето на текстот е според утврдени стандарди, на пример „ѓ=gj“, „ж=zh“, „s=dz“, „њ=nj“, „ќ=kj“, „ч=ch“, „џ=dj“ и „ш=sh“. Ова е прикажано во *Табела 1.*

Табела 1. Транслитерација од кирилица во латиница во стандардни услови
Table 1. Transliteration from Cyrillic to Latin alphabet under standard circumstances

Кирилица	*Латиница*	*Процес на транслитерација*	
А а	A a	авантура	avantura
Б б	B b	борба	borba
В в	V v	вести	vesti
Г г	G g	град	grad
Д д	D d	дрво	drvo
Ѓ ѓ	Gj gj	ѓавол	gjavol
Е е	E e	елен	elen
Ж ж	Zh zh	жонглер	zhongler
З з	Z z	збор	zbor
S s	Dz dz	sид	dzid
И и	I i	имот	imot
J j	J j	јубилеј	jubilej
К к	K k	коска	koska
Л л	L l	леден	leden
Љ љ	Lj lj	љубичица	ljubichica
М м	M m	манастир	manastir
Н н	N n	нога	noga
Њ њ	Nj nj	коњ	konj
О о	O o	облека	obleka
П п	P p	песна	pesna
Р р	R r	разговор	razgovor
С с	S s	соба	soba
Т т	T t	торба	torba
Ќ ќ	Kj kj	ќумур	kjumur
У у	U u	умерено	umereno
Ф ф	F f	форма	forma

Табела 2. Повеќезначна транслитерација од кирилица во латиница во нестандардни услови
Table 2. Ambiguous Transliteration from Cyrillic to Latin alphabet under non-standard circumstances

Кирилица	*Латиница*	*Процес на транслитерација*	
А а	A a	авантура	avantura
Б б	B b	борба	borba
В в	V v	вазна	vazna
Г г	G g	град	grad
Д д	D d	дрво	drvo
Ѓ ѓ	G g	ѓавол	gavol
Е е	E e	елен	elen
Ж ж	Z z	жонглер	zongler
З з	Z z	збор	zbor
S s	Z z	sид	zid
И и	I i	имот	imot
J j	J j	јубилеј	jubilej
К к	K k	коска	koska
Л л	L l	леден	leden
Љ љ	L l	љубичица	lubicica
М м	M m	манастир	manastir
Н н	N n	нога	noga
Њ њ	Nj nj	коњ	konj
О о	O o	облека	obleka
П п	P p	песна	pesna
Р р	R r	разговор	razgovor
С с	S s	соба	soba
Т т	T t	торба	torba
Ќ ќ	K к	ќумур	kumur
У у	U u	умерено	umereno
Ф ф	F f	форма	forma

Х х	H h	хумор	humor
Ц ц	C c	цвеќе	cvekje
Ч ч	Ch ch	човек	chovek
Џ џ	Dj dj	џамија	djamija
Ш ш	Sh sh	шеќер	shekjer

Х х	H h	хумор	humor
Ц ц	C c	цвеќе	cveke
Ч ч	C c	човек	covek
Џ џ	J j	џамија	jamija
Ш ш	S s	шеќер	seker

Но во нестандардни услови, кои се многу чести за македонски услови, пишувањето букви на латиница може да имаат две значења во кирилица. Така ако имаме латинско „s“ тоа може да биде кирилско за „ш“ или „с“, или ако имаме напишано латинско „z“ тоа може да биде кирилско за „з“ или „ж“ итн. Со ова се појавува нејасност во смислата на реченициte. Ова е прикажано во *Табела 2.*

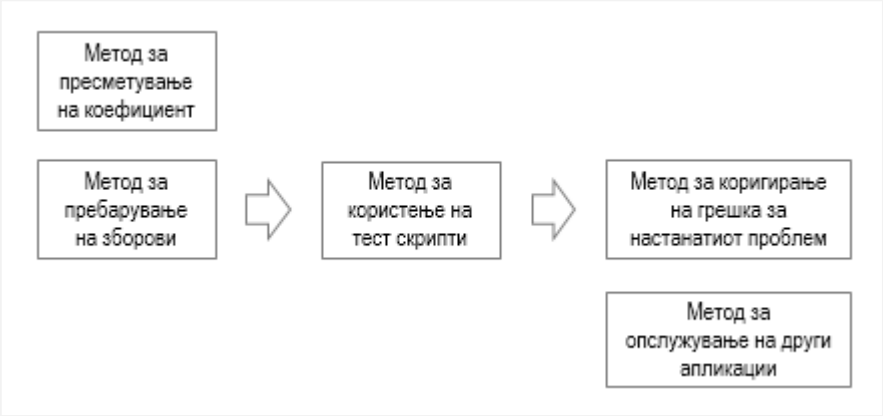
Така, многу често како резултат се добиваат зборови со повеќе значења. На пример, од „vesti“ се добива „вешти“ или „вести“, од „dokazi“ се добива „докази“ или „докажи“. Има уште многу други вакви примери. Овие случаи ги нарекуваме „транслитерација со повеќе значења“.

4 МЕТОДИ НА ИСТРАЖУВАЊЕ

Магистерскиот труд опфаќа неколку фази кои ќе се одвиваат според дадената листа на изведување:

- Анализа на јазични писма при транслитерација;
- Анализа на типови на податоци корисни за транслитерација;
- Анализа на граматички правила за реченициte од македонскиот јазик;
- Анализа на методите за транслитерација на цели реченици;
- Тестирање на претходно дефинирани алгоритами и скрипти;

За да се овозможи сето тоа применети се неколку методи и тоа за: пресметување на коефициент; пребарување на зборови во база на податоци; користење на тест скрипти; коригирање на грешка за настанатиот проблем; опслужување на други апликации, прикажани во редослед на Слика 2.



Слика 2. Редоследно прикажување на методите на истражување
Figure 2. Sequential overview of research methods

- Метод за пресметување на коефициент се применува со веќе направената скрипта (koeficient.php) за време на истражувањето, со користење на програма за пишување на програмски код и програма за комуникација со база на податоци. Коефициентот, како дел од главниот алгоритам за повеќезначна транслитерација, при решавање на двозначноста на зборовите користи вредност заокружена на шеста децимала со помош на PHP параметарот round кој може да прима негативни и нулти вредности.

- Метод за пребарување на зборови е најчестиот начин за процесирање на податоци во самиот алгоритам. Тој е овозможен преку зададени функции кои се повторуваат на одреден дел од програмскиот код. Во зависност од зборовите за пребарување се прави селектирање на атрибутите во табели преку упити за кои базата на податоци дава чисти резултати. Доколку базата на податоци дава нулта вредност за пребараниот збор, во тој случај тој останува во почетна состојба без транслитерација на неговите знаци.

- Метод за користење на тест скрипти (citaj_recenici.php, pdf_sodrzina.php и pecati.php) се однесува на рачно и автоматско тестирање, со следење на резултати преку цели и процентуални вредности. Со секој задоволителен излез од тест скриптата (citaj_recenici.php), се прави надградба на главниот алгоритам. Сите овие скрипти се креирани и тестирани во повеќе етапи, според алгоритамот за транслитерација. Скриптата (citaj_recenici.php) се стартува рачно и нејзиното временско работење ќе

зависи од: *конфигурацијата на базата, програмскиот код на скриптата, конфигурацијата на серверот и големината на зададената содржина*.

- Метод за коригирање на грешка за настанатиот проблем секогаш се применува при појава на грешка во крајниот резултат од некоја тест скрипта. Грешките во нашиот случај, најчесто се појавуваат кога речениците содржат: специјални знаци, децимални броеви и римски броеви. Како пример може да се земе римскиот број „V“ кој после транслитерација во кирилица се добива буквата „В“. За да се поправат грешките треба да се промени дел од програмскиот код и со тоа да се подобри функционалноста на алгоритамот.

- Метод за опслужување на други апликации овозможува пристап до транслитерирана содржина преку модул за транслитерација. Модулот има програмски код кој може да се интегрира во било која веб базирана апликација и може да работи за големи содржини[2]. Друг пристап се овозможува преку кориснички интерфејс каде корисниците можат да внесуваат содржини до 1500 зборови. Нашата верзија на алгоритамот за транслитерација е тестирана на Linux Debian Server со поддршка на PHP и MySQL технологиите. Без разлика на писмото на внесената содржина, таа секогаш мора да содржи точка како означување на завршеток на реченицте. Во случај да не содржи, алгоритамот автоматски ја додава ознаката, без исклучок на „?!” итн.

5 ИСТРАЖУВАЊЕ СО ПРИМЕНА НА ТРАНСЛИТЕРАЦИЈА

Во овој магистерски труд се направени повеќе истражувања во делот на повеќезначна транслитерација на цели реченици од латиница во кирилица. Според бројот на тестови кои се направени во различни етапи, за секое истражување се употребени различни типови на речници.

[2] Содржини кои може да имаат повеќе од 1500 зборови

5.1 Пребарување на повеќезначни зборови во отворен речник (ОР)

Првото направено истражување се однесува на пронаоѓање зборови кои можат да имаат повеќе од едно значење во дадено писмо.

За таа цел се користи зададена база (отворен речник[3]) со 261460 зборови од македонскиот јазик вклучувајќи: глаголи, имиња на локации и личности, придавки, именки и многу други поими, а со помош на изработена скрипта за транслитерација (vnesi_dvoznacni.php) е добиен резултат со повеќе од 5000 зборови со повеќезначна транслитерација. Тоа значи дека имаме резултат во кој повеќе од 2% од зборовите во речникот напишани на латиница при нивно транслитерирање во кирилица добиваат повеќе од едно значење. Овој процент на транслитерирани зборови е релативно голем во делот на структурата на самиот јазик. При пребарување на таквите зборови, со даден веб сервис кој може да прави разлика помеѓу кирилица и латиница, се појавува нивна повеќезначност.

5.2 Пребарување на повеќезначни зборови во толковен речник (ТР)

Направено е второ истражување за транслитерација во кое сега е применета друга база (толковен речник[4]) со 66338 зборови од македонскиот јазик вклучувајќи: глаголи, именки, придавки, прилози и заменки.

Преку процесот за добивање на двозначни зборови, добиен е резултат со повеќе од 1000 зборови со повеќезначна транслитерација. Тоа значи дека имаме резултат во кој 1,5% од зборовите во речникот напишани на латиница при нивно транслитерирање во кирилица добиваат повеќе од едно значење. Оттука се забележува дека е добиен помал процент на зборови во однос на претходното истражување, во кое беше применета база (ОР) со 261460 зборови. Доколку се споредат резултатите од истражувањата, покрај добиениот процентот на зборови, ќе се појави разлика и во користењето на поимите, односно транслитерација на зборови кои не содржат имиња на локации и личности. Тоа значи дека се користат само оние поими од македонскиот јазик од кои може да се добијат правилните значења при транслитерација на цели реченици.

[3] преземен од OpenOffice Dictionary, (http://www.openoffice.org/)
[4] преземен од Дигитален речник на македонскиот јазик, (http://www.makedonski.info/)

5.3 Споредување на ОР и ТР

Бидејќи некои зборови, кои се претходно добиени и зачувани во база на податоци, немаат смисла во една реченица при транслитерација, направено е рачно чистење на таквите зборови. Почетната состојба на двозначните поими во првиот тип на речник изнесуваше 2430 поими, кои после процесот на бришење се намалени на 647 двозначни поими или 1300 транслитерирани двозначни зборови[5] содржани во табелата за двозначност. Ваквиот добиен резултат остварува разлика од 73%, односно три пати помалку добиени двозначни поими[6]. Во вториот тип на речник немаме чистење на зборови. Ова е прикажано во следната табела.

Табела 3. Разлика на двозначни поими после рачно чистење на табелата за двозначност
Table 3. Difference between words with two meanings after manual cleaning of the table for ambiguity

	Почетна состојба	**Нова состојба**	**Разлика во %**
I – Отворен речник			
Двозначни поими	2430	647	-73,37
Вкупно зборови	261460	258553	-1,11
II – Толковен речник			
Двозначни поими	504	504	0
Вкупно зборови	66338	66338	0

5.4 Комбинирање на ОР и ТР

Следното тестирање се однесува на спојување на двата типа на речници, односно на отворениот и толковниот речник. Овде задолжително се креирани четири дополнителни табели и тоа: две табели за двозначност и две табели за содржина на соодветните зборови од речниците. За таа цел се направени две помали скрипти и тоа: за внесување на двозначни поими[7] и бришење на дупликати[8] на таквите поими. Во нашата база стандардно еден двозначен поим има два двозначни збора. После процесот на спојување на речниците и бришењето на дупликати, добиени се 1066 двозначни поими. Во новата

[5] двозначни зборови или кирилски зборови (на пример: докажи, докази) кои се добиени после транслитерацијата на двозначниот поим (на пример: dokazi).
[6] двозначен поим или латиничен поим (на пример: dokazi) кога се транслитерира во кирилица има повеќе од едно значење (на пример: докажи, докази).
[7] vnesi_dvoznacni.php
[8] brisi_duplikati_vo_dvoznacnost_spoeni.php

комбинирана табела за содржина се додадени сите овие двозначни поими и плус останатите зборови кои не се дупликат, и со тоа се добива разлика од -20,32% или намален фонд од 324891 на 258883 зборови. Ова е прикажано во следната табела.

Табела 4. Резултат на новодобиените табели при комбинирање на ОР и ТР
Table 4. Results of the newly obtained tables after combination of open dictionary and expository dictionary

ОР + ТР = КР (Комбиниран)	**Со дупликати**	**Без дупликати**	**Разлика во %**
Двозначни поими	1151	1066	-7,38
Вкупно зборови	324891	258883	-20,32

Предложениот алгоритам за повеќезначна транслитерација во овој труд може да трансформира цели реченици од латинско во кирилско писмо. Пред да се стартува овој алгоритам и да се добие краен резултат, мора да се поминат некои одредени етапи. За таа цел се користат неколку помошни скрипти и тоа: за автоматско читање на цели реченици од PDF документи[9], за одредување на коефициент на повеќезначните зборови[10], одредување на значењето на зборовите во целите реченици и решавање на повеќезначна транслитерација[11].

5.5 Автоматско читање на цели реченици од дигитални содржини

Ваквиот вид на читање овозможува добивање на цели реченици, а истовремено и одредување на позицијата на ограничени зборови. Кога се зборува за ограниченост, се однесува на одделување на важните поими од цела реченица, а тоа се зборови кои се наоѓаат лево и десно од повеќезначниот поим. Досега со помош на изработената скрипта (citaj_recenici.php) автоматски се прочитани над 300 PDF документи. Истата скрипта е применета кај три типа на речници, а тоа се: *ОР, ТР и КР*. За секој речник е креирана база на податоци со идентично креирани табели.

Според анализите кои се направени за прочитаните PDF документи, се добиени резултати со нумерички податоци и тоа:

[9] citaj_recenici.php
[10] koeficient.php
[11] resenie_za_dvoznacnost.php

- вкупно двозначни зборови;
- вкупно реченици прочитани во сите PDF документи;
- вкупно зборови прочитани во сите PDF документи;
- вкупно реченици со повеќезначни зборови;
- вкупно поими[12] (лево, десно) од двозначните зборови;
- вкупен број на повторување на сите поими (лево, десно) од двозначните зборови.

Таквите податоци се добиени преку процесот на внесување и одредување на содржина од надворешни извори, а тоа се однесува на текстови од PDF документи и веб сајтови.

Резултати

Следните резултати се добиени врз основа на прочитаните PDF документи со содржини напишани на кирилица и при нивно читање без употреба на транслитерација од кирилица во латиница, директно за секој пронајден збор се прави селектирање во база, се инкрементира определена вредност и се меморира во табелата за резултати.

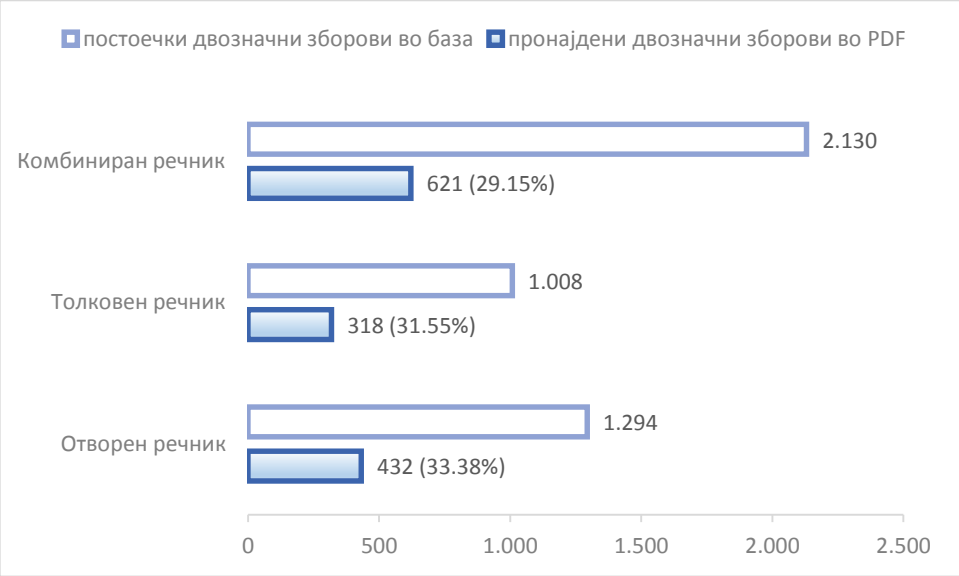
Табела 5. Резултати добиени при читање на реченици од PDF документи
Table 5. Results obtained after reading sentences from PDF documents

	Отворен речник	**Толковен речник**	**Комбиниран речник**
вкупно двозначни зборови	432 / 1294	318 / 1008	621 / 2130
вкупно реченици прочитани во сите PDF документи	329.958	329.350	393.034
вкупно зборови прочитани во сите PDF документи	4.187.998	4.241.813	4.574.189
вкупно реченици со повеќезначни зборови	36.746	43.635	60.692
вкупно поими (лево, десно) од двозначните зборови	60.365	68.082	102.821
вкупен број на повторување на сите поими (лево, десно) од двозначните зборови	229.315	276.528	355.336

[12] поими се однесува на сите кирилски зборови кои се наоѓаат во речникот

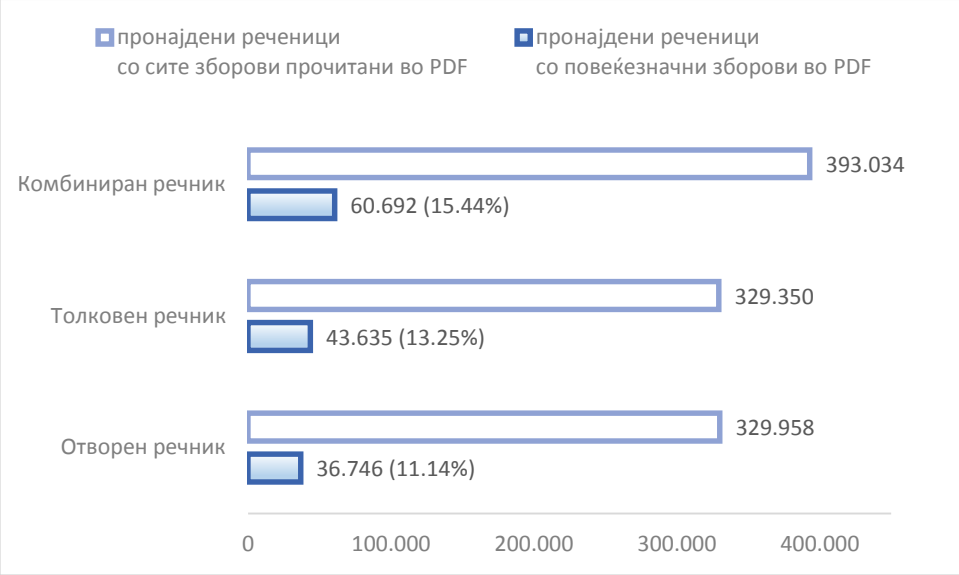
Секој двозначен збор може да се појави повеќе пати во една реченица. За таквиот тип на зборови се направени пресметки и се добиени три различни резултати, т.е. за секој речник посебно. Скриптата работи според правилото читај - најди – запиши и со неа се добиени следните податоци:

- Во отворениот речник од вкупно 647 постоечки базирани двозначни поими на латиница или 1.294 двозначни зборови на кирилица, со процесот на читање се добиени само 33,38% или тоа се 432 двозначни зборови. Кај толковниот речник од 1.008 двозначни зборови се добиени 318 или 31,55%, и за комбинираниот од 2.130 само 29,15% или 621 двозначен збор.



Слика 3. Процент на двозначни зборови (кои се прочитани во PDF документи) од вкупно двозначни зборови кои се веќе постоечки во база
Figure 3. Percentage of words with two meanings (read from PDF documents) from the total number of words with two meanings existing in the database

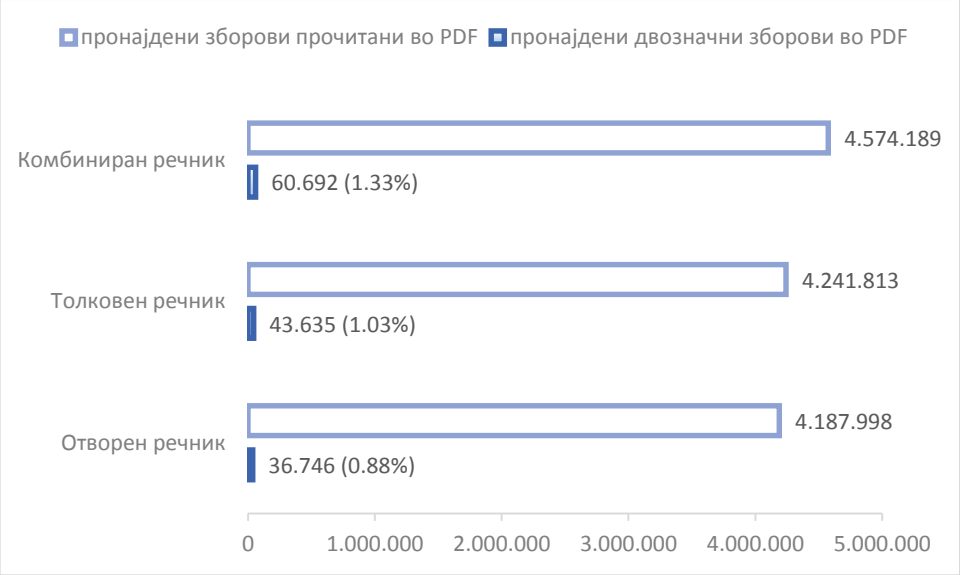
- Со фактот дека една реченица може да има најмалку еден двозначен збор, оттука може да се земе во предвид и процентот на двозначни реченици. Па затоа кај отворениот речник имаме вкупно 329.958 прочитани реченици во сите PDF документи, од кои 11,14% се реченици со двозначни зборови односно 36.746 такви реченици. Кај толковниот речник имаме 13,25% од 329.350, и кај комбинираниот 15,44% двозначни реченици од вкупно 393.034 реченици.



Слика 4. Процент на двозначни реченици од вкупно реченици прочитани во сите PDF документи

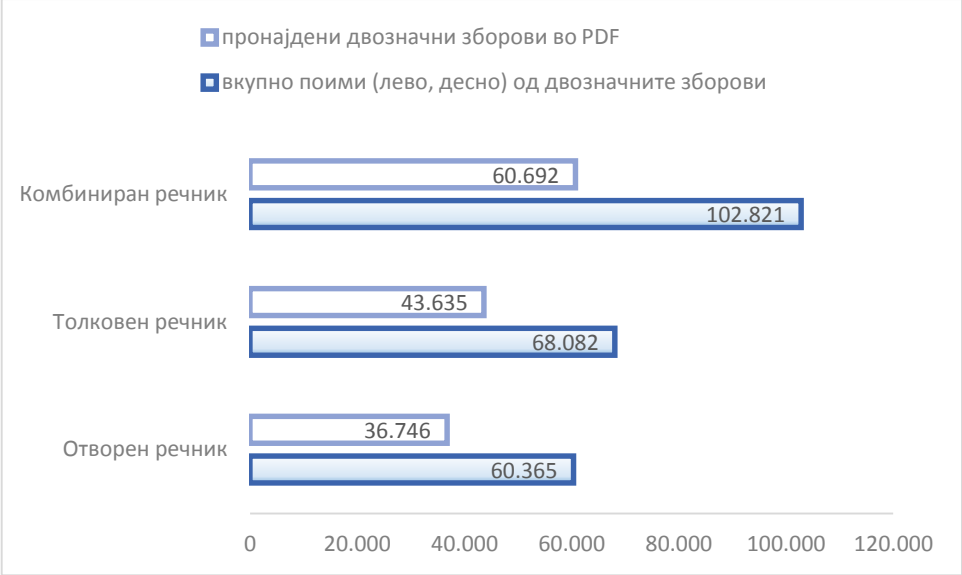
Figure 4. Percentage of sentences with two meanings from the total number of sentences read in all PDF documents

- Секоја реченица е составена од зборови и затоа за секој збор може да се издвојат неколку податоци. Бидејќи зборовите се повторуваат во текот на автоматското читање, вредностите за секој збор постојано се зголемуваат. Интересен е податокот за обичните зборови и двозначните поими за кои во истражувањето се добиени резултати со број поголем од еден. Кај отворениот речник се прочитани и како бројка се внесени во база вкупно 4.187.998 зборови. Според тој податок е издвоен процент на двозначни зборови. Тоа значи дека имаме 0,88% или 36.746 двозначни зборови појавени во вкупниот број на зборови прочитани во сите PDF документи. Кај толковниот речник се внесени 4.241.813 зборови и од нив само 1,03% или 43.635 се двозначни. Во комбинираниот речник за разлика од претходните два речника, имаме 1,33% застапеност на двозначни зборови, односно 60692 од вкупно 4.574.189 зборови.



Слика 5. Процент на двозначни зборови од вкупно зборови прочитани во сите PDF документи
Figure 5. Percentage of words with two meanings from the total number of words read in all PDF documents

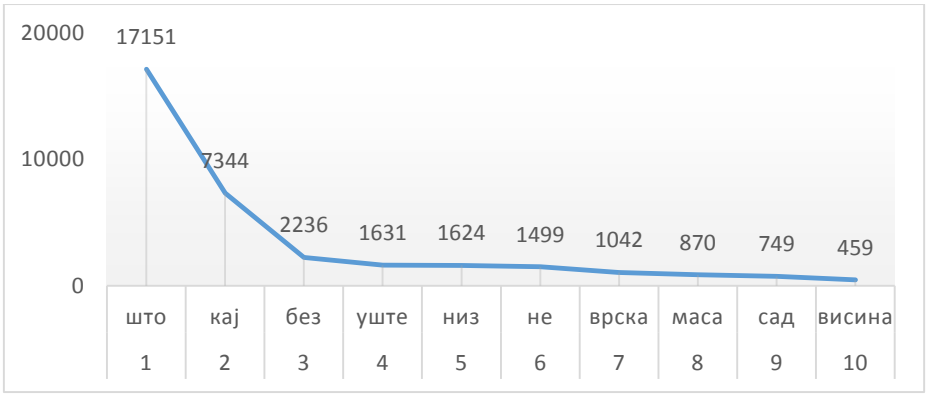
- Следниот добиен податок се однесува на позиционирани места во речениците, и како таков може лесно да се спореди со претходно прикажаните резултати. Податокот треба да функционира само во услови кога зборот е двозначен, што значи дека лесно може да добие позиција и број. Тој број се однесува на вкупно поими кои се наоѓаат лево и десно од двозначните зборови, а позицијата преку процесот за одредување на позиција на ограничените зборови изнесува различно за секој речник. Излезот на скриптата (citaj_recenici.php) резултираше кај отворениот речник со вредност од 60365 поими, кои се наоѓаат од -3 до +3 позиција од двозначниот збор. Толковниот речник иако има помал фонд на двозначни зборови складирани во база, сепак за разлика од претходниот, има поголем број на поими, односно 68082 поими, и за комбинираниот имаме 102821 поими кои се наоѓаат лево и десно од двозначниот збор.



Слика 6. Вкупно поими (лево, десно) од двозначните зборови
Figure 6. Total number of words (left, right) from words with two meanings

- Покрај сите добиени вредности, исто така како важен податок се издвојува и бројот на застапеноста на двозначните поими. Од вкупно 432 (во ОР), 318 (во ТР) и 621 (во КР) постоечки двозначни поими пронајдени во PDF документите, според анализите кои се применети овде, најзастапени се зборовите кои имаат најголема вредност, односно тоа е бројот на прочитани повеќезначни зборови во секоја реченица. Овде следуваат најзастапените двозначни зборови добиени од автоматско прочитаните документи.

Слика 7. 10 најзастапени двозначни зборови од ОР
Figure 7. 10 most common words with two meanings in the open dictionary



Слика 8. 10 најзастапени двозначни зборови од ТР
Figure 8. 10 most common words with two meanings in the expository dictionary

Слика 9. 10 најзастапени двозначни зборови од КР
Figure 9. 10 most common words with two meanings in the combined dictionary

Кога станува збор за застапеност на зборовите, добиен е уште еден резултат кој сега е пресметан според поимите кои се наоѓаат лево и десно од двозначните зборови. Тоа значи дека се добиваат бројки за секој поим со кои може лесно да се потенцираат позициите на зборовите и нивните повторувања. Овие резултати се прикажани во следната табела.

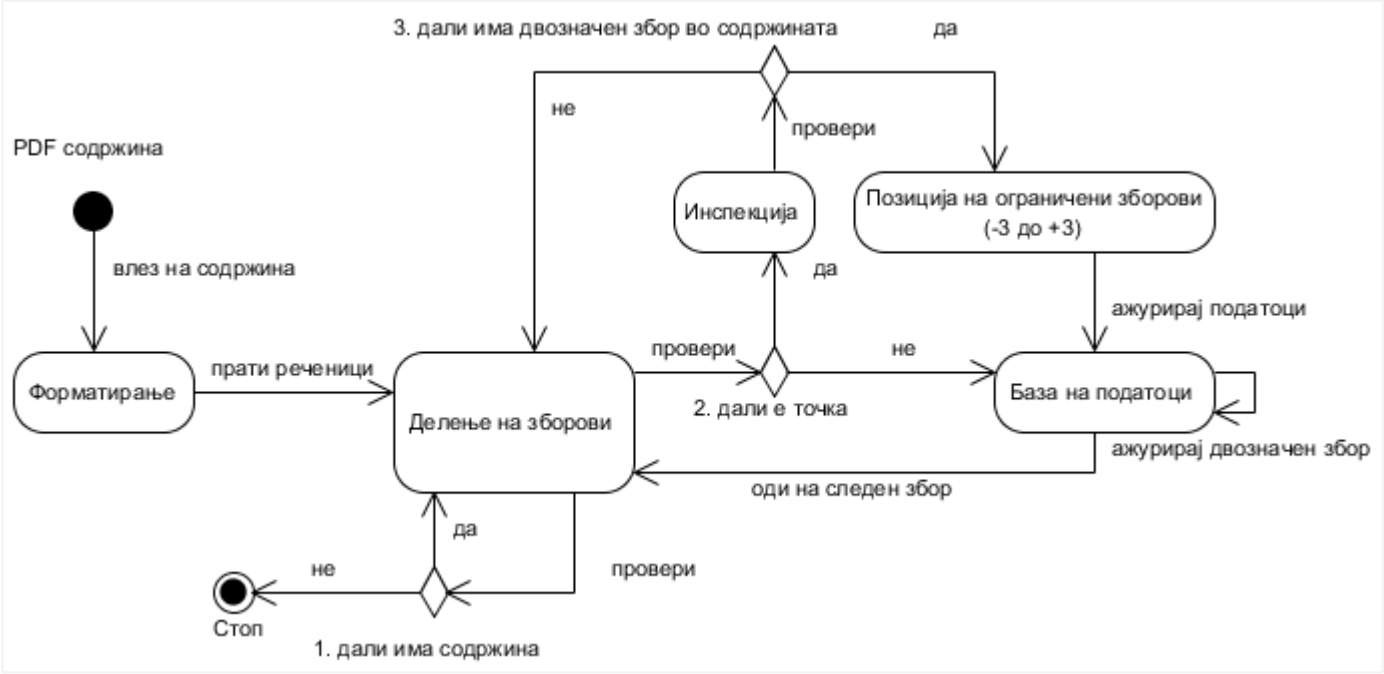
Табела 6. Најзастапени поими лево и десно од двозначните зборови според повторувањето кај отворениот тип на речник
Table 6. Most common terms left and right from the words with two meanings after repetition in the open dictionary

Id	двозначен збор	поим (лево, десно)	minus3	minus2	minus1	plus1	plus2	plus3	повторувања
261095	што		928	3612	1481	142	544	1102	21465
261095	што	се	440	149	30	4041	832	745	6237
261095	што	на	1075	1071	98	162	609	1260	4275
261095	што	е	356	341	168	1770	316	207	3158
261095	што	и	708	563	87	161	287	591	2397
261095	што	како	66	33	2014	33	51	120	2317
261095	што	со	169	345	956	142	146	541	2299
261095	што	во	293	300	27	351	334	607	1912
261095	што	да	142	112	1	59	821	463	1598
261095	што	тоа	63	38	1212	78	80	33	1504
261095	што	од	239	349	49	71	321	344	1373
261095	што	за	221	277	188	70	151	457	1364
261095	што	при	7	6	1076	48	17	56	1210
261095	што	каде	4	0	1087	0	5	3	1099
261095	што	го	117	6	0	580	232	122	1057
261095	што	ги	47	23	0	527	209	157	963
261095	што	така	18	15	803	3	21	12	872
261095	што	не	55	104	16	385	151	103	814
261095	што	ќе	25	6	0	574	101	57	763
261095	што	затоа	4	6	694	0	0	2	706
250228	уште	се	28	170	230	134	91	37	690
261095	што	ја	35	6	0	420	120	69	650
261095	што	она	2	23	500	0	12	2	539
71764	значи	дека	2	0	1	496	7	3	509
111223	маса	на	40	229	9	140	9	77	504
254525	целата		7	17	1	0	132	55	485
261095	што	може	4	0	2	296	99	57	458
261095	што	поради	12	184	248	4	4	2	454
222264	сега		20	52	24	39	13	32	427
261095	што	до	56	21	4	18	161	160	420

- Сите досегашни резултати прикажани преку дијаграми се дел од нашите истражувања и според нив се направени споредувања во однос на складираните податоци во база.

5.6 Статистичка обработка на автоматско прочитаните цели реченици со употреба на база на податоци

Добиените резултати во потточка (5.5) се вредности кои се добиени во овој дел. Процесот за внесување на податоци е применет кај трите типа на речници со помош на скриптата citaj_recenici.php. Во текот на овој процес, секој прочитан поим во дадена реченица се проверува дали постои во база и според тоа се добиваат филтрирани зборови. Тоа се однесува и на знаците (%&()0123456789 итн.), кои не се земени во предвид што се наоѓаат во база на податоци. На овој начин полесно може да се решава повеќезначната транслитерација. Начинот на внесување на податоци во база може да се види во следниот дијаграм (Слика 10).



Слика 10. Добивање на филтрирани зборови и нивно внесување во база на податоци
Figure 10. Obtaining filtrated words and their entering in database

За подетално да се разбере постапката за внесување на поими во база на податоци, подолу е прикажан резултат кој е добиен со помош на скриптата за внесување на податоци директно во база со веќе зададени параметри. Таквите параметри се едноставни за користење и можат да се применат овде во делот за транслитерацијата на цели реченици од латиница во кирилица. Скриптата може да работи на повеќе нивоа и тоа за: препознавање на реченица, мали и големи букви, цели и децимални броеви, знаци и празни места. Преку овие нивоа се добива резултат кој придонесува синтаксичко разбирање на една

транслитерирана реченица. Во следниот поднаслов (Резултат) е опишан главниот дел на процесот за добивање на вредности на параметрите за: повторувања и позицијата на зборовите кои се наоѓаат пред и после двозначниот збор.

Резултат

За подетален опис на резултатот, креиран е еден PDF документ со кирилска содржина и од него со скриптата citaj_recenici.php се вчитуваат реченици кои содржат двозначни зборови. Во нашиот случај се прочитани 34 реченици кои на крајот завршуваат со точка, а поимите кои се наоѓаат на позиција три места пред и после двозначниот збор се инкрементираат за 1 во база на податоци, а со тоа се добива и параметарот n_i - повторување на ист поим на одредена позиција, кој ќе биде опишан во потточката (5.7). Во продолжение е прикажана една реченица со кирилски поими кои веќе постојат во база на податоци и истите ќе ги анализираме:

- „На минатата забава во Скопје , овие вешти жени им помагаа на старите лица од дневниот центар.”

Основно нешто е препознавањето на поимите кои имаат две значења кога ќе се транслитерираат од латиница во кирилица. Во овој случај немаме транслитерирање на поими туку директно разгледување на прочитаните кирилски поими. Тука се препознаени 3 двозначни зборови и тоа: *забава, вешти и жени*. Тие зборови скриптата лесно може да ги препознае во базата на податоци и за сите да добие индекс. Ова може да се прикажи во следната табела:

Табела 7. Претходници и следбеници на двозначниот збор
Table 7. Ancestors and followers of the word with two meanings

двозначни поими - латиница	двозначни зборови - кирилица	-3	-2	-1	0	1	2	3
zabava	жабава, **забава**		на	минатата	**забава**	во	Скопје	
vesti	вести, **вешти**	Скопје		овие	**вешти**	жени	им	помагаа
zeni	**жени**, зени		овие	вешти	**жени**	им	помагаа	на

Во секоја реченица може да има најмалку еден повеќезначен збор или пак ниту еден таков збор. Во дадениот пример секој од таквите зборови имаат свои

претходници и следбеници, односно поими кои стојат пред и после двозначниот збор во реченицата. Минималниот број на поими во една реченица секогаш треба да е поголем од еден за да се овозможи формирање на опсег на различни поими од -3 до +3 позиција од самиот двозначен збор кој се наоѓа на 0 позиција. Од примерот со реченицата да го земеме двозначниот збор „забава“ и да ги преземеме вредностите, односно бројот на позиционираните поими кои може да се повторуваат или пак да содржат некое двозначење. Оттука се гледа дека од -2 до +2 позиција се наоѓаат (на, минатата, во, Скопје), додека пак останатите позиции се празни. Истите овие поими може да се наоѓаат и во други реченици со истиот двозначен збор на истата позиција. За да се комплетира процесот за внесување на податоци, секогаш повторувањата на позиционираните поими треба да се внесат во соодветна табела во база. Наредните два двозначни збора „вешти“ и „жени“ позиционираат некои поими кои се наоѓаат и во други реченици. Таков е примерот со зборот „жени“, кој се повторува на -3 и +1 позиција вкупно четири пати, во случај кога двозначниот збор е „вешти“ со индекс 21103. Оттука може да се забележи дека двозначните зборови може да се наоѓаат еден до друг и во тој случај без разлика дали позицинираниот поим е двозначен, тој останува ист без да се одредува неговото значење. Ваквото претставување на вредностите е прикажано во следната табела.

Табела 8. Резултат од 34 прочитани кирилски реченици од примерот за позиционирање на поими
Table 8. Result from 34 read Cyrillic sentences from the example of words positioning

id двозначен поим	поим	minus3	minus2	minus1	plus1	plus2	plus3	повторувања
54839	минатата	0	0	1	0	0	0	1
54839	на	0	1	0	0	0	0	1
54839	во	0	0	0	1	0	0	1
54839	скопје	0	0	0	0	1	0	1
21103	овие	0	1	1	0	0	0	2
21103	Скопје	1	0	0	0	0	0	1
21103	жени	1	0	0	3	0	0	4
21103	им	0	0	0	0	1	0	1
21103	помагаа	0	0	0	0	0	1	1
53747	вешти	0	0	3	0	0	1	4
53747	овие	0	1	0	0	0	0	1
53747	им	0	0	0	1	0	0	1
53747	помагаа	0	0	0	0	1	0	1
53747	на	1	0	0	1	0	1	3

5.7 Чекори за одредување на правилен повеќезначен збор во цели реченици

Во овој дел се опишани четирите чекори преку кои се одредува точниот повеќезначен збор кој се добива после транслитерацијата на дадени реченици (на пр. од „dokazi“ преминува во „докази“ и „докажи“). Тоа е објаснето во следните чекори:

Чекор 1: Барање за транслитерација на содржина од корисникот до скрипта1 (скрипта за читање и ограничување на цели реченици);

Чекор 2: Замена на загради[13] на двозначниот збор и ограничување на зборови и нивно повеќезначно одредување во дадената содржина;

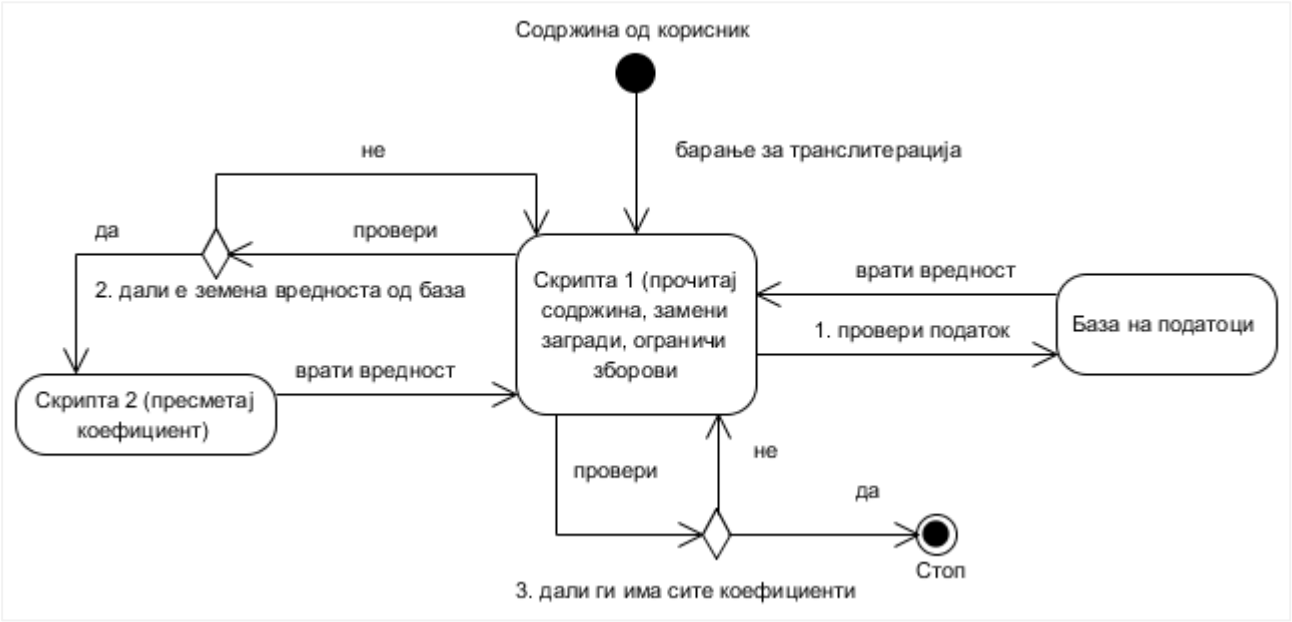
Чекор 3: Добивање на коефициент со помош на скрипта 2 (скрипта за пресметување на коефициент);

- Во овој чекор се пресметува и бројот на реченици со двозначните зборови кои се добиени со помош на повторување на секој поим од реченициte.

Чекор 4: Прикажување на резултатите на кориснички интерфејс.

За да се појасни ова следува дијаграм со примена на коефициент за добивање на крајните резултати (Слика 11).

[13] на пример [докажи][докази]



Слика 11. Одредување на точни повеќезначни зборови во цели реченици со помош на скрипта 1 и скрипта 2
Figure 11. Obtaining precise ambiguous words in full sentences using script 1 and script 2

Содржината која е внесена од корисникот во корисничкиот интерфејс на веб сервисот, како барање за транслитерација, секогаш се процесира низ две скрипти и тоа:

- Скрипта за читање и ограничување на целите реченици (Скрипта 1);
- Скрипта за пресметување на коефициент (Скрипта 2).

За да се пресмета главниот коефициент, најпрво чекор 1 заедно со содржината внесена од корисникот преминува во чекор 2, каде се вчитува таа содржина. Овде се вклучува скрипта 1 која истовремено ги проверува сите зборови во база и на секој од нив се доделува вредност (број на повторувања) во база на податоци. Откако сите зборови ќе добијат соодветна вредност, може слободно да се премине на чекор 3, што означува почеток за пресметување на коефициентот. Овде се вклучува скрипта 2 која понатаму враќа вредности на скрипта 1. Доколку сите коефициенти од содржината се пресметани, чекор 3 преминува на чекор 4, што значи дека е одреден коефициентот на повеќезначните зборови.

5.8 Дефиниција на формули за добивање на коефициент за транслитерација со примена на условна веројатност и Бајесови формули

Го разгледуваме проблемот на двозначна транслитерација и притоа користиме алгоритам во којшто на преден план е примената на условната веројатност и Бајесовите формули. Пред да го опишеме начинот на којшто го решаваме проблемот, ќе дадеме кратка дефиниција на условната веројатност и на Бајесовите формули.

Нека X и Y се случајни променливи кои се дефинирани на ист простор на веројатност Ω, и нека Y е дискретна случајна променлива. Нека R_Y е множеството вредности на случајната променлива Y т.е. $\{y \in R \mid P(Y = y) > 0\}$. R_Y е конечно или пребројиво множество.

Ако $a \in R$ и $y \in R_Y$, дефинираме условна распределба со:

$$F(a \mid y) = P(X \le a \mid Y = y) = \frac{P(X \le a \mid Y = y)}{P(Y = y)} .$$

Разгледуваме реченици напишани на кирилица во коишто се среќава одреден двозначен збор. Нека двозначниот збор во оригиналната реченица се наоѓа на $i-$тата позиција. Во процесот на транслитерација разгледуваме вкупно N реченици во кои се среќава двозначниот збор X_i и нека претпоставиме дека двозначниот збор има две значења Y и Z. Нека X_{i-k} е случајна променлива која е дефинирана како бројот на појавувања на даден збор на k-тата позиција, лево од двозначниот збор во N реченици во кои се среќава двозначниот збор. Нека X_{i+k} е случајна променлива која е дефинирана како бројот на појавувања на даден збор на k-тата позиција десно од двозначниот збор во N реченици во кои се среќава двозначниот збор.

$$\begin{array}{ccccccccc} X_{i-3} & X_{i-2} & X_{i-1} & & X_i & & X_{i+1} & X_{i+2} & X_{i+3} \\ \downarrow & \downarrow & \downarrow & & \square\ \square & & \downarrow & \downarrow & \downarrow \\ X_{i-3} & X_{i-2} & X_{i-1} & Y & & Z & X_{i+1} & X_{i+2} & X_{i+3} \end{array}$$

Единствената промена во процесот на транслитерација ќе биде направена кај двозначниот збор X_i на i-тата позиција. Неговиот транслитериран збор би бил Y или Z.

Го разгледуваме влијанието на зборовите во реченицата во којашто се среќава двозначниот збор, коишто се најдени на три позиции пред двозначниот збор и три позиции после двозначниот збор.

Нека H_{i-k} го означува случајниот настан: збор којшто се наоѓа на k-тата позиција лево од двозначниот збор X_i, и нека H_{i+k} го означува настанот: збор којшто се наоѓа на k-тата позиција десно од двозначниот збор.

$$P(Y) = P(H_{i-3}) \cdot P_{H_{i-3}}(Y) + P(H_{i-2}) \cdot P_{H_{i-2}}(Y) + P(H_{i-1}) \cdot P_{H_{i-1}}(Y) + \\ + P(H_{i+1}) \cdot P_{H_{i+1}}(Y) + P(H_{i+2}) \cdot P_{H_{i+2}}(Y) + P(H_{i+3}) \cdot P_{H_{i+3}}(Y)$$

т.е. $P(Y) = \sum_{k=i-3}^{i+3} P(H_k) \cdot P_{H_k}(Y)$ имајќи во предвид дека $P(H_i) \cdot P_{H_i}(Y) = 0$, бидејќи на i-тата позиција се наоѓа двозначниот збор X_i.

$$P(H_{i-3}) \cdot P_{H_{i-3}}(Y) = P(H_{i-3}) \cdot \frac{P(YH_{i-3})}{(H_{i-3})} = P(YH_{i-3})$$

Но сакаме да одредиме кои од зборовите има најголемо влијание во процесот на транслитерација т.е. во кој од зборовите би бил транслитериран двозначниот збор. Користиме Бајесови формули:

$$P_{X_i}(Y) = \frac{P(Y) \cdot P_Y(X_i)}{P(X)}.$$

Во нашиот алгоритам $P_{X_i}(Y) = \frac{P(Y) \cdot P_Y(X_i)}{P(X)} + \lambda$, и алгоритамот не е осетлив на вредноста на λ. Во нашиот алгоритам $\lambda = 0.5$.

На истиот начин одредуваме $P_{X_i}(Z) = \frac{P(Z) \cdot P_Y(X_i)}{P(X)} + \lambda$.

Ако $P_{X_i}(Y) > P_{X_i}(Z)$ тогаш двозначниот збор X ќе биде транслитериран во Y , во спротивно X ќе биде транслитериран во Z.

5.8.1 Параметри за решавање на повеќезначна транслитерација

Покрај параметрите во *Табела 8* постојат и други параметри чии вредности се добиени со помош на ново дефинираните формули и истите ќе бидат овде опишани. Параметрите кои се потребни за решавање на повеќезначна транслитерација од латиница во кирилица се карактеризираат според обемот на реченици кои се прочитани од надворешните извори. Тие параметри се однесуваат на:

- K - коефициент за одредување на двозначниот збор;
- n_i - повторување на ист поим на одредена позиција пред или по двозначниот збор во повеќе реченици;
- N - вкупниот број на реченици во кои првиот активен двозначен збор се повторува;
- M - вкупниот број на реченици во кои вториот двозначен збор се повторува.

Во оваа магистерска се истражувани и добиени три формули во кои се користат овие наведени параметри. Тоа значи дека имаме три различни дефиниции за секоја формула посебно, преку кои се добиваат различни коефициенти за двозначните зборови при транслитерација.

Формула 1. *Коефициент во кој n_i претставува вкупно повторување на секој збор посебно без разлика на која позиција се наоѓа од двозначниот поим, во сите досега прочитани реченици*

$$K = \sum_{i=1}^{6} n_i * \frac{1}{N+M} + 0.5$$

Формула 2. *Коефициент во кој n_i претставува вкупно повторување на секој збор посебно во зависност од дадената позиција (-3 -2 -1 +1 +2 +3), во сите досега прочитани реченици*

$$K = \sum_{i=1}^{6} k_i$$

$$k_i = \sum_{i=1}^{6} n_i * \frac{1}{N+M} + 0.5$$

Формула 3. *Коефициент во кој n_i претставува вкупно повторување на секој збор во зависност од дадената позиција (-3 -2 -1 +1 +2 +3) во моментално разгледуваната реченица, чии вредности се добиени од сите досега прочитани реченици*

$$K = \sum_{i=1}^{6} n_i * \frac{1}{N+M} + 0.5$$

Коефициент претставува вредност со кој се одредува точниот двозначен збор, а со тоа се добива и значењето на цела реченица при транслитерација од латиница во кирилица. За добивање на коефициент се вклучени сите параметри од формула 1, но најважно е сите тие претходно да се внесат во база на податоци, бидејќи алгоритамот за транслитерација го одредува значењето со коефициентот преку тие параметри. За да се појасни добивањето на коефициентот, во потточка (5.8.2) следуваат примери каде подетално се опишани параметрите.

5.8.2 Равенства за транслитерирање на двозначни поими

Во оваа точка е опишан примерот во кој се земени во предвид 34 реченици во кои има двозначни поими. Секоја од нив е решена на три различни начини, бидејќи се применети три различни формули. Една од тие цели реченици е реченицата:

- “Na minatata **zabava** vo Skopje , ovie **vesti zeni** im pomagaa na starite lica od dnevniot centar.“

Оваа реченица е напишана на латиница, и во неа се наоѓаат повеќе зборови, од кои само три се издвојуваат како поими со повеќе од едно значење. Тоа се ”*zabava”, “vesti”, “zeni”,* кои при транслитерација во кирилица, се добиваат двозначните зборови: „забава, жабава”; „вести, вешти”; „жени, зени”. Алгоритамот работи со позиционирање на зборови кои се наоѓаат пред и после двозначниот збор. Ако се индексира поимот “zabava” во база, за него автоматски може да се издвојат неколку податоци и тоа: бројот на повторување на двозначните зборови „забава, жабава”, бројот на реченици во кои се наоѓаат тие зборови и позиционираните места на тие зборови.

После мапирањето на зборовите од латиница во кирилица добиен е зборот „минатата” кој се наоѓа на минус првата позиција или пред двозначниот збор „забава”. Во нашиот случај тој се повторува само еднаш, што значи дека $n_3 = 1$.

За да се добијат и другите параметри како што е M, потребно е индексирање и на зборот „жабава”. Тоа може да се види во Табела 9 каде што $N = 1, M = 0$. Вредноста на $M = 0$, бидејќи зборот „жабава” не постои во ниту една реченица, во која, тој бил прочитан од PDF документите и зачуван во база. Според формула 1, бидејќи -3 и +3 позиција се празни, вредноста за нив останува 0.5. За да се добие целиот коефициент K на двозначниот збор „забава”, потребно е да се пресметаат и преостанатите вредности за зборовите „на, во, Скопје”. Истиот чекор треба да се повтори и за двозначниот збор „жабава” со зборовите „на, минатата, во, Скопје”. Сите овие параметри доколку се применат во формулата, каде n_i е вкупното повторување на секој збор посебно без разлика на која позиција се наоѓа од двозначниот поим, во сите досега прочитани реченици, се добива следниот резултат:

$$K_{\text{забава}} = \text{празно}_{-3} + \text{на}_{-2} + \text{минатата}_{-1} + \text{во}_{+1} + \text{Скопје}_{+2} + \text{празно}_{+3}$$
$$= \left(0 * \frac{1}{1+0} + 0.5\right) + \left(1 * \frac{1}{1+0} + 0.5\right) + \left(1 * \frac{1}{1+0} + 0.5\right) + \left(1 * \frac{1}{1+0} + 0.5\right)$$
$$+ \left(1 * \frac{1}{1+0} + 0.5\right) + \left(0 * \frac{1}{1+0} + 0.5\right) = 0.5 + 1.5 + 1.5 + 1.5 + 1.5 + 0.5 = 7$$

$$K_{\text{жабава}} = \text{празно}_{-3} + \text{на}_{-2} + \text{минатата}_{-1} + \text{во}_{+1} + \text{Скопје}_{+2} + \text{празно}_{+3}$$
$$= \left(0 * \frac{1}{0+1} + 0.5\right) + \left(0 * \frac{1}{0+1} + 0.5\right) + \left(0 * \frac{1}{0+1} + 0.5\right) + \left(0 * \frac{1}{0+1} + 0.5\right)$$
$$+ \left(0 * \frac{1}{0+1} + 0.5\right) + \left(0 * \frac{1}{0+1} + 0.5\right) = 3$$

Овие добиени коефициенти се однесуваат само на двозначниот поим “zabava”, што значи дека преостануваат уште четири вакви проверки за преостанатите два поими “vesti и zeni”. Поголемата добиена вредност од $K_{\text{забава}}$ и $K_{\text{жабава}}$ претставува решение за правилниот двозначен збор, кој во нашиот случај е зборот „забава”. Откако ќе се добијат и наредните коефициенти $K_{\text{вешти}}, K_{\text{вести}}, K_{\text{жени}}$ и $K_{\text{зени}}$, тогаш може слободно да се провери дали транслитерираната реченица е точна. Ова е прикажано во следната Табела.

Табела 9. Резултат на параметрите за транслитерација според првата формула (Формула 1)
Table 9. Result from transliteration parameters after the first formula (Formula 1)

двозначен збор	**поим**	**-3**	**-2**	**-1**	**+1**	**+2**	**+3**	**Повторувања** (n_i)	$\frac{1}{N+M}$	$n_i * \frac{1}{N+M} + 0.5$
забава	минатата	0	0	1	0	0	0	1	1	1.5
забава	на	0	1	0	0	0	0	1	1	1.5
забава	во	0	0	0	1	0	0	1	1	1.5
забава	Скопје	0	0	0	0	1	0	1	1	1.5
вешти	овие	0	1	1	0	0	0	2	0.043	0.586
вешти	Скопје	1	0	0	0	0	0	1	0.043	0.543
вешти	жени	1	0	0	3	0	0	4	0.043	0.672
вешти	им	0	0	0	0	1	0	1	0.043	0.543
вешти	помагаа	0	0	0	0	0	1	1	0.043	0.543
жени	вешти	0	0	3	0	0	1	4	0.25	1.5
жени	овие	0	1	0	0	0	0	1	0.25	0.75
жени	им	0	0	0	1	0	0	1	0.25	0.75
жени	помагаа	0	0	0	0	1	0	1	0.25	0.75
жени	на	1	0	0	1	0	1	3	0.25	1.25
вести	на	0	0	0	1	0	0	1	0.043	0.543

Табела 10. Приказ на вкупниот број на прочитани реченици во кои се наоѓаат двозначните поими
Table 10. Overview of the total number of read sentences that contain word with two meanings

индекс	**двозначен збор**	**двозначен поим**	**N или M**
54839	забава	zabava	1
21103	вешти	vesti	13
53747	жени	zeni	4
236012	сто	sto	1
215017	раце	race	1
182155	пошта	posta	1
261095	што	sto	2
238861	така	taka	1
20687	вести	vesti	10

Овие резултати се само вредности кои се добиени со помош на Формула 1. Наредната формула се однесува на коефициент K кој е составен од под коефициенти k_i. Во овие под коефициенти се користат истите параметри како во Формула 1, во кои сега n_i е вкупно повторување на секој збор посебно во зависност од дадената позиција (-3 -2 -1 +1 +2 +3), во сите досега прочитани реченици.

Следува равенство на само еден коефициент:

$$K_{\text{забава}} = (k_1 + k_2 + k_3 + k_4 + k_5 + k_6) = 0 + 4 + 4 + 4 + 4 + 0 = 16$$

$$k_1 = \text{празно}_{-3} + \text{празно}_{-2} + \text{празно}_{-1} + \text{празно}_{+1} + \text{празно}_{+2} + \text{празно}_{+3}$$
$$= \left(0 * \frac{1}{0+0} + 0\right) + \left(0 * \frac{1}{0+0} + 0\right) + \left(0 * \frac{1}{0+0} + 0\right) + \left(0 * \frac{1}{0+0} + 0\right)$$
$$+ \left(0 * \frac{1}{0+0} + 0\right) + \left(0 * \frac{1}{0+0} + 0\right) = 0$$

$$k_2 = \text{на}_{-3} + \text{на}_{-2} + \text{на}_{-1} + \text{на}_{+1} + \text{на}_{+2} + \text{на}_{+3}$$
$$= \left(0 * \frac{1}{1+0} + 0.5\right) + \left(1 * \frac{1}{1+0} + 0.5\right) + \left(0 * \frac{1}{1+0} + 0.5\right) + \left(0 * \frac{1}{1+0} + 0.5\right)$$
$$+ \left(0 * \frac{1}{1+0} + 0.5\right) + \left(0 * \frac{1}{1+0} + 0.5\right) = 0.5 + 1.5 + 0.5 + 0.5 + 0.5 + 0.5 = 4$$

$$k_3 = \text{минатата}_{-3} + \text{минатата}_{-2} + \text{минатата}_{-1} + \text{минатата}_{+1} + \text{минатата}_{+2} + \text{минатата}_{+3}$$
$$= \left(0 * \frac{1}{1+0} + 0.5\right) + \left(0 * \frac{1}{1+0} + 0.5\right) + \left(1 * \frac{1}{1+0} + 0.5\right) + \left(0 * \frac{1}{1+0} + 0.5\right)$$
$$+ \left(0 * \frac{1}{1+0} + 0.5\right) + \left(0 * \frac{1}{1+0} + 0.5\right) = 0.5 + 0.5 + 1.5 + 0.5 + 0.5 + 0.5 = 4$$

$$k_4 = \text{во}_{-3} + \text{во}_{-2} + \text{во}_{-1} + \text{во}_{+1} + \text{во}_{+2} + \text{во}_{+3}$$
$$= \left(0 * \frac{1}{1+0} + 0.5\right) + \left(0 * \frac{1}{1+0} + 0.5\right) + \left(0 * \frac{1}{1+0} + 0.5\right) + \left(1 * \frac{1}{1+0} + 0.5\right)$$
$$+ \left(0 * \frac{1}{1+0} + 0.5\right) + \left(0 * \frac{1}{1+0} + 0.5\right) = 0.5 + 0.5 + 0.5 + 1.5 + 0.5 + 0.5 = 4$$

$$k_5 = \text{Скопје}_{-3} + \text{Скопје}_{-2} + \text{Скопје}_{-1} + \text{Скопје}_{+1} + \text{Скопје}_{+2} + \text{Скопје}_{+3}$$
$$= \left(0 * \frac{1}{1+0} + 0.5\right) + \left(0 * \frac{1}{1+0} + 0.5\right) + \left(0 * \frac{1}{1+0} + 0.5\right) + \left(0 * \frac{1}{1+0} + 0.5\right)$$
$$+ \left(1 * \frac{1}{1+0} + 0.5\right) + \left(0 * \frac{1}{1+0} + 0.5\right) = 0.5 + 0.5 + 0.5 + 0.5 + 1.5 + 0.5 = 4$$

$$k_6 = \text{празно}_{-3} + \text{празно}_{-2} + \text{празно}_{-1} + \text{празно}_{+1} + \text{празно}_{+2} + \text{празно}_{+3}$$
$$= \left(0 * \frac{1}{0+0} + 0\right) + \left(0 * \frac{1}{0+0} + 0\right) + \left(0 * \frac{1}{0+0} + 0\right) + \left(0 * \frac{1}{0+0} + 0\right)$$
$$+ \left(0 * \frac{1}{0+0} + 0\right) + \left(0 * \frac{1}{0+0} + 0\right) = 0$$

Кај овие равенства се применети истите параметри со различни вредности во зависност од позицијата на зборовите. Секоја празна позиција е означена со црвена боја, што значи дека вредноста е еднаква на 0. Коефициентот $K_{\text{забава}}$ е добиен како збир од шест под коефициенти, а одлуката за избор на точниот двозначен збор во крајниот резултат ќе зависи од вредноста на другиот коефициент $K_{\text{жабава}}$. Наредните коефициенти се решаваат на истиот начин како коефициентот кој е опишан погоре. Сите вредности може да се преземат и да се пресметаат од следната Табела.

Табела 11. Резултат на параметрите за транслитерација според втората формула (Формула 2)
Table 11. Result from transliteration parameters after the second formula (Formula 2)

двозначен збор	**поим (лево, десно)**	**-3 (n_1)**	**-2 (n_2)**	**-1 (n_3)**	**+1 (n_4)**	**+2 (n_5)**	**+3 (n_6)**	**Повторувања**	k_i
забава	минатата	0	0	1	0	0	0	1	4
забава	на	0	1	0	0	0	0	1	4
забава	во	0	0	0	1	0	0	1	4
забава	Скопје	0	0	0	0	1	0	1	4
вешти	овие	0	1	1	0	0	0	2	-
вешти	Скопје	1	0	0	0	0	0	1	-
вешти	жени	1	0	0	3	0	0	4	-
вешти	им	0	0	0	0	1	0	1	-
вешти	помагаа	0	0	0	0	0	1	1	-
жени	вешти	0	0	3	0	0	1	4	-
жени	овие	0	1	0	0	0	0	1	-
жени	им	0	0	0	1	0	0	1	-
жени	помагаа	0	0	0	0	1	0	1	-
жени	на	1	0	0	1	0	1	3	-
вести	на	0	0	0	1	0	0	1	-

Во оваа табела се прикажани само четири резултати со иста вредност, бидејќи лево и десно од двозначниот збор се позиционирани поимите само по еднаш. Како пример да го земеме поимот „на“. Тој поим се наоѓа само на -2 позиција и за него автоматски $n_2 = 1$. Сега таа вредност кога ќе се собери со 0.5 и плус нултите вредности од позициите -3, -2, 1, 2, 3 со додадениот параметар 0.5 на секој од нив, се добива вредност 4 за даденият под коефициент k_2. Последната формула применува n_i како вкупно повторување на секој збор во зависност од дадената позиција (-3 -2 -1 +1 +2 +3) во моментално разгледуваната реченица, чии вредности се добиени од сите досега прочитани реченици. Во продолжение е прикажано равенството за добивање на коефициентите $K_{\text{вести}}$ и $K_{\text{вешти}}$.

$$K_{\text{вешти}} = \text{Скопје}_{-3} + \text{празно}_{-2} + \text{овие}_{-1} + \text{жени}_{+1} + \text{им}_{+2} + \text{помагаа}_{+3}$$
$$= \left(0 * \frac{1}{13+10}\right) + 0 + \left(1 * \frac{1}{13+10}\right) + \left(3 * \frac{1}{13+10}\right) + \left(1 * \frac{1}{13+10}\right)$$
$$+ \left(1 * \frac{1}{13+10}\right) = 0 + 0 + 0.0434 + 0.1304 + 0.0434 + 0.0434 = 0.2606$$

$$K_{\text{вести}} = \text{Скопје}_{-3} + \text{празно}_{-2} + \text{овие}_{-1} + \text{жени}_{+1} + \text{им}_{+2} + \text{помагаа}_{+3}$$
$$= 0 + 0 + 0 + 0 + 0 + 0 = 0$$

Овие равенства прикажуваат еден видлив резултат од зададените двозначни зборови. Тоа значи дека првиот двозначен збор „вешти“ е решение за латинско напишаниот двозначен поим “vesti“.

Во главната табела за Формула 3 од која се преземаат податоците, не се внесуваат дополнително двете колони како во претходните две табели, поради несоодветно прикажување на резултатите.

6 РЕЗУЛТАТИ ОД ИСТРАЖУВАЊАТА

Во текот на истражувањето на повеќезначната транслитерација, направени се повеќе тестови. Тие се однесуваат на секој специфичен дел, и тоа на два начина: рачно и автоматско. Рачно или мануелно тестирање е применето најмногу кај кратките форми од PHP скриптите и во алгоритамот за транслитерација. Тоа се применува при креирањето на почетна верзија на некоја скрипта, а таков е примерот за подготовка и споредба на:

- транслитериран и не транслитериран поим,
- двата коефициента за двозначниот поим,
- реченици со случаен избор итн.

6.1 Автоматско тестирање на цели реченици со примена на формули

Автоматското тестирање секогаш се стартува со претходно креирана PHP скрипта од која, во зависност од зададеното барање, се добиваат брзи резултати. Такво тестирање најмногу е направено кај речениците со случаен избор, односно на веќепрочитаните реченици од PDF документи кои се внесени во база на податоци. Тестирани се 300 реченици од вкупно 36.746 двозначни реченици, кои се поделени во неколку групи и тоа:

- реченици со најголемо повторување на двозначни зборови,
- реченици со средно повторување и
- реченици со најмало повторување.

Овие повторувања се однесуваат на застапеноста на двозначните зборови кои се зачувани во база на отворениот речник (OP). Реченициите со најголемо повторување се наоѓаат табеларно во база на редна позиција од 1-100, средните од 101-200, и најмалите на позиција од 201-300. Тестирањето во овој случај претставува транслитерација на цели реченици со повеќезначни поими на кирилица, што значи дека оригиналните реченици кои се тестирани се на кирилица. Транслитерацијата не се однесува на реченици со латинско писмо, бидејќи процентот на веќе пронајдени двозначни поими е помало од 100%. Овој процент не е исполнет максимално, бидејќи застапеноста на двозначните зборови во македонските текстови е значително мала или скоро да не постои. После неколку математички анализи и постапки за приближно добивање на точен резултат при решавање на повеќезначна транслитерација на цели реченици од латиница во кирилица, се креирани три различни формули, кои се објаснети во претходните точки. Секоја формула посебно е тестирана на 300 реченици со помош на алгоритамот за транслитерација. Добиени се три различни резултати и тие се опишани во следните потточки.

Резултат од тестираните реченици со примена на Формула 1 и OP

Овој резултат е добиен според Формула 1 за транслитерација на цели реченици, односно формулата во која n_i претставува *Вкупно повторување на секој збор посебно без разлика на која позиција се наоѓа од двозначниот поим, во сите досега прочитани реченици*. Тоа е прикажано со следната Табела.

Табела 12. Приказ на $\boldsymbol{n_i}$ во Формула 1
Table 12. Overview of $\boldsymbol{n_i}$ in Formula 1

двозначен збор	поим (лево, десно)	-3	-2	-1	+1	+2	+3	Повторувања (n_i)	$\frac{1}{N+M}$	$n_i * \frac{1}{N+M} + 0.5$
-	-	-	-	-	-	-	-	-	-	-

После тестирањето на сите групи на реченици е добиен резултат со позитивни и негативни вредности. Позитивните се однесуваат на бројот на реченици кои успешно поминале низ алгоритамот за транслитерација. Тоа значи дека споредувањето на оригиналната и транслитерираната реченица е еднакво на 1. Бројот на такви реченици изнесува за група 1 (↓1 негативна, ↑99 позитивни), за група 2 (↓3 негативни, ↑97 позитивни) и за група 3 (↓5 негативни, ↑95 позитивни).

Ова е прикажано во следната табела (Табела 13), а визуелно на (Слика 27. Прилог 1).

Табела 13. Негативни резултати добиени со примена на Формула 1
Table 13. Negatives results obtained by application of Formula 1

Групи / Повторување	**Реченици за тест**	**Негативни [реден бр.]**
Група 1 / најголемо	1-100	1 [60]
Група 2 / средно	101-200	3 [119, 151, 189]
Група 3 / најмало	201-300	5 [201, 203, 224, 229, 285]
	Вкупно	9

Резултат од тестираните реченици со примена на Формула 2 и ОР

При тестирање на истите реченици е добиен и вториот резултат каде е употребена Формула 2, односно формулата во која n_i претставува *Вкупно повторување на секој збор посебно во зависност од дадената позиција (-3 -2 -1 +1 +2 +3), во сите досега прочитани реченици*. Тоа е прикажано со следната Табела.

Табела 14. Приказ на $\boldsymbol{n_i}$ во Формула 2
Table 14. Overview of $\boldsymbol{n_i}$ in Formula 2

двозначен збор	поим (лево, десно)	-3 (n_1)	-2 (n_2)	-1 (n_3)	+1 (n_4)	+2 (n_5)	+3 (n_6)	Повторувања	k_i
-	-	-	-	-	-	-	-	-	-

Бидејќи применуваме друга формула автоматски се добија и други вредности кои се приближно исти со првиот резултат (Формула 1). Тоа се однесува на сите три групи во кои имаме позитивни и негативни вредности за реченициите за транслитерација. Бројот на таквите реченици изнесува: за група 1 (↓1 негативна, ↑99 позитивни), за група 2 (↓3 негативни, ↑97 позитивни) и за група 3 (↓4 негативни, ↑96 позитивни). Ова е прикажано во следната табела (Табела 15), а визуелно на (Слика 28. Прилог 2).

Табела 15. Негативни резултати добиени со примена на Формула 2
Table 15. Negatives results obtained by application of Formula 2

Групи / Повторување	**Реченици за тест**	**Негативни [реден бр.]**
Група 1 / најголемо	1-100	1 [60]
Група 2 / средно	101-200	3 [119, 151, 189]
Група 3 / најмало	201-300	4 [201, 203, 229, 285]
	Вкупно	8

Резултат од тестираните реченици со примена на Формула 3 и ОР

Овој резултат е малку поразличен од претходните два, бидејќи е употребена трета формула каде n_i претставува *Вкупно повторување на секој збор во зависност од дадената позиција (-3 -2 -1 +1 +2 +3) во моментално разгледуваната реченица, чии вредности се добиени од сите досега прочитани реченици*.

Табела 16. Приказ на $\boldsymbol{n_i}$ во Формула 3
Table 16. Overview of $\boldsymbol{n_i}$ in Formula 3

двозначен збор	поим (лево, десно)	-3 (n_1)	-2 (n_2)	-1 (n_3)	+1 (n_4)	+2 (n_5)	+3 (n_6)	Повторувања	$\frac{1}{N+M}$	$n_i * \frac{1}{N+M} + 0.5$
-	-	-	-	-	-	-	-	-	-	-

Овде се добиени исто така позитивни и негативни вредности за реченициите кои се тестирани, што значи дека е добиен резултат од различни групи на реченици. Бројот на таквите реченици изнесува за група 1 (↓1 негативна, ↑99 позитивни), за група 2 (↓4 негативни, ↑96 позитивни) и за група 3 (↓6 негативни, ↑94 позитивни). Ова е прикажано во следната табела (Табела 17), а визуелно на (Слика 29. Прилог 3).

Табела 17. Негативни резултати добиени со примена на Формула 3
Table 17. Negatives results obtained by application of Formula 3

Групи / Повторување	**Реченици за тест**	**Негативни [реден бр.]**
Група 1 / најголемо	1-100	1 [60]
Група 2 / средно	101-200	4 [119, 151, 159, 189]
Група 3 / најмало	201-300	6 [201, 229, 232, 233, 275, 285]
	Вкупно	11

Резултат од тестираните реченици со примена на Формула 2 и ново креираниот речник (комбиниран плус речник, КПР)

Досега се добиени три резултати во кои е применет само ОР. Тоа е така земено, бидејќи процентот на добиените двозначни зборови, во споредба со ТР и КР, е најголем. Сите досега добиени резултати имаат по неколку негативни реченици. Со цел да се намалат овие негативности направена е уште една тест база на податоци наречена комбиниран плус речник или КПР.

Овде се применува комбинираниот речник со вкупно 258883 фонд на зборови. Затоа е креирана скрипта која ќе има две можности и тоа за:

- читање на PDF содржини (двозначни зборови итн.);
- внесување на нови зборови кои не постојат во базата на податоци.

Новопрочитаните поими може да бидат на латиница и кирилица, со исклучок на знаци и броеви. Во зависност од тоа дали тие се на позициите пред и после двозначниот збор, ќе зависи и внесувањето на нови зборови во база на податоци. Со новата скрипта се добиени плус 99598 поими, и ако тие се соберат со почетната состојба ќе се добијат вкупно 358481 поими. Исто така и овде се тестирани истите реченици на кирилица како во претходните 3 тестирања.

Овде е употребена формулата во која n_i претставува *Вкупно повторување на секој збор посебно во зависност од дадената позиција (-3 -2 -1 +1 +2 +3), во сите досега прочитани реченици.* После тестирањето на 300 кирилски реченици, фондот на поими иако е зголемен, сепак бројот на негативните реченици не е намален. Причина за тоа се новите зборови кои се внесени во база на податоци.

Како позитивна страна се покажа резултатот за кој се тестирани истите реченици, но сега тие се напишани на латиница. После транслитерацијата, бројот на негативните реченици овде се намалува поради новите додадени зборови кои се на латиница.

Одовде може да се заклучи дека со новокреираната база на податоци, позитивни резултати може да се добијат, доколку оригиналната реченица за транслитерација е на латиница. За нашиот веб сервис тоа е во предност, бидејќи неговата основна функција е транслитерација на цели реченици од латиница во кирилица. Резултатите од тестирањето се прикажани во Табела 18.

Табела 18. Негативни резултати добиени со примена на Формула 2 и додавање на нови поими во КР
Table 18. Negatives results obtained by application of Formula 2 and adding new terms in combined dictionary

Групи / Повторување	**Реченици за тест**	**Негативни при транслитерација на кирилски реченици**	**Негативни при транслитерација на латински реченици**
Група 1 / најголемо	1-100	4	6
Група 2 / средно	101-200	13	16
Група 3 / најмало	201-300	17	23
	Вкупно	34	45

Прашања:

Зошто се појавуваат повеќе негативни резултати за разлика од претходните[14], при транслитерација на кирилски реченици, иако имаме поголем фонд на зборови?

Зошто се појавуваат помалку негативни резултати за разлика од претходните, при транслитерација на реченици со латинско писмо?

Одговорите на овие прашања се објаснуваат со следниот пример.

Да го земеме зборот “kvalitativnite” кој е прочитан од PDF документите и проверен во база како не постоечки, каде автоматски е внесен како нов запис во главната табела со зборови во колоната во која внесуваат зборови на кирилица. Потоа тој збор е транслитериран во латиница и е внесен во друга колона (latinica). Тоа значи дека прочитаниот збор “kvalitativnite” останува ист и после транслитерацијата.

Наредниот збор кој е прочитан во PDF документите е „квалитативните” и тој исто така не се наоѓа во базата на податоци, односно во колоната (zbor), па затоа е внесен како нов запис. После транслитерацијата во латиница е добиен зборот “kvalitativnite”, којшто исто така е запишан како нов запис во другата колона (latinica) од табелата за зборови.

[14] (Формула 1, Формула 2, Формула 3)

Сега постојат два различни поими во колоната (zbor), но со иста вредност во колона (latinica).

id	zbor	latinica
264859	kvalitativnite	kvalitativnite
281273	квалитативните	kvalitativnite

Тоа значи дека, кога се проверува зборот “kvalitativnite” во колона (latinica), скриптата автоматски ја зема вредноста од колоната (zbor), и ја става во крајното решение на тестираната реченица.

Една од 300-те реченици за тестирање во која се појавува зборот „квалитативните”, односно “kvalitativnite”, е следната:

Оригинална реченица = „Да се картира некоја појава значи да се претстави нејзината местоположба, просторниот распоред, квантитативните и квалитативните белези.”

Кога оваа реченица поминува низ чекорите во тест скриптите за транслитерација, се проверува секој збор од реченицата и како резултат се добива:

Транслитерирана реченица = „Да се картира некоја појава значи да се претстави нејзината местоположба, просторниот распоред, квантитативните и kvalitativnite белези.”

Од овде може да се забележи дека е добиена негативна вредност за зборот “kvalitativnite”, а позитивна за другите транслитерирани зборови. Ако истата реченица за тестирање е напишана на латиница, после транслитерирањето се добива крајното решение. Доколку се споредат оригиналната и транслитерираната реченица, ќе се добие повторно негативен резултат.

Оригинална реченица = “Da se kartira nekoja pojava znaci da se pretstavi nejzinata mestopolozba, prostorniot raspored, kvantitavnite I kvalitativnite belezi.“

Транслитерирана реченица = „Да се картира некоја појава значи да се претстави нејзината местоположба, просторниот распоред, квантитативните и kvalitativnite белези.”

За да се реши овој настанат проблем, целата табела со зборови треба рачно да се провери и дел од зборовите да се избришат. Бришењето треба да се однесува само на зборовите од македонскиот јазик, кои во колона (zbor) се напишани на латиница, додека другото останува исто.

6.2 Анализа на автоматско тестираните реченици

Анализата се однесува на првите три резултати кои се тестирани автоматски и за нив ќе направиме неколку споредби во делот на групирање на реченициte. Како што е споменато имаме на располагање 300 кирилски реченици и секоја од нив има свое значење. Тоа значење се однесува пред сè на двозначните зборови, но имаме и зависност од повторувањето на тие зборови во досега внесените реченици во база на податоци.

Секоја реченица може да биде различно изградена граматички, без разлика на специјалните знаци и бројки, и за секоја од нив постои групирање според фреквенцијата на двозначните зборови во база. Ако ги разгледаме подетално резултатите, ќе може да се забележи помала или поголема разлика во добиените вредности кои можат да бидат позитивни или негативни. Дали тестираните реченици се добиени позитивно на кирилица или латиница ќе зависи од тоа дали ги имаме сите двозначни поими внесено во база.

Со добиените резултати имаме добиени некои приближни решенија за транслитерација на цели реченици, но доколку во овој случај се тестираат реченици напишани на латиница негативните вредности на реченициte во крајниот резултат ќе се зголемат. Доколку се направи споредба на добиените резултати, ќе се добие графичка крива за секоја група на застапеност на двозначен збор, која на некои места од графикот, во иста точка, се поклопува со друга крива, што значи дека имаме исти вредности за дадените реченици.

Негативни резултати

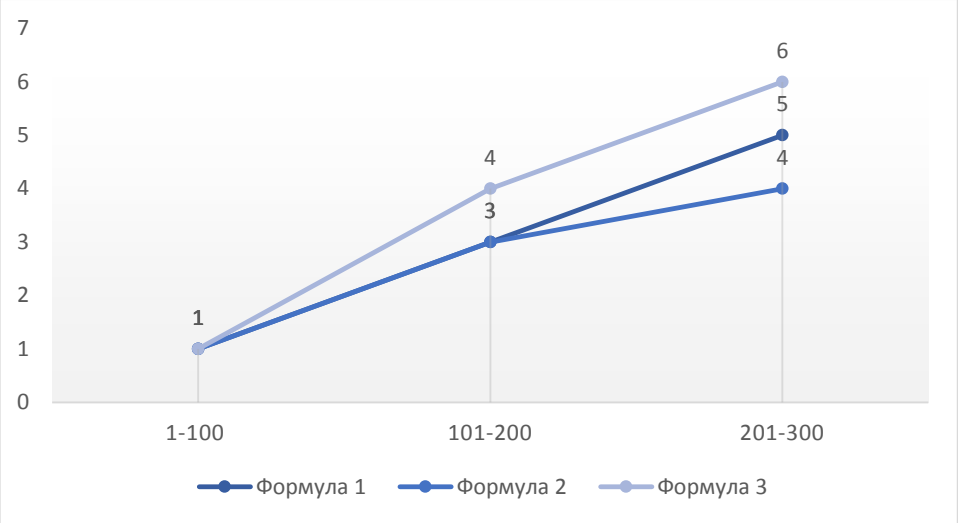
Првата негативна реченица која спаѓа во групата најголемо повторување е со реден број 60, и таа се појавува во сите претходно добиени резултати (Табелите 13, 15, 17). Таа има негативна вредност, бидејќи не може да се одреди точниот двозначен поим, а тоа е поради еднаквите вредности на коефициентите. Претходните 59 и наредните 40 реченици се успешно поминати низ алгоритамот за транслитерација, преку сите три формули, без никаква грешка. Сите овие реченици спаѓаат во првата група.

Следни еднакви грешки кај трите резултати се појавуваат на 119, 151 и 189 реченица, поради несоодветно добиените двозначни зборови. Дополнително имаме уште една негативна вредност за 159 реченица, но овој пат само во третиот резултат (Табела 17). Овие реченици спаѓаат во делот на втората група.

Во последната група ја имаме 201 реченица која се повторува во сите три резултати; потоа 203, 224, 229 и 285 реченица со Формула 1; 203, 229 и 285 со Формула 2 и на крај е 229, 232, 233, 275, 285 со Формула 3. Сите овие вредности се прикажани во следната Табела.

Табела 19. Споредување на негативните реченици од претходно добиените негативни резултати
Table 19. Comparison of the negative sentences from previously obtained negative results

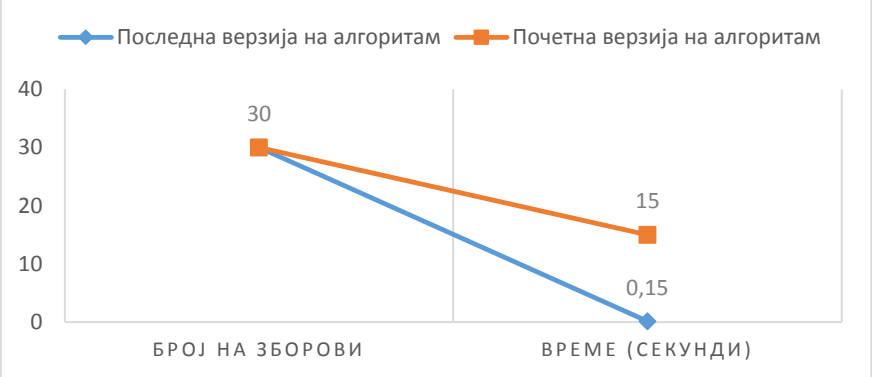
Повторување	Формула 1	Формула 2	Формула 3	Повторување
најголемо	1 [**60**]	1 [**60**]	1 [**60**]	[60]
средно	3 [**119**, **151**, **189**]	3 [**119**, **151**, **189**]	4 [**119**, **151**, 159, **189**]	[119, 151, 189]
најмало	5 [**201**, 203, 224, **229**, **285**]	4 [**201**, 203, **229**, **285**]	6 [**201**, **229**, 232, 233, 275, **285**]	[201, 229, 285]
Негативни	9	8	11	7



Слика 12. Графички приказ на негативните вредности од 300 реченици при користење на три различни формули
Figure 12. Graphical display of the negative values in 300 sentences using three different formulas

6.3 Анализа на временскиот интервал на алгоритамот

Алгоритамот за транслитерација е програмиран и дизајниран со технологиите PHP, MySQL, jQuery, HTML, за креирање на скрипти и други дигитални ресурси. За краток временски период е направена почетната верзија на алгоритамот за транслитерација на цели реченици. Тоа е првиот експериментален обид од кој се добија позитивни резултати во делот на семантиката, а негативни за временскиот интервал на извршување на алгоритамот. Веќе во вториот обид се промени состојбата во делот на времето, но не и на семантиката. Состојбата на тие експериментални обиди може да се види на следниот дијаграм.



Слика 13. Временски интервал на алгоритамот при изведување на ист процес (почетен, краен)
Figure 13. Time interval for operating the same process in algorithm (begin, end)

Дијаграмот ни прикажува една состојба за алгоритамот, каде што ист процес, се извршува за различен временски интервал. Тоа е така поради не оптимизираниот код и базата со податоци. Двете верзии на алгоритамот се однесуваат на различни транслитерирани реченици со просек околу 30 зборови за секоја од нив. Почетната верзија започнува со користење на база во која атрибутите од табелите не се индексирани, и затоа кога се транслитерираат реченициите, се добива побавно крајниот резултат, односно 15 секунди за секоја реченица одделно. После неколку експериментални обиди е добиена и последната верзија на алгоритамот каде е значително подобрена, што значи дека со додаденото индексирање е добиен 99% побрз резултат во споредба од почетна верзија. Сите резултати се добиени со многубројни тестирања преку специјално изработени скрипти. За да се добие точен резултат потребно е оригиналната реченица да се спореди со транслитерираната реченица, односно тие две да бидат еднакви. Еднаквоста се однесува на точност и преклопување во делот на празни места, специјални знаци, бројки итн.

На следниот дијаграм (Слика 14) е прикажано работењето на алгоритамот за транслитерација, кој е поставен на Linux сервер. Во нашиот случај имаме две почетни вредности, едната е за време, додека другата за број на содржина.

1. Во првиот случај алгоритамот стартува во 09:46h, а завршува во 09:51h. За време од 5 минути се прочитани и транслитерирани 3267 зборови од еден PDF документ. Почетната состојба на зборовите е 2.834.584, а после поминатото време се зголемува на 2.837.851 зборови.

2. Во вториот случај процесот на алгоритамот се извршува 5 часа и 16 минути, и за тоа време се прочитани вкупно 206737 зборови. Тоа значи дека бројот на зборовите на почеток започнува со 2.837.851 кој после поминатото време се зголемува на 3.044.588.

Ако се направи просек на овие вредности, ќе се добие бројка од 10 транслитерирани зборови за време од една секунда.

Слика 14. Резултат на прочитани и транслитерирани зборови во зависност од времето
Figure 14. Result from transliterated words depending on the time

7 ФУНКЦИОНАЛНИ ОСОБИНИ НА АЛГОРИТАМОТ ЗА ТРАНСЛИТЕРАЦИЈА

Врз основа на водечките технологии, алгоритамот за повеќезначна транслитерација на цели реченици овозможува пребарување, конвертирање, изведување на проверки на знаци и бројки, и транслитерација на двозначни поими. Тој може да работи со точност од 98% преку база на податоци фокусирани на глаголи, придавки, сврзници и др. Неодредеността и неточноста на алгоритамот се коригира со бројот на прочитани документи, реченици и зборови.

Во ова поглавје е даден детален опис на алгоритамот за транслитерација со примери за:

- комбинирани букви од македонската азбука;
- добивање на транслитерирани цели реченици од латиница во кирилица во кои е одредено значењето на зборовите;
- функција за ограничување на зборовите;
- ER дијаграм.

7.1 Дефиниција на корпус

Корпус е збирка на јазични текстови во електронска форма, избрани според надворешните критериуми, за да го застапуваат јазикот како извор на податоци за лингвистички истражувања. (John Sinclair, 1933-2007).

Тој е оптимизиран за пребарување и анализа на одредени зборови или фрази кои се наоѓаат во база на податоци. Корпус може да содржи текстови од еден јазик (еднојазичен корпус) или текст на податоци на повеќе јазици (повеќејазичен корпус). Во текстови спаѓаат книги, списанија итн.

Големината на корпусот зависи од потребата за која е наменет. Постојат илјадници корпуси во светот, но повеќето од нив се креирани за специјални истражувања и не се достапни во јавноста.

Корпус за нашето истражување претставува множество од македонско кирилски зборови напишани во дигитална форма во база на податоци. Тие се преземени од различни извори на интернет и служат за одредени анализи и тестирања.

Прашања кои треба да се одговорат:

- Како да се измери основната фреквенција на појавувањето на зборовите?

- Како да се нормализираат податоците што сакате да ги анализирате?

- Како да се измери односот помеѓу зборовите и фразите во корпусот?

Одговорот е да се користи статистика и веројатност. Анализа се врши со помош на компјутер, односно со специјални скрипти. Во тоа е предноста, бидејќи за неколку секунди истражувачот добива информации. Ако тоа мора рачно да се анализира, потребни се неколку часа или денови. Во текот на истражувањето со помош на скриптите, од 300 прочитани документи се избројани околу 4,5 милиони зборови појавени во 300.000 реченици.

Со цел да се обучува алгоритамот да класифицира и да бележи елементи во текстот, треба да се знае природата на корпусот.

7.2 Транслитерација со комбинирање на знаци

Алгоритамот за транслитерација се состои од повеќе делови за добивање на правилната реченица. Од 31 буква од македонската азбука само 5 букви (s, k, z, c, g) можат да се најдат во ситуација на двозначност кога тие ќе се напишат во латинско писмо. Тоа значи дека алгоритамот ќе мора да прави комбинации со овие букви кои се составен дел во една реченица. Доколку се транслитерира зборот “transliteracija”, сега може само буквите (t, r, a, n, l, i, t, e, r, a, i, j, a) да се конвертираат директно во кирилско писмо, додека останатите букви (s, c) се комбинираат во (с, ц ; ш, ц ; с, ч ; ш, ч). Кога ќе се добие правилниот збор, кој се проверува во базата на податоци, се оди на наредниот поим од реченицата. Во нашиот алгоритам се земени во предвид максимум 20 комбинирани букви од (s, k, z, c, g) кои може да се наоѓаат во еден поим од реченицата. Тоа значи дека ќе може секој од овие букви да се повторуваат по 4 пати во еден збор. Следува приказ на *Табела 7* со сите комбинации кои може да се применат при транслитерација на цели реченици.

Табела 20. Максимално комбинирани букви за еден двозначен поим напишан на латиница
Table 20. Maximum of combined letters for one word with two meanings written on Latin alphabet

Буква	**Бр. на букви**	**Бр. на комбинации**	**Комбинации**
s (с, ш)	1	2	(с) ; (ш)
	2	4	(с , с) ; (с , ш) ; (ш , с) ; (ш ,ш)
	3	8	(с , с , с) ; (с , с , ш) ; (с , ш , с) ; (ш ,с , с) ; (с , ш , ш) ; (ш , с , ш) ; (ш , ш , с) ; (ш ,ш , ш)
	4	15	(с , с , с , с) ; (с , с , с , ш) ; (с , с , ш , с) ; (с , ш , с , с) ; (ш , с , с ,с) ; (с , с , ш , ш) ; (с , ш , ш , с) ; (ш , ш , с , с) ; (с , ш , с , ш) ; (ш , с , ш , с) ; (с , ш , ш , ш) ; (ш , с , ш , ш) ; (ш , ш , с , ш) ; (ш , ш , ш , с) ; (ш , ш , ш , ш)
k (к, ќ)	1	2	(к) ; (ќ)
	2	4	(к , к) ; (к , ќ) ; (ќ , к) ; (ќ ,ќ)
	3	8	(к , к , к) ; (к , к , ќ) ; (к , ќ , к) ; (ќ ,к , к) ; (к , ќ , ќ) ; (ќ , к , ќ) ; (ќ , ќ , к) ; (ќ ,ќ , ќ)
	4	15	(к , к , к , к) ; (к , к , к , ќ) ; (к , к , ќ , к) ; (к , ќ , к , к) ; (ќ , к , к ,к) ; (к , к , ќ , ќ) ; (к , ќ , ќ , к) ; (ќ , к , ќ , к) ; (к , ќ , к , ќ) ; (ќ , к , ќ , ќ) ; (к , ќ , ќ , ќ) ; (ќ , ќ , к , ќ) ; (ќ , ќ , ќ , к) ; (ќ , ќ , ќ , ќ)
z (з, ж)	1	2	(з) ; (ж)
	2	4	(з , з) ; (з , ж) ; (ж , з) ; (ж ,ж)
	3	8	(з , з , з) ; (з , з , ж) ; (з , ж , з) ; (ж ,з , з) ; (з , ж , ж) ; (ж , з , ж) ; (ж , ж , з) ; (ж ,ж , ж)

	4	15	(з,з,з,з);(з,з,з,ж);(з,з,ж,з); (з,ж,з,з);(ж,з,з,з);(з,з,ж,ж); (з,ж,ж,з);(ж,ж,з,з);(з,ж,з,ж); (ж,з,ж,з);(з,ж,ж,ж);(ж,з,ж,ж); (ж,ж,з,ж);(ж,ж,ж,з);(ж,ж,ж,ж)
с (ц, ч)	1	2	(ц);(ч)
	2	4	(ц,ц);(ц,ч);(ч,ц);(ч,ч)
	3	8	(ц,ц,ц);(ц,ц,ч);(ц,ч,ц);(ч,ц,ц); (ц,ч,ч);(ч,ц,ч);(ч,ч,ц);(ч,ч,ч)
	4	15	(ц,ц,ц,ц);(ц,ц,ц,ч);(ц,ц,ч,ц); (ц,ч,ц,ц);(ч,ц,ц,ц);(ц,ц,ч,ч); (ц,ч,ч,ц);(ч,ч,ц,ц);(ц,ч,ц,ч); (ч,ц,ч,ц);(ц,ч,ч,ч);(ч,ц,ч,ч); (ч,ч,ц,ч);(ч,ч,ч,ц);(ч,ч,ч,ч)
g (г, ѓ)	1	2	(г);(ѓ)
	2	4	(г,г);(г,ѓ);(ѓ,г);(ѓ,ѓ)
	3	8	(г,г,г);(г,г,ѓ);(г,ѓ,г);(ѓ,г,г); (г,ѓ,ѓ);(ѓ,г,ѓ);(ѓ,ѓ,г);(ѓ,ѓ,ѓ)
	4	15	(г,г,г,г);(г,г,г,ѓ);(г,г,ѓ,г); (г,ѓ,г,г);(ѓ,г,г,г);(г,ѓ,ѓ,ѓ); (г,ѓ,ѓ,г);(ѓ,ѓ,г,г);(г,ѓ,г,ѓ); (ѓ,г,ѓ,г);(г,ѓ,ѓ,ѓ);(ѓ,г,ѓ,ѓ); (ѓ,ѓ,г,ѓ);(ѓ,ѓ,ѓ,г);(ѓ,ѓ,ѓ,ѓ)

Направена е една анализа на поими од македонскиот јазик и како резултат се добиени мал број на поими кои содржат повеќе од 3 исти знаци напишани на латиница кои при транслитерација се конвертираат во кирилско писмо. Тоа е прикажано во следната Табела.

Табела 21. Максимално содржани знаци во поими напишани на латиница
Table 21. Maximum contained characters of words written on Latin alphabet

Поим на латиница	**Бр. на исти знаци во еден поим**	**Транслитерација**
soslusas	4 (s)	сослушаш
soslusuvas	4 (s)	сослушуваш
sustestvenost	4 (s)	суштественост
skokotkajki	4 (k)	скокоткајќи
celicarnici	3 (c)	челичарници
karakterizirajki	3 (k)	карактеризирајќи

Транслитерацијата во овој случај прави конвертирање на кирилските двозначни знаци кои се дел од еден збор, додека зборот може да се повторува повеќе од еден пат во самата реченица. Колку повеќе двозначни зборови се наоѓаат во една реченица, толку алгоритамот за транслитерација има повеќе процеси за обработка на дадена реченица. Ако се земе на пример зборот

“soslusas” веднаш се забележуваат знаците кои може да се комбинираат за да се најде точното значење при транслитерација. Таквиот карактер е латинската буква “s” и тука се појавува 4 пати, но ако погледнеме во Табела 20, од неа може да се добијат 15 различни комбинации од зборот (soslusas), но само еден збор може да е точен, а тоа е кирилскиот збор „сослушаш”.

Тоа значи промена на латинските знаци кои може директно или преку процес на транслитерација да се трансформираат во кирилица. Кога станува збор за директна транслитерација тоа се однесува на зборовите кои немаат повеќезначност како на пример „круг, музика, деца итн.”, со што може лесно и директно да се добие соодветниот поим, но под услов да се наоѓа во базата на податоци. Доколку тој поим не постои во базата на податоци, во тој случај останува непроменет или не транслитериран. Кога станува збор за базата, во неа се наоѓаат неколку табели со податоци од кои може лесно да се препознае двозначниот збор. Со тоа ќе може процесот на транслитерација истовремено да се поврзе со тие табели и да ги добива сите податоци од нив.

7.3 Транслитерација со одредување на значењето на зборовите во целите реченици

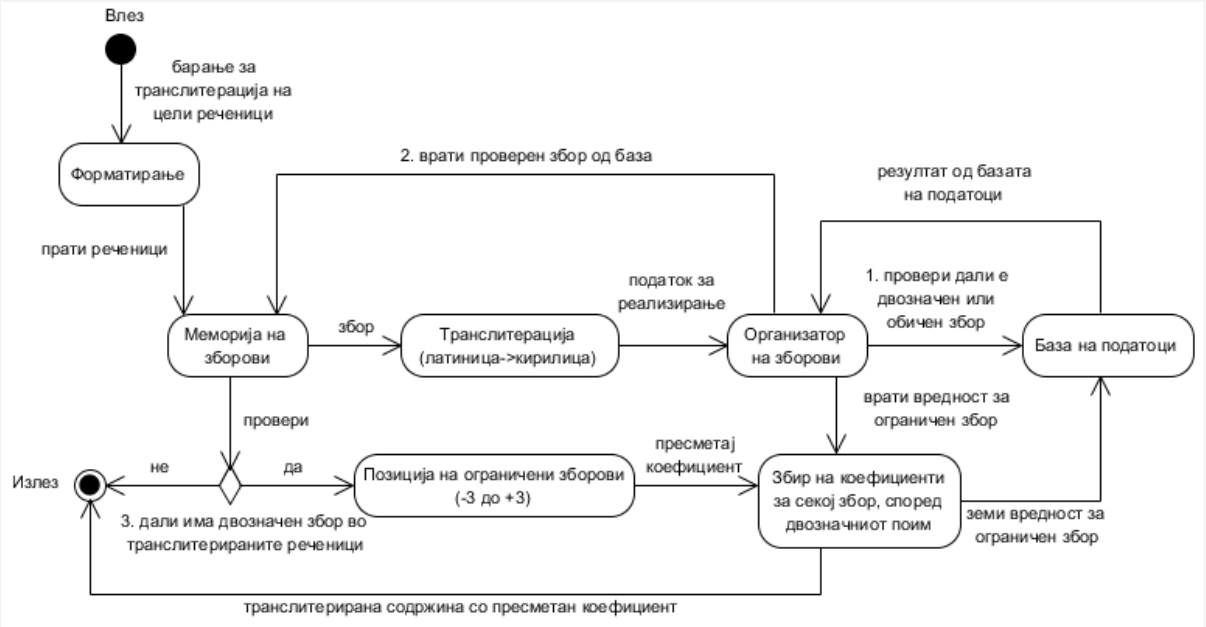
Значењето на зборовите се одредува преку алгоритам за повеќезначна транслитерација. Претходниот алгоритам работеше врз основа на цели реченици[15] кои беа целосно внесени во база. При секое транслитерирање на бараната содржина од страна на корисникот, алгоритамот имаше за цел да ги провери сите зборови во сите реченици, и на крај да пресмета соодветен коефициент[16]. Ефектот за добивање на брз резултат на овој начин се покажа негативен.

За да се овозможи подобар и побрз пристап до податоците креирани се дополнителни табели во кои сега наместо цели реченици се зачувуваат само зборовите кои се наоѓаат пред и после двозначниот збор. Во табелите, на тие зборови им се доделени вредности: за позициите, за повторувањата на поимите како и за добиените коефициенти.

[15] Над 2000 реченици

[16] Коефициент за одредување на точниот двозначен збор со помош на одредена формула

Ова е веќе реализирано, а како решение за тоа се примерите кои се прикажани во претходните точки. Секоја креирана табела во база на податоци е спремна за праќање на информации до алгоритамот кој може лесно да одреди кои од зададените двозначни поими кои се транслитерираат од латиница во кирилица треба да се стават во значењето на една реченица. Едно такво одредување на значењето на зборовите може да се прикажи со следниот дијаграм (Слика 15).



Слика 15. Добивање на транслитерирани цели реченици од латиница во кирилица, во кои е одредено значењето на зборовите
Figure 15. Obtaining transliterated full sentences from Latin to Cyrillic alphabet, with determined meaning of the words

Кога станува збор за неодреденост на реченициите, се однесува на транслитерирани зборови од латиница во кирилица, во кои се наоѓаат и неопределени двозначни поими. Таков пример може да се види од следната реченица:

- „На минатата [жабава] [забава] во Скопје овие [вести] [вешти] [жени] [зени] им помагаа на ...”

Оваа реченица преку процесот за транслитерација, со помош на комбинираните зборови кои се проверуваат директно во база, не може целосно да се определи значењето на целите реченици, бидејќи двозначните поими кои

се наоѓаат во единични загради, треба да се одредат преку збирот на коефициентите на сите зборови од целата реченица.

Овој веб сервис овозможува кориснички интерфејс во кој се поставува барањето за повеќезначна транслитерација на цели реченици од латиница во кирилица. Обемот на реченициите за транслитерација е ограничен, па затоа меморијата за привремено зачувување на зборови може добро да функционира во вакви услови. Меморија на зборови се однесува на нетранслитерирани зборови кои треба се обработат и на зборови кои се веќе спремни за одредување на значењето на реченициите. Секој збор треба да се праќа од меморијата до делот за транслитерација од латиница во кирилица. Овде се транслитерираат само оние знаци кои се напишано во стандардни услови, а останатите знаци се препуштаат на наредните чекори за транслитерација. Организаторот на зборови овозможува: зборовите за транслитерација да се проверат во базата на податоци, да преминат директно во меморија или пак да се комбинираат преостанатите знаци од зборовите, после транслитерацијата на знаци во стандардни услови. На крај се пресметува коефициентот за да може лесно да се одреди значењето на целата реченица. Овој чекор се овозможува доколку во содржината за транслитерација се наоѓа барем еден двозначен поим. Затоа меморијата на зборови дава можност целата транслитерирана содржина заедно со неопределените двозначни зборови, преку позиционирање на зборови кои се 3 места пред и 3 после, да се земат вредности кои претходно се зачувани во база на податоци. Преку тие вредности со одредена формула се добива крајниот коефициент, а со тоа се одредува и двозначниот збор, кој е дел од транслитерираната содржина. Доколку транслитерираната содржина не содржи двозначни зборови последниот чекор за позиција на ограничени зборови не се користи, па затоа се преминува директно на излез од алгоритамот за транслитерација.

7.4 Функции на алгоритамот за повеќезначна транслитерација

Секој алгоритам своите процеси може да ги управува со програмски код напишан со услови, циклуси, случаи и поврзана база. Ваквите претставувања на програмскиот код може, но и не мора да се применат во зададените функции. Нашиот алгоритам работи со помош на функции, што значи директно може се

вклучат во главниот код. Пример за една таква функција која може да ги ограничува зборовите е прикажана на следната слика (Слика 16).

```
function ogranicuvanje_na_zborovite($recenica, $dvoznacen_zbor){
    mb_internal_encoding("UTF-8");

     $rezultat = "";
     $recenica_od_internet = explode(" ", mb_strtolower($recenica));
     $pozicija = array_search($dvoznacen_zbor, $recenica_od_internet);

     $brojac_na_zborovi = count($recenica_od_internet);
     $razlika = $brojac_na_zborovi - $pozicija;

     if($pozicija > 2){$pozicija_levo = $pozicija-3;} else {$pozicija_levo = 0;}
     if($razlika >= 3){$pozicija_desno = $pozicija+4;} else {$pozicija_desno = $brojac_na_zborovi;}

     if(!empty($pozicija)){
       for($i = $pozicija_levo; $i < $pozicija; $i++){ $rezultat .= $recenica_od_internet[$i].' '; }
       for($j = $pozicija + 1; $j < $pozicija_desno; $j++){ $rezultat .= $recenica_od_internet[$j].' '; }
     }
     else{
       for($i = $pozicija_levo; $i < $pozicija; $i++){ $rezultat .= $recenica_od_internet[$i].' '; }
       for($j = $pozicija + 1; $j < $pozicija_desno; $j++){ $rezultat .= $recenica_od_internet[$j].' '; }

     }
     return $rezultat;
   }
```

Слика 16. Функција за ограничување на зборовите кај повеќезначна транслитерација со Формула 1 или Формула 2
Figure 16. Function of limitation in words with ambiguous transliteration with Formula 1 and Formula 2

Бидејќи функциите придонесуваат олеснување и разбирливост во еден програмски код, па затоа во нашиот алгоритам е применета една од повеќето функции, а тоа е функцијата за ограничување. Со неа се овозможува определување на зборови кои според правилото на позиција се наоѓаат три места пред и после двозначниот збор. Со оглед на тоа што секоја PHP команда работи со латинска поддршка, во оваа функција е употребена опцијата за декодирање UTF-8, со која се овозможува пребарување на кирилски двозначни поими во база на податоци. Оваа функција е многу важна во некои делови од алгоритамот, бидејќи може да ги филтрира сите непотребни зборови во целите реченици за транслитерација, што значи дека времето на работење на алгоритамот е оптимално искористено. Оваа функција е употребена најмногу во алгоритамот за транслитерација каде се применува Формула 1 или Формула 2.

Во текот на истражувањето за примена на Формула 3 е креирана исто така една функција, (Слика 17) која ги содржи истите параметри како во претходната функција, но различно дефинирана во условите.

```
function ogranicuvanje_na_zborovite($recenica, $dvoznacen_zbor){
    mb_internal_encoding("UTF-8");

     $rezultat = "";
     $recenica_od_internet = explode(" ", mb_strtolower($recenica));
     $pozicija = array_search($dvoznacen_zbor, $recenica_od_internet);

     $brojac_na_zborovi = count($recenica_od_internet);
     $razlika = $brojac_na_zborovi - $pozicija;

     if($pozicija > 2){$pozicija_levo = $pozicija-3;} else {$pozicija_levo = 0;}
     if($razlika >= 3){$pozicija_desno = $pozicija+4;} else {$pozicija_desno = $brojac_na_zborovi;}

     if(!empty($pozicija)){
      for($i = $pozicija_levo; $i < $pozicija; $i++){
                    if($recenica_od_internet[$i]!='.' and $recenica_od_internet[$i]!='')
                          $rezultat .= $recenica_od_internet[$i].'-'.($pozicija-$i).' ';
      }
      for($j = $pozicija + 1; $j < $pozicija_desno; $j++){
                    if($recenica_od_internet[$j]!='.' and   $recenica_od_internet[$j]!='')
                          $rezultat .= $recenica_od_internet[$j].'+'.($j-$pozicija).' ';
      }
     }
     else{
      for($i = $pozicija_levo; $i < $pozicija; $i++){ $rezultat .= $recenica_od_internet[$i]."+$i "; }
      for($j = $pozicija + 1; $j < $pozicija_desno; $j++){
                    if($recenica_od_internet[$j]!='.' and $recenica_od_internet[$j]!='')
                          $rezultat .= $recenica_od_internet[$j]."+$j ";
           }

     }

     return $rezultat;

}
```

Слика 17. Функција за ограничување на зборовите кај повеќезначна транслитерација со Формула 3

Figure 17. Function of limitation in words with ambiguous transliteration with Formula 3

Во оваа функција новите услови се најчесто за проверка на позицијата и за додавање ознака на секој збор во реченицата. Други најчесто користени функции кои можат повеќе од еднаш да се појават при процесот на алгоритамот за транслитерација се следниве:

ознака на функција	опис
najdi_zbor_vo_sodrzina1	пребарување на соодветниот збор
najdi_dvoznacen_zbor	пребарување на соодветниот двозначен збор
zameni_zagradi	одредување на правилниот двозначен збор
presmetaj_koeficient	одредување на коефициент од главната табела со резултати
vo_kir	конвертирање на буквите од латиница во кирилица
vo_lat	конвертирање на буквите од кирилица во латиница
znaci	форматирање 2 (специјални знаци)
znaci_nazad	форматирање 1 (специјални знаци)

7.5 Решавање на повеќезначна транслитерација со примена на база на податоци

Базите денес се основа на многу апликации. Се користат за чување на податоци, претставување на податоци пред клиентите на веб апликациите и како поддршка на други комерцијални процеси. Базите се користат и при многу научни истражувања. Нашиот веб сервис сите информации ги одобрува со помош на база на податоци. Секоја база има свои извори на кои може да се поставуваат упити. Тие извори се однесуваат на т.н. табели во кои се наоѓаат податоци, односно атрибути со сопствено име и тип. Секој од атрибутите најчесто се разликува според типот на податоци кој може да биде: цел број, децимален број, текст или пак според клучот кој може да биде единствен за секој внесен податок во табелата. Како пример да ја земеме нашата база за толковниот речник во кој се креирани 4 табели и тоа:

- **Табела со содржина (sodrzina_recnik2)** во која се наоѓаат 66091 зборови од македонскиот јазик. Секој од тие зборови има свој индекс кој во табелата е означен како примарен клуч и важи во табелата за кирилица и латиница. Бидејќи речникот содржи зборови напишани само на кирилица, автоматски таквите зборови се внесуваат во табелата кај едниот атрибут (zbor), а потоа истите се конвертираат во латиница и се внесуваат во истата табела кај друг атрибут (latinica). Конвертирањето се однесува на нестандардно прикажување на буквите „ѓ=g“, „ж=z“, „ќ=k“, „ч=c“ и „ш=s“, додека останатите букви „а=a“, „б=b“, „ѕ=dz“, „њ=nj“, „џ=dj“ итн.” остануваат исти. Пример да ја земеме именката „куќа” која после конвертирање од кирилица во латиница се добива „kuka“. Од вака добиените зборови лесно може да се добијат двозначни поими.

- **Табела за двозначност (dvoznacnost_recnik2)** во која, преку споредување на зборови со латинско писмо од табелата со содржина, се добиени двозначни поими. На пример именката „kuka“ преку одредена скрипта за добивање двозначност на поими се добива {*kuka} –> {кука, куќа*}. Доколку ги потенцираме според атрибути, зборот „kuka“ се наоѓа кај (latinicni_poimi), додека „кука, куќа“ кај (dvoznacni_poimi).
- **Табела за број на реченици (br_recnici_recnik2)** во која се креирани три атрибути и тоа: атрибут за индекс на записите во табелата со примарен клуч, индекс на кирилски зборови според двозначниот поим, и вкупниот број на кирилски зборови според двозначниот поим. Редниот број за секој збор се добива од табелата со содржина (речник) во кои се наоѓаат зборови користени во македонскиот јазик. За да постои овој индекс во табелата за број на реченици, претходно е потребно да се прочитаат PDF документите, и во нив да се најдат зборовите според двозначните поими, да се изброи вкупниот број на таквите зборови, а потоа автоматски да се премине до редниот број. На пример зборот „куќа“ добиен од двозначниот поим „kuka“, со индекс 26695 се повторува 73 пати.
- **Табела со резултати (rezultati_recnik2)** се состои од неколку атрибути, потребни за решавање на повеќезначна транслитерација. За да се добијат вредностите за секој од овие атрибути, претходно се прави внесување на податоци во претходно опишаните табели, па потоа дел од тие податоци се користат и во оваа табела. Секој збор во оваа табела има свој индекс, преку кој се добива крајниот коефициент. При читање на реченици од надворешни извори се добиваат двозначни поими, а преку нив се добиваат и други зборови. Тоа значи дека се добиваат и внесуваат само зборови кои се наоѓаат на позиција три места пред и три места после двозначните поими. Сите овие добиени зборови може да се повторуваат во повеќе прочитани реченици, што значи бројачот секогаш се зголемува за единица. На крај со помош на скрипта за решавање на коефициент се добива коефициентот за секој збор во оваа табела.

Сите овие табели се прикажани во следниот дијаграм (ER дијаграм), од кој може лесно да се прочитаат соодветните ознаки на табелите, атрибутите, типот на секој атрибут и поврзаноста на секоја табела. Шематскиот приказ или

дијаграм е креиран во MySQLWorkbench програмата која овозможува исто така и приказ на состојбата на вредностите во сите креирани табели.

Слика 18. ER дијаграм на толковниот речник
Figure 18. ER diagram of expository dictionary

Останати атрибути во дадениот дијаграм се следните:

Табела (sodrzina_recnik2)

- reden_br (идентификационен број на секој кирилски збор од речникот);
- sostojba (има вредност 0 или 1 во зависност од тоа дали зборот е транслитериран во латиница);

Табела (rezultati_recnik2)

- id_r (инкрементална вредност)
- id_dv1 (идентификационен број на двозначниот поим);
- minus3, minus2, minus1, plus1, plus2, plus3 (број на повторувања на зборовите според позицијата);
- povtoruvanja (вкупен број на повторување на сите поими (лево, десно) од двозначните зборови).

8 ВЕБ СЕРВИСИ

Веб сервис е модуларна апликација која може да биде објавена, лоцирана и пристапна од било која точка на интернетот или локалната мрежа. Претставува метод на комуникација помеѓу две апликации или електронски уреди на интернет. Тоа е збир на стандарди каде апликациите мора да бидат усогласени со цел да се постигне интероперабилност низ Веб.

Цел на веб сервисите е да се овозможи поврзување на дистрибуирани софтверски компоненти без оглед на која платформа се спроведуваат, кој програмски јазик се користи, како и платформата на која тие се извршуваат.

8.1 REST (REpresentational State Transfer)

Веб сервисите обично се поврзани со SOAP (Simple Object Access Protocol). Но, постои и друг начин на реализирање на сервисите, а тоа е REST (REpresentational State Transfer) архитектурата. Тоа ќе биде начинот на комуникација помеѓу клиентите и серверот на нашиот веб сервис со помош на HTTP протокол. Ресурсите кои се праќаат и примаат како барање и одговор, може да бидат текстуални податоци, слика или бројки. REST користи пристап сличен како CRUD пристапот кај SQL јазикот во релациони бази на податоци и за секоја операција користи метод преку HTTP протоколот.

Табела 22. HTTP методи
Table 22. HTTP methods

HTTP	CRUD
POST	креирање, ажурирање, бришење
GET	земање (читање)
PUT	креирање, ажурирање
DELETE	бришење

За земање на ресурси се користи GET, за бришење DELETE, а за креирање и ажурирање POST и PUT методи. REST е ориентиран на ресурси и користи чисти URL адреси.

На пример од:

```
http://kokino.ugd.edu.mk/transliteracija/start.php?sodrzina=zbor
```

преминува во:

```
http://kokino.ugd.edu.mk/transliteracija/zbor
```

8.2 Споредување на веб сервиси со други технологии

- CORBA (Common Object Request Broker Architecture) е дизајнирана 1990 и овозможува механизам за креирање на клиент/сервер апликации во хетерогена средина (Marolt, 1996). CORBA на почетокот немала заштита при комуникација и трансфер на податоци. Како резултат на тоа, мрежните администратори би требало, портите за комуникација помеѓу сервисите да ги ограничат за да не дојде до напад од страна на хакерите. Во поновата верзија на CORBA тоа е поправено. CORBA е дизајнирана да работи со сите јазици и затоа користи IDL (Interface Definition Language) за конвертирање на објекти од еден јазик во друг.

- RMI (Remote Method Invocation) е Јава специфициран механизам за клиент/сервер повици. Разликата помеѓу овие две технологии е тоа што RMI се користи во Java-to-Java архитектури, но за IDL тоа не е потребно.

- DCOM (Distributed Common Object Model) е Microsoft механизам за remote повици. Тој може да се употреби во различни јазици (Visual C++, Visual Basic, C# итн.), но само на Microsoft платформа.

- HTTP трансакциска архитектура работи со помош на код кој работи на сервер како што е Apache. Клиентите комуницираат преку HTTP или HTTPS. Барањето оди до серверот, додека одговорот од серверот до клиентот се дава во вид на HTML или XML. Со користење на HTTPS (SSL енкрипција на HTTP) се обезбедува сигурност при преносот на податоците.

- ASP и PHP се технологии кои користат HTTP базиран сервис.

    - ASP (Active Server Pages) е креиран за Microsoft и базиран на Visual Basic јазик, може да работи на различни сервери и е Windows базирано решение.

    - PHP е креиран за Open Source и Linux/Unix и може да користи shell команди, што значи може да работи на повеќе сервери, но најчесто на Linux/Unix машини.

8.3 Веб сервис за транслитерација на цели реченици од латиница во кирилица

Нашиот веб сервис овозможува транслитерација на два начина. Едниот начин е директен приказ на транслитерирани содржини преку кориснички интерфејс, а другиот начин е преку модул за транслитерација, без интерфејс. Ова може да се применува во различни области, каде сетовите на податоци им се составени од македонски содржини напишани на латиница. Таквите содржини се прикажани во различни веб апликации, и доколку тие се во голема количина, не би можело да се менаџираат за краток временски период. За таа цел нашиот веб сервис овозможува побрз и поедноставен начин за управување со податоците, каде што за голем број на содржини ќе може да се користи начинот за транслитерација на барање од други апликации (модул), а за мал број на содржини преку кориснички интерфејс.

8.3.1 Кориснички интерфејс на веб сервисот

Сервисот се користи преку веб и може да опслужува задоволителен број на корисници. На секој корисник се прикажува интерфејсот кој е многу едноставен и лесен за употреба.

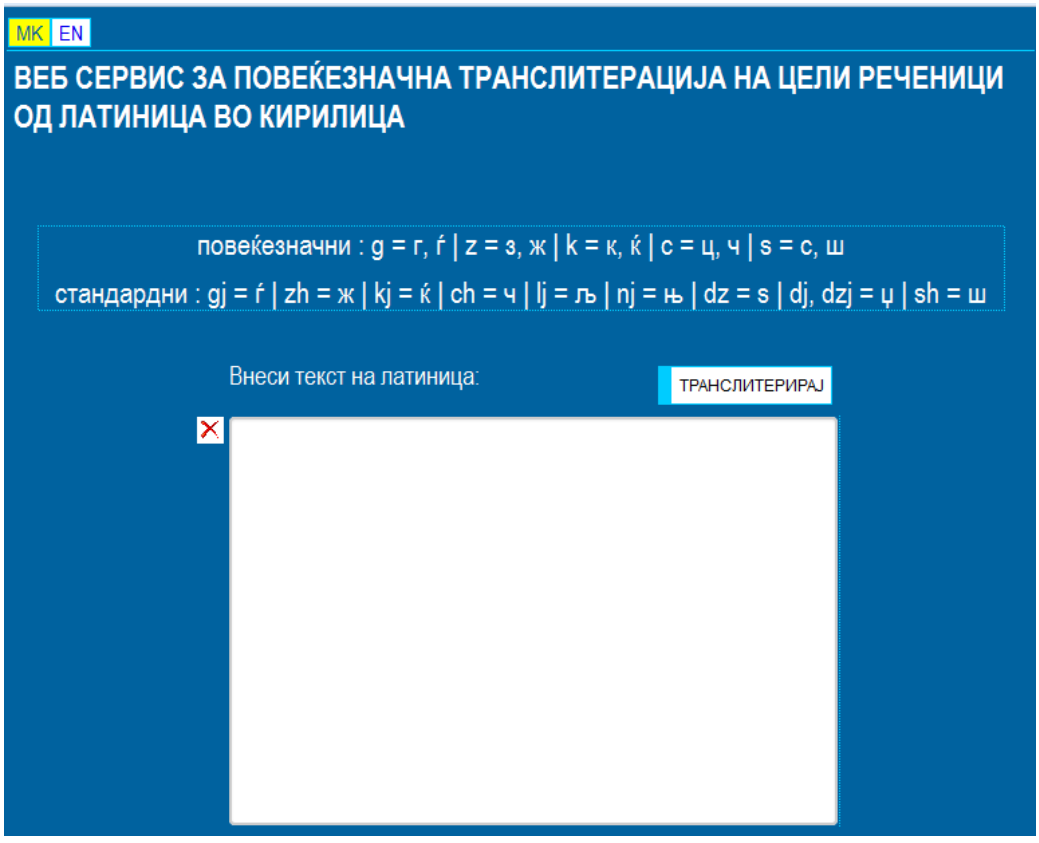
Овој веб сервис е прикачен на нашиот тест сервер на УГД и до него може да се пристапи преку линкот: http://kokino.ugd.edu.mk/transliteracija/.

Подетален опис

Кога се посетува веб сервисот, најпрво се отвора почетен прозорец на кој се прикажуваат неколку ентитети од HTML страницата. Тоа се однесува на:

- насловот на веб сервисот;
- копчињата за настани (Транслитерирај, Ново(X));
- место за внесување на реченици.

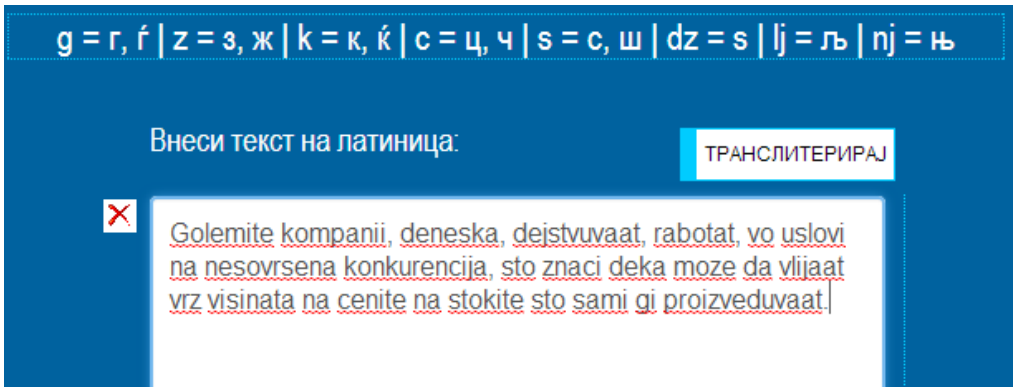
Настанот за транслитерација се вклучува кога копчето (Транслитерација) ќе се кликне барем еднаш. Со тоа се активираат сите скрипти и тоа: за пресметување на зборови (алгоритам, PHP, MySQL), за стил и за приказ на резултат (CSS3, jQuery, JavaScript). Тоа е прикажано на следната слика.



Слика 19. Приказ на формата за внесување на реченици за транслитерација
Figure 19. Overview of the form for entering sentences for transliteration

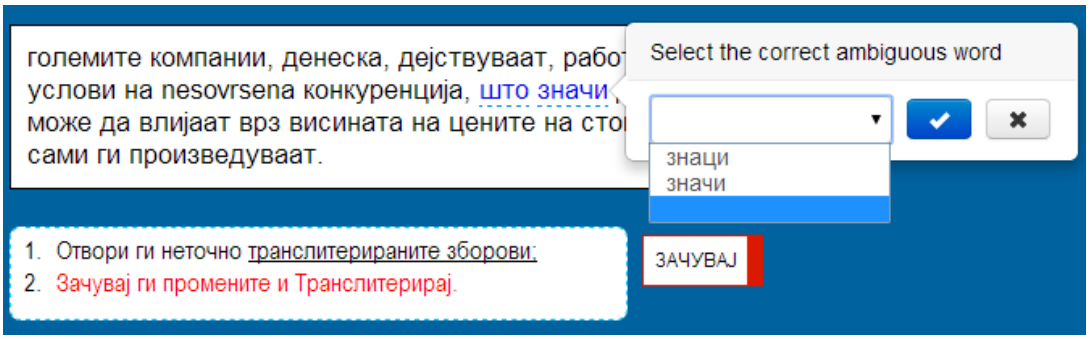
Наредeн чекор е транслитерација на внесената содржина по барање од корисникот. Овде може да се транслитерираат содржини до 1500 зборови заедно со знаци, со што се ограничува пристапот за трошење на перформансите

на серверот. После кликање на копчето (Транслитерирај) се вчитува ограничената содржина и започнува процесот за транслитерација со помош на изработениот алгоритам. Во зависност од големината на содржината ќе зависи и процесирањето на алгоритамот. На следната слика, како пример е прикажана една реченица, која е преземена од интернет и за неа ќе направиме неколку споредби.



Слика 20. Внесена реченица како барање за транслитерација
Figure 20. Entered sentence as a request for transliteration

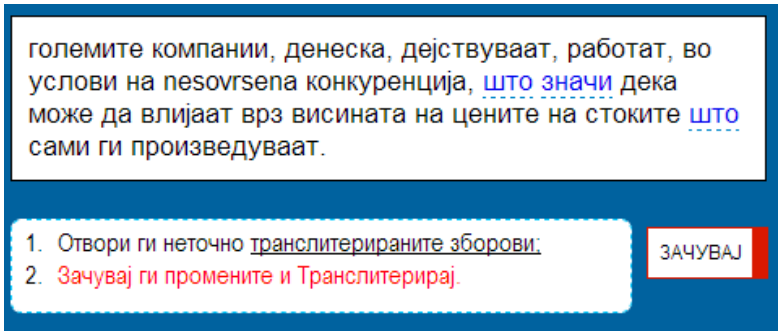
Процесот на транслитерација трае дел од секунда и како резултат се добива реченица, која во зависност од двозначноста се појаваат зборови подвлечени со испрекината линија и означени со сина /плава боја (што, значи, што). Таквите зборови се двозначни и на секој од нив може да се направи корекција, доколку во реченицата немаат точно значење. Корекцијата е овозможена на секој двозначен збор со неколку кликања (Слика 21).



Слика 21. Корекција на реченица со помош на предложени двозначни зборови
Figure 21. Correction of sentence using proposed words with two meanings

Начинот на кој може да се прави промени е преку Popup прозорецот на кој е дадена листа со двозначни зборови (знаци, значи). Во примерот за корекција е

земен вториот подвлечен двозначен збор „значи“, и доколку треба да се коригира се селектира точниот збор и се потврдува со копчето за валидност.



Слика 22. Приказ на крајниот резултат на транслитерирана реченица
Figure 22. Overview of the end result in transliterated sentence

На Слика 22 е прикажан резултатот од транслитерацијата, односно:

- транслитерираната реченица која во даден блок е составена од непознати зборови (nesovrsena), двозначни зборови (што, значи, што), и обични зборови. Во крајниот резултат, непознатите зборови остануваат исти, бидејќи не постојат во база на податоци;
- бројот на двозначни зборови кои се подвлечени со испрекината линија;
- бројот на реченици кои се пресметани според точките кои се наоѓаат на крајот од секоја реченица.

Сите овие транслитерирани содржини може да се зачуваат во база на податоци со само едно кликање на копчето (Зачувај). После овој чекор се активира настанат за невидливост на тоа копче, и автоматски се овозможува самоучење[17] на главниот алгоритамот за транслитерација.

Внесувањето на нетранслитерирани реченици во текст полето може да се меша на кирилица и латиница. Таа опција е овозможена, бидејќи не претставува проблем за добивање на точен резултат. Тоа се должи на алгоритамот за транслитерација кој на почетокот прави конвертирање во латиница на сите знаци од зборовите (на пример: работат->rabotat->работат), а потоа и во

[17] Самоучење претставува акција која позиционира поими лево и десно од двозначните зборови и автоматски ги зачувува во база. Тоа се стартува кога корисникот сака да ја зачува транслитерираната содржина со копчето „Зачувај“. Оваа акција е применета во првиот дел од истражувањето кога автоматски се читаа содржини од PDF документи, но без кориснички интерфејс.

кирилица (на пример: rabotat->rabotat->работат). Во овој кориснички интерфејс е овозможена и јазична поддршка на македонски (МК) и англиски (EN).

8.3.2 Модул за транслитерација на веб сервисот

Овој модул е направен со цел транслитерирањето на содржини да се овозможува преку барање од други апликации. Тој е направен да биде компатибилен во сите PHP скрипти кој се наоѓаат во апликациите на различни сервери. Модулот за транслитерација (Слика 23) е тестиран на една експериментална веб апликација на која се објавуваат прашања и коментари. Апликацијата е поврзана со база на податоци и преку неа е овозможено објавување на содржини во реално време. Интегрирањето на дополнителниот програмскиот код (Слика 24) во самата апликација, ќе овозможи пристап до тој модул, со што автоматски ќе се ажурира базата на сервисот (самоучење) и базата на апликацијата. Оваа примена на модулот за транслитерација е во моментот кога корисниците внесуваат содржини во веб апликацијата. Следува приказ на модулот за транслитерација на големи содржини.

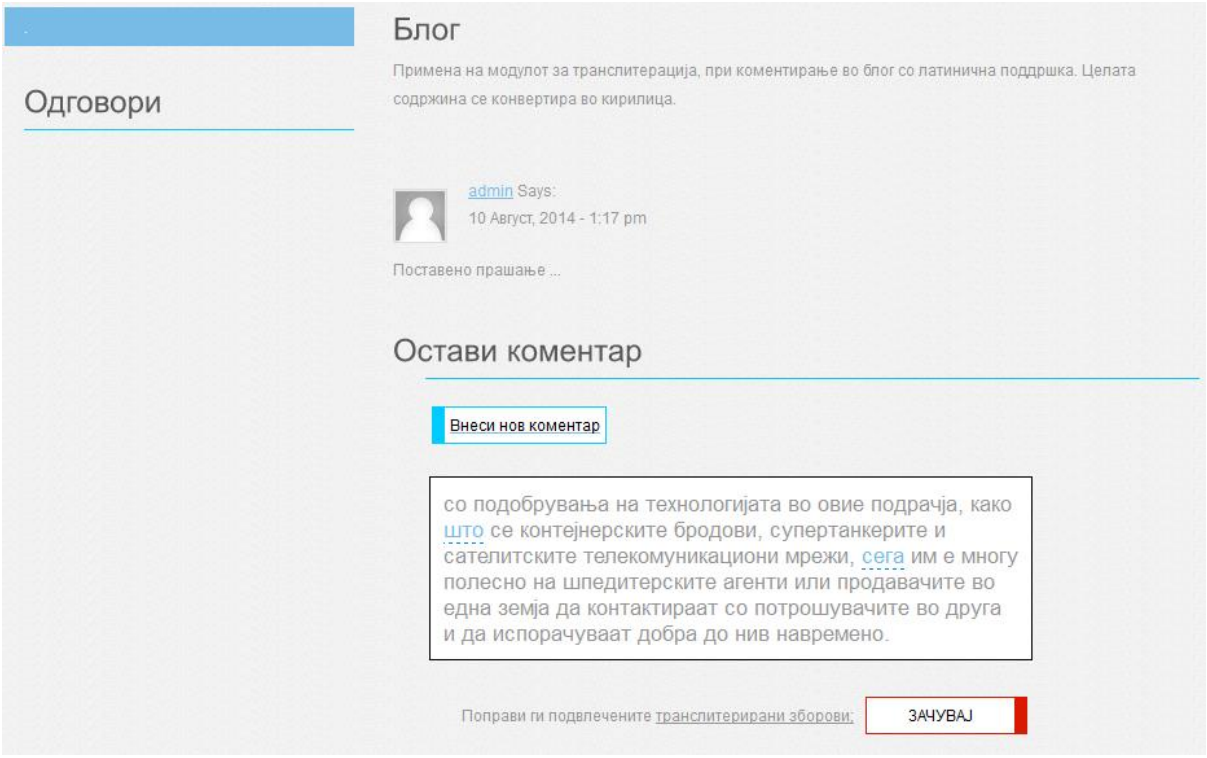
```php
<?php
 error_reporting(0);
 header('Content-Type: text/html; charset=utf-8');
 mb_internal_encoding("UTF-8");

 $sodrzina = $_POST['sodrzina'];
 mysql_real_escape_string($sodrzina);

 if(isset($sodrzina) and $sodrzina!=''){

    echo web_service_transliteration($sodrzina);

 }

 function web_service_transliteration($sodrzina){

    if(isset($sodrzina) and $sodrzina!=''){

       $sodrzina = urlencode($sodrzina);
       $sodrzina = str_replace(array("\n", "\r"), ' ', $sodrzina);

       $link = file_get_contents("http://kokino.ugd.edu.mk/transliteracija/s_get/$sodrzina");

       return $link;
    }
 }

?>
```

Слика 23. Програмски код на модулот за транслитерација на веб сервисот
Figure 23. Program code in the module for web service transliteration

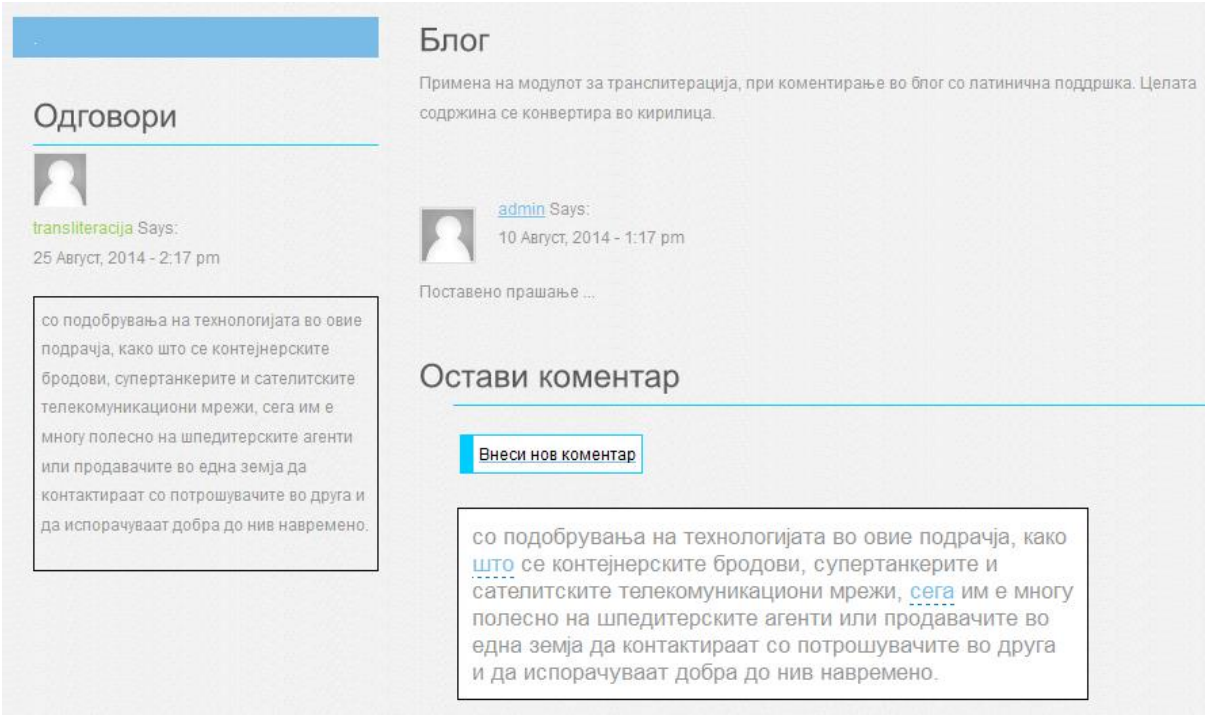
```
jQuery(function(){

        $("#novo").hide();
        var pole_so_sodrzina = "textarea#sodrzina";
        var kopce_za_transliteracija = "input#transliteriraj";

        var pristap_do_veb_servis = "http://kokino.ugd.edu.mk/transliteracija/s_post";
        var pristap_do_tvoj_link = "http://kokino.ugd.edu.mk/veb-sajt/index_trans.php";

         $(kopce_za_transliteracija).on("click", function(){

            $(pole_so_sodrzina).hide();
            $(kopce_za_transliteracija).hide();
            $("#novo").show();

            var sodrzina_za_transliteriranje =  $(pole_so_sodrzina).val();

            $.post(pristap_do_veb_servis, {sodrzina : sodrzina_za_transliteriranje, link : "0"},
function(data){

                        $("#kraen_rezultat").html(data);

                            $("#dodadi").on("click", function(){

                                    var zacuvaj_sodrzina = $("#div2").text();

                                   $.post(pristap_do_tvoj_link, {transliterirana_sodrzina :
            zacuvaj_sodrzina}, function(data){

                                    });

                                });
                });

            });

       });
```

Слика 24. Програмски код за пристап до модулот за транслитерација
Figure 24. Program code for approach to the module for transliteration

На (Слика25 и Слика26) е прикажано како корисниците внесуваат коментари, и како содржините се зачувуваат во база на податоци.



Слика 25. Транслитерација на барање од други веб апликации пред ажурирање на експериментална база на податоци
Figure 25. Transliteration upon request from other web applications before updating the experimental database



Слика 26. Транслитерација на барање од други веб апликации после ажурирање на експериментална база на податоци
Figure 26. Transliteration upon request from other web applications after updating the experimental database

Но постои и друг начин за примена на модулот, а тоа е кога во Веб апликациите се наоѓаат веќе внесени македонски содржини напишани на латиница, што значи дека ќе се автоматизира повикувањето и ажурирањето на содржини во апликацијата. Поради тоа е потребно дополнително ажурирање на модулот за транслитерација (Слика 23) за да може да функционира правилно.

9 ЗАКЛУЧОК

Овој магистерски труд е резултат на повеќемесечни истражувања и анализи на феноменот транслитерација која во принцип може да конвертира букви од едно писмо во друго. Заклучивме дека дел од македонските зборови кои се напишани на латиница кога ќе се транслитерираат во кирилица добиваат повеќе од едно значење (dokazi: докажи, докази). Оваа појава доведува до нејасност на зборовите во цели реченици кога се читаат од страна на корисниците.

Затоа е предложен нов алгоритам кој ќе може да го реши проблемот на повеќезначна транслитерација. Алгоритамот функционира со функции кои можат да го пронајдат двозначниот збор во дадена реченица и да направат ограничување на позициите на три зборови кои се наоѓаат пред и после него во дадена реченица. За реализирање на транслитерацијата користени се трите типа на речници и тоа: ОР; ТР; и КР.

Со ново креирани скрипти најпрво е направено пребарување на двозначни зборови во секој речник внесен во база на податоци. Потоа се направени вчитувања на содржини од PDF документи и за истите се добиени резултати кои се зачувани во соодветни табели во база на податоци.

Резултатите се однесуваат на:

- вкупно двозначни зборови;
- вкупно реченици прочитани во сите PDF документи;
- вкупно зборови прочитани во сите PDF документи;
- вкупно реченици со повеќезначни зборови;
- вкупно поими (лево, десно) од двозначните зборови;
- вкупен број на повторување на сите поими (лево, десно) од двозначните зборови.

Од добиените резултати може да се заклучи дека кај ОР 2% од зборовите се двозначни, додека кај ТР само 1,5%. Во текот на истражувањето се дефинирани три формули за транслитерација со примена на условна веројатност и Бајесови

формули. Постапката за решавање на повеќезначна транслитерација продолжува со тестирање на формулите и анализа на добиените резултати.

Процесот на тестирање се овозможи преку 300 реченици од вкупно 36.746 двозначни реченици напишани на латиница и кирилица, и беа поделени на:

- реченици со најголемо повторување на двозначни зборови,
- реченици со средно повторување и
- реченици со најмало повторување.

Анализата е направена врз основа на позитивните и негативните резултати кои се добиени при тестирање на реченициите. Дали тестираните реченици се добиени позитивно на кирилица или латиница ќе зависи од тоа дали ги имаме сите двозначни поими внесено во база. Во овој магистерски труд сите резултати се претставени преку дијаграми и табели.

9.1 Дискусија

Сервисот за транслитерација е фокусиран на анализа на содржини кои се праќаат преку корисничко поле, а потоа и на конвертирање букви со претходно дефинирани методи.

За да може сето ова да функционира потребен е алгоритам за повеќезначна транслитерација кој ќе работи според начинот за определување на повеќезначни реченици.

Тој начин се однесува на идентификација на двозначен поим во дадената реченица, и позиционирање на зборовите кои се наоѓаат 3 места пред и 3 места после поимот.

Ова е услов како да се добие приближно добар резултат. Спротивно од тоа постојат и грешки при транслитерација кои создаваат негативен резултат.

Грешките се појавуваат најчесто кога реченициите содржат: специјални знаци, децимални броеви и римски броеви.

Напорите за во иднина треба да се насочат кон:

- Позиционирање на зборовите од 3 на 5 места пред и после двозначниот поим;

- Зголемување на двозначни зборови во база на податоци, преку дополнително прочитани содржини;

- Анализа и тестирање на поголем број реченици (>300);

- Можност за внесување или транслитерирање на поголем број на зборови кои се внесуваат во корисничкиот дел за транслитерација;

- Подобрување на формулите за транслитерација.

- Автоматско предвидување на зборовите кои по грешка се внесуваат од корисниците во формите за пишување на текст (точно: pregled; грешка: prgled, pergled).

ПРИЛОЗИ

Следуваат неколку прилози во кои се прикажани реченици кои се тестирани и анализирани преку различни скрипти. Трите тестови кои се направени претходно, овде се прикажани табеларно со различни ознаки. Во секој прилог, покрај зададените реченици, прикажан е и вкупниот број на позитивни и негативни резултати.

	Оригинална реченица		Транслитерација на оргинална реченица (напишана на кирилица) Вкупно реч.: 300 Позитивни: 291 Негативни: 9
58	ѓубрењето со јаглен диоксид се применува во оранжерии и пластеници кога денот е сончев и топол (околу пладне), со зголемување на концентрацијата на со2 на 0, 2-0, 5%, а објектот се држи затворен 24 часа за да се продолжи поволното дејство на со2.	58	ѓубрењето со јаглен диоксид се применува во оранжерии и пластеници кога денот е сончев и топол (околу пладне), со зголемување на концентрацијата на со2 на 0, 2-0, 5%, а објектот се држи затворен 24 часа за да се продолжи поволното дејство на со2.
59	подгревање на семето се врши во термостат на температура која постепено се зголемува, од 50-60ос, 3-5 часа, со почесто мешање на истото.	59	подгревање на семето се врши во термостат на температура која постепено се зголемува, од 50-60ос, 3-5 часа, со почесто мешање на истото.
60	дополнителното осветлување се изведува веднаш со самото никнење, со тоа што вкупната должина на денот (природно и вештачко светло) за домат и пиперка да биде 16-18 часа, а за краставица 14-16 часа.	60	дополнителното осветлување се изведува веднаш со самото никнење, со тоа што вкупната должина на денот (природно и вештачко светло) за домат и пиперка да биде 16-18 часа, а за краставица 14-16 (часа).
61	целата територија на државата на нутс ниво 1 е една единица.	61	целата територија на државата на нутс ниво 1 е една единица.
62	во текот на одреден временски период тие имаат стапки на пораст на сопствениот производ знатно поголем од просечната стапка како на индустријата во целина така и на целото стопанство, односно целата национална економија.	62	во текот на одреден временски период тие имаат стапки на пораст на сопствениот производ знатно поголем од просечната стапка како на индустријата во целина така и на целото стопанство, односно целата национална економија.
63	степенот на органската поврзаност во прв ред зависи од разновидноста и квалитетот на сообраќајната инфраструктура, при што не помало значење имаат и водотеците како важен инфраструктурен и технолошки фактор на целата содржина на развојните осовини.	63	степенот на органската поврзаност во прв ред зависи од разновидноста и квалитетот на сообраќајната инфраструктура, при што не помало значење имаат и водотеците како важен инфраструктурен и технолошки фактор на целата содржина на развојните осовини.

Слика 27. Прилог 1 - Транслитерација на реченици напишани на кирилица со примена на Формула 1

Figure 27. Annex 1 – Transliteration of sentences written in Cyrillic alphabet using Formula 1

	Оригинална реченица		Транслитерација на оргинална реченица (напишана на кирилица) Вкупно реч.: 300 Позитивни: 292 Негативни: 8
200	влијание врз гласот: поради растење на грклanot и здебелување на лигавицата на гласните жици, гласот станува подлабок, типично машки.	200	влијание врз гласот: поради растење на грклanot и здебелување на лигавицата на гласните жици, гласот станува подлабок, типично машки.
201	молекулите на тропомиозинот меѓусебно се поврзани во долга нишка, а две нишки градат дупли хеликс.	201	молекулите на тропомиозинот меѓусебно се поврзани во долга нишка, а две ниски градат дупли хеликс.
202	создавање и градба на нишките-секоја нишка почнува со здружување на крајот на едната со крајот на другата опашка на миозинскиот молекул.	202	создавање и градба на нишките-секоја нишка почнува со здружување на крајот на едната со крајот на другата опашка на миозинскиот молекул.
203	во влакната на скелетните мускули напречните мостови се насочени кон 6 тенки нишки кои ја опкружуваат секоја дебела нишка.	203	во влакната на скелетните мускули напречните мостови се насочени кон 6 тенки ниски кои ја опкружуваат секоја дебела нишка.
204	кога мускулот ќе се надразни се зголемува концентрацијата са2+ јони во миоплазмата што причинува промена во градбата на миофибрилите, допуштајќи притоа напречниот мост да се врзе со тенката нишка.	204	кога мускулот ќе се надразни се зголемува концентрацијата са2+ јони во миоплазмата што причинува промена во градбата на миофибрилите, допуштајќи притоа напречниот мост да се врзе со тенката нишка.
205	ваквата промена на положбата на напречното мовче создава сила која го придвижува главчето од тенката кон дебелата нишка.	205	ваквата промена на положбата на напречното мовче создава сила која го придвижува главчето од тенката кон дебелата нишка.
206	имено, во состав на хормозомите се констатира присуство на нуклеопротеинско влакно или нишка составена од днк, како и протеини хистони и нехистони, во однос 1: 1.	206	имено, во состав на хормозомите се констатира присуство на нуклеопротеинско влакно или нишка составена од днк, како и протеини хистони и нехистони, во однос 1: 1.
207	на пример, amazon ќе ве извести со маил кога новите книги од вашата омилена тема или од ваш омилен автор ќе бидат публикувани.	207	на пример, amazon ќе ве извести со маил кога новите книги од вашата омилена тема или од ваш омилен автор ќе бидат публикувани.

Слика 28. Прилог 2 - Транслитерација на реченици напишани на кирилица со примена на Формула 2

Figure 28. Annex 2 – Transliteration of sentences written in Cyrillic alphabet using Formula 2

	Оригинална реченица		Транслитерација на оргинална реченица (напишана на кирилица) Вкупно реч.: 300 Позитивни: 289 Негативни: 11
228	затоа, како ниско поставена ушна школка се смета онаа кај која најниската точка се наоѓа пониско од усниот агол или ако горниот раб на ушната школка се наоѓа под замислената хоризонтална линија која поминува низ надворешниот агол на окото.	228	затоа, како ниско поставена ушна школка се смета онаа кај која најниската точка се наоѓа пониско од усниот агол или ако горниот раб на ушната школка се наоѓа под замислената хоризонтална линија која поминува низ надворешниот агол на окото.
229	ушниот приврзок (ресичката-lobulus auriculae) сраснат со вратот е минор-аномалија кога ресичката на ушната школка (lobulus auriculae) не виси на долниот дел од аурикулата, туку е врзана за вратот, а понекогаш и насочена кон горе и кон назад.	229	(усниот) приврзок (ресичката-lobulus auriculae) сраснат со вратот е минор-аномалија кога ресичката на ушната школка (lobulus auriculae) не виси на долниот дел од аурикулата, туку е врзана за вратот, а понекогаш и насочена кон горе и кон назад.
230	ќе можат да ги пријават сите статусни промени за даночни цели, како на пример: промена на податоци за контакт и адреса, пријава/одјава на фискален апарат, алокација на фискална каса во деловните единици и др.	230	ќе можат да ги пријават сите статусни промени за даночни цели, како на пример: промена на податоци за контакт и адреса, пријава/одјава на фискален апарат, алокација на фискална каса во деловните единици и др.
231	бидејќи во кимберлитските бречи не се забележуваат појавите на контактните промени, може да се заклучи дека магмата во голема мера била оладена, во вид на каса.	231	бидејќи во кимберлитските бречи не се забележуваат појавите на контактните промени, може да се заклучи дека магмата во голема мера била оладена, во вид на каса.
232	на големината на мултипликаторот влијаат и одлуките на депозитните банки за тоа колку средства ќе држат на својата жиро сметка или во каса.	232	на големината на мултипликаторот влијаат и одлуките на депозитните банки за тоа колку средства ќе држат на својата жиро сметка или во каша.
233	зголемувањето на износот во каса го намалува монетарниот мултиплокатор.	233	зголемувањето на износот во каша го намалува монетарниот мултиплокатор.
	освен ова и увозниците на земјата во која курсевите на странските валути растат ќе настојуваат да ги		освен ова и увозниците на земјата во која курсевите на странските валути растат ќе настојуваат да ги

Слика 29. Прилог 3 - Транслитерација на реченици напишани на кирилица со примена на Формула 3

Figure 29. Annex 3 – Transliteration of sentences written in Cyrillic alphabet using Formula 3

Следуваат уште неколку прилози (Слика 30, 31, 32, 33) во кои се прикажани нумерички податоци изразени во цели броеви и во проценти. Тие се добиени врз основа на прочитаните содржини од различни PDF документи, со помош на автоматска скрипта. За време на работењето на скриптата можеше да се види во живо секој нов податок кој е внесен во базата. Резултати подолу може да се прикажуваат во било кој пребарувач визуелно со помош на PHP и MySQL технологиите. Наредните прилози се однесуваат на добиените резултати кои се зачувани во четири различни речници: отворен, толковен, комбиниран и комбиниран плус.

Отворен речник | Дигитален речник | Комбиниран речник | Комбиниран речник - Plus | Решение 2 - наша формула | Решение 1 - формула +0.5 | Решение 2 - формула +0.5 | Решение 3 - формула +0.5

Податоци за Речник 1 (Отворен речник)

*Вкупно двозначни зборови *(или 647 двозначни поими - sega = [сега, шега]) :*

432 од 1294

33.38 %

Процент на двозначни реченици од Вкупно реченици прочитани во сите пдф документи :

36746 од 329958

11.14 %

Процент на двозначни зборови од Вкупно зборови прочитани во сите пдф документи :

36746 од 4187998

0.88 %

Вкупно реченици со повеќезначни зборови :

36746

Вкупно поими (лево, десно) од двозначните зборови :

60365

Вкупен број на повторување на сите поими (лево, десно) од двозначните зборови :

229315

10 најзастапени двозначни зборови :

Приказ на сите пронајдени двозначни зборови во текстовите:

id_dvozn_poimi	двозначен збор	бројач
223590	сеф (шеф)	6
260825	штеп (степ)	2
178073	постава (поштава)	20
124113	наведувајќи (наведувајќи)	1
138485	нишкиот (ниските)	1
223614	сефови (шефови)	1
220732	сведоците (сведочите)	5
188522	преписи (препиши)	5
97994	козара (кожара)	2
182250	пошти (пости)	1
180158	потпиши (потписи)	1
245555	ќош (кош)	1
254523	целава (челава)	1
105021	лажам (лазам)	1
113364	мешам (месам)	2
245420	ќенан (кенан)	5
57443	загрижуваат (загризуваат)	3
105051	лажењето (лазењето)	1
57418	загрижи (загризи)	1
260699	штавена (ставена)	1
33500	грижел (гризел)	3
190337	пресечи (пресеци)	1
115271	множински (мнозински)	3
97919	кожарска (козарска)	1
10599	белезите (бележите)	1
117363	мусев (мушев)	1
108014	лозана (ложана)	10
111902	машине (масине)	1
216233	резе (реже)	3
4373	ангелот (ангелот)	1
219060	руси (руши)	2

Слика 30. Прилог 4 – Извештај за податоците од отвореният речник кои се добиени преку автоматска скрипта за читање на текст содржини

Figure 30. Annex 4 – Report on data from the open dictionary obtained by automatic script for reading text contents

Слика 31. Прилог 5 – Извештај за податоците од толковниот речник кои се добиени преку автоматска скрипта за читање на текст содржини

Figure 31. Annex 5 – Report on data from the expository dictionary obtained by automatic script for reading text contents

Слика 32. Прилог 6 – Извештај за податоците од комбинираниот речник кои се добиени преку автоматска скрипта за читање на текст содржини

Figure 32. Annex 6 – Report on data from the combined dictionary obtained by automatic script for reading text contents

Слика 33. Прилог 7 – Извештај за податоците од комбинираниот плус речник кои се добиени преку автоматска скрипта за читање на текст содржини
Figure 33. Annex 7 – Report on data from the combined plus dictionary obtained by automatic script for reading text contents

Следните прилози (Слика 34, 35) се однесуваат на програмски код од jQuery скриптите кои се битни за функционалноста на веб сервисот за транслитерација.

```
jQuery(function($){

  kopce1();
  kopce1_en();

});

function kopce1(){

   $("input#prikazi").on('click', function(){

      var ime = $("textarea#ime").val();

      $.post('start', {sodrzina : ime}, function(data){

         $('#od-baza').html(data);

      });

   });

}

function kopce1_en(){

   $("input#prikazi_en").on('click', function(){

      var ime = $("textarea#ime").val();

      $.post('start', {sodrzina_en : ime}, function(data){

         $('#od-baza').html(data);

      });

   });

}
```

Слика 34. Прилог 8 – jQuery функции за стартување на алгоритамот за транслитерација кај кориснички интерфејс на веб сервисот
Figure 34. Annex 8 – jQuery functions for starting transliteration algorithm by user interface of web service

```
jQuery(function(){

    function kopce2(){
      $("input#dodadi").on('click', function(){

        $('#dialog').remove();
        $('input#dodadi').remove();

        var kopce_dodadi = $("#div2").text();

        $.post('insert.php', {sodrzina : kopce_dodadi}, function(data){
          alert(data);
        });

      });


    }

    kopce2();

    $('#dialog').hide();
    $('#dialog').fadeIn();
    $('input#dodadi').hide();
    $('input#dodadi').fadeIn();

    $("#div2").hide();
    $('#div2').fadeIn('slow');

});
```

Слика 35. Прилог 9 – jQuery функција за додавање на нова содржина во база на податоци, преку кориснички интерфејс на веб сервисот

Figure 35. Annex 9 – jQuery function for adding new content in database, by user interface of the web service

КОРИСТЕНА ЛИТЕРАТУРА

Biemann, C. (2012). *Structure Discovery in Natural Language.* Berlin: Springer.

Cao, N. X., Pham, N. M., & Vu, Q. H. (2010). Comparative Analysis of Transliteration Techniques based on Statistical Machine Translation and Joint-Sequence Model. *Proceedings of the 2010 Symposium on Information and Communication Technology* (pp. 59-63). ACM.

Chiang, D. (2012). *Grammars for Language and Genes.* Berlin: Springer.

Chinnakotla, M. K., Damani, O. P., & Satoskar, A. (2010). Transliteration for Resource-Scarce. *ACM Transactions on Asian Language Information Processing (TALIP), 9*(4), 30.

Clark, A., Fox, C., & Lappin, S. (2010). *The Handbook of Computational Linguistics and Natural Language Processing.* Blackwell Publishing Ltd.

Daniela, O., Burger, S., & Weilhammer, K. (2000). What are transcription errors and Why are they made? *LREC.*

Dyer, C. (2010). A formal model of ambiguity and its applications in machine translation.

Eiji, A., & ABEKAWWA, T. (2009). Fast decoding and Easy Implementation:. *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration*, (pp. 65-68).

Feldman, L., Lukatela, G., & Turvey, M. T. (1985). Effects of phonological ambiguity on beginning readers of Serbo-Croatian. *Journal of experimental child psychology* , 492-510.

Filipović Đurđević, D., Milin, P., & Beth Feldman, L. (2013). Bi-alphabetism: A window on phonological processing. *Psihologija*, 421–438.

Fukunishi, T., Finch, A., Yamamoto, S., & Sumita, E. (2013). A Bayesian Alignment Approach to Transliteration Mining. *ACM Transactions on Asian Language Information Processing (TALIP), 12*(3), 22.

Hsu, C. C., & Chen, C. H. (2010). Mining Synonymous Transliterations from the World Wide Web. *ACM Transactions on Asian Language Information Processing (TALIP), 9*(1), 27.

Hsu, C. C., Chen, C. H., Shih, T. T., & Chen, C. K. (2010). Measuring Similarity between Transliterations. *ACM Transactions on Asian Language Information Processing (TALIP), 6*(1), 20.

Hudeček, L. (1987). Transliteracija i Transkripcija. *Rasprave Instituta za hrvatski jezik i jezikoslovlje, 13*(1), 19-30.

Indurkhya, N., & Damerau, F. J. (2010). *Natural Language Processing, Second Edition.* Chapman & Hall/CRC .

Jong-Hoon, O., Key-Sun, C., & Hitoshi, I. (2006). A Comparison of Different Machine Transliteration Models. *Journal of Artificial Intelligence Research*, 119–151.

Jong-Hoon, O., Uchimoto, K., & Torisawa, K. (2009). Machine transliteration using target-language grapheme and phoneme: multi-engine transliteration approach. *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration*, (pp. 36-39).

Kang, I.-H., & Kim, G. (200). English-to-Korean Transliteration using Multiple Unbounded. *Proceedings of the 18th conference on Computational linguistics*, (pp. 418-424).

Layton, T. L., Crais, E. R., & Watson, L. R. (2000). *Early Language Impairment in Children: Nature.* Delmar Thomson Learning.

Lopez, A. (2008). Statistical machine translation. *ACM Computing Surveys, 40*(3), 1-49.

Lukatela, G., Turvey, M. T., Feldman, L. B., Carello, C., & Katz, L. (1989). Context Effects in Bi-alphabetical Word Perception. *Journal of Memory and Language, 28*(2), 214–236.

Luke, W., & Thomson, L. (2003). *PHP and MySQL Web development.* Sams Publishing.

Manning, C. D., & Suetze, H. (1999). *Foundations of Statistical Natural Language Processing.* London: The MIT Press.

Marolt, M. (1996). CORBA - A Framework for Development od Distributed Applications. *Electrotechnical Review, 63*, 286-291.

Nasreen, A., & Larkey, L. S. (2003). Statistical transliteration for English-Arabic cross language information retrieval. *Proceedings of the twelfth international conference on Information and knowledge management* (pp. 139-146). ACM.

Olive, J., Christianson, C., & McCary, J. (2011). *Handbook of Natural Language Processing and Machine Translation.* Berlin: Sringer.

Pervouchine, V., Li, H., & Lin, B. (2009). Transliteration Alignment. *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, (pp. 136-144).

Prochasson, E., Kageura, K., Morin, E., & Aizawa, A. (2008). Looking for Transliterations in a Trilingual English, French and Japanese. *LREC Workshop on Comparable Corpora (LREC'08).* Language Resources and Evaluation Conference.

Rico, S. (2009). *Syntactically Enriched Statistical Machine Translation.* Zurich.

Sajjad, H., Fraser, A., & Schmid, H. (2011). An Algorithm for Unsupervised Transliteration Mining with an Application. *HLT '11 Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, *1*, pp. 430-439.

Saldanha, G. (n.d.). Principles of corpus linguistics and their application to translation studies research. Birmingham .

Šimičević, G., & Boljanović, A. (2009). Transcription and Transliteration in a Computer. *INFuture2009: Digital Resources and Knowledge Sharing*, (pp. 365-374). Zagreb.

Spasov, S., & Zdravev, Z. (2013). Web service for ambiguous transliteration of full sentences from latin to cyrillic alphabet. *JS, 1*(1).

Zhang, M., Zhang, L., Li, H., Kumaran, A., & Liu, M. (2012). Report of NEWS 2012 Machine Transliteration Shared Task. *NEWS '12 Proceedings of the 4th Named Entity Workshop*, (pp. 10-20).

Дигитален речник на македонскиот јазик. (n.d.). Retrieved from www.makedonski.info.

Стојанче Спасов

Веб сервис за повеќезначна транслитерација на цели реченици
од латиница во кирилица

Универзитет „Гоце Делчев“ - Штип