УНИВЕРЗИТЕТ „ГОЦЕ ДЕЛЧЕВ" - ШТИП
ФАКУЛТЕТ ЗА ИНФОРМАТИКА

ГОДИШЕН ЗБОРНИК
2012
YEARBOOK
2012

ГОДИНА 1   VOLUME I

GOCE DELCEV UNIVERSITY - STIP
FACULTY OF COMPUTER SCIENCE

УНИВЕРЗИТЕТ „ГОЦЕ ДЕЛЧЕВ“ – ШТИП
ФАКУЛТЕТ ЗА ИНФОРМАТИКА

ГОДИШЕН ЗБОРНИК
2012
YEARBOOK
2012

СОДРЖИНА
CONTENT

WEB SERVICE FOR AMBIGUOUS TRANSLITERATION OF FULL SENTENCES FROM LATIN TO CYRILLIC ALPHABET

Stojance Spasov[1]   **Zoran Zdravev**[1]

stojance.spasov@ugd.edu.mk   zoran.zdravev@ugd.edu.mk

[1]Faculty of Computer Science, University "Goce Delcev" - Stip

Abstract

Introduction to transliteration as a process discovers the ability to differ and convert symbols from one language with different meaning to other languages. The intelligent algorithm for detecting Latin and Cyrillic alphabet has a need of minimal steps for transliteration, in the cases when the words from one language can have more than one meaning. This paper pays attention to the forms of transliteration of full sentences, which shall be used for Macedonian texts written in Latin alphabet on social networks, web-sites, old magazines etc.

Keywords: algorithm, Macedonian texts, Latin and Cyrillic alphabet, more than one meaning.

Introduction

The transliteration is a process used for mapping of words from one system of writing to another. This procedure is widely applied in Macedonian alphabet that uses Cyrillic letters, due to the circumstances when the writing of letters in Latin alphabet is required. With the assistance of modern techniques and technologies, the process of easier and faster transliteration under standard circumstances is enabled. But under non-standard circumstances, common for Macedonian area, together with the grammar rules, there is often a misunderstanding, especially regarding the meaning and ambiguity of words. For example from "kuka" we derive "кука" or "куќа", from "sok" we derive "сок' or "шок" and so on. It mostly happens in international communication, where different technologies, without Cyrillic input language, are used for sending messages and mails from mobile phones, writing on social networks etc. Therefore, regarding transliteration with different meanings, researches of the purpose and meaning of the words in the sentence are required as well as algorithm that will solve this problem. This could greatly contribute to Macedonian alphabet, especially for terms written in Latin alphabet. The purpose of this research is creating a web based service i.e. defining an intelligent algorithm that will enable transliteration of full sentences from Latin to Cyrillic alphabet as well as transliteration upon request of other applications.

Transliteration

The words from any language should sometimes be written in other alphabet. Mostly this happens under Macedonian circumstances as well, when the words are written in Latin alphabet. This is due to the fact that communication devices do not have Cyrillic input language. When sending an email from mobile phone written in Latin alphabet for example. The transliteration may be reversible and convert terms from one alphabet to another. The transliteration is not always a simple process for realization, because there is not a difference drawn between writing Macedonian alphabet with Latin letters. Therefore, there are rules according to several standards. The Macedonian transliteration is standardized with ISO R9:1968. This system was adapted and adopted in 1970 by the Macedonian Academy of Sciences and Arts and it is considered as officially accepted in the Republic of Macedonia. There are so far transliteration algorithms invented that work perfectly under standard circumstances. Standard circumstances are those when the writing of the texts follows already established standards such as „ѓ=gj“, „ж=zh“, „ѕ=dz“, „њ=nj“, „ќ=kj“, „ч=ch“, „џ=dj“ и „ш=sh“. This

presentation of the letters and diagraphs is simpler. However, there is other type of writing in which letters are called diacritics presented with a special sign, for example for letters ѓ, ќ, ч, ж, ш, s, џ - (ǵ, ḱ, č, ž, š, dz, dž). This is demonstrated in *Table 1.*

Table 1. Transliteration from Cyrillic to Latin alphabet under standard circumstances

Cyrillic	Latin	Process of transliteration	
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	ѕид	dzid
И и	I i	имот	imot
Ј ј	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

Table 2. Ambiguous Transliteration fr Cyrillic to Latin alphabet under non-stand circumstances

Cyrillic	Latin	Process of transliteratic	
А а	A a	авантура	avant
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	Gj gj	ѓавол	gavol
Е е	E e	елен	elen
Ж ж	Z z	жонглер	zongl
З з	Z z	збор	zbor
S s	Z z	ѕид	zid
И и	I i	имот	imot
Ј ј	J j	јубилеј	jubile
К к	K k	коска	koska
Л л	L l	леден	leden
Љ љ	L l	љубичица	lubici
М м	M m	манастир	mana
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
	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

о	О	O o	облека	oblek
п	П	P p	песна	pesn
р	Р	R r	разговор	razgo
с	С	S s	соба	soba
	Т т	T t	торба	torba
	Ќ ќ	K к	ќумур	kumu
	У у	U u	умерено	umer
ф	Ф	F f	форма	forma
	Х х	H h	хумор	humo
ц	Ц	C c	цвеќе	cveke
ч	Ч	C c	човек	covek
џ	Џ	J j	џамија	jamija
ш	Ш	S s	шеќер	seker

Under non-standard circumstances, very common in Macedonian area, the writing of letters in Latin alphabet can have two meanings in Cyrillic alphabet. So, for the Latin "s" we have Cyrillic "ш" or "с" and for Latin "z" we can have Cyrillic "з" or "ж". Hence, the sentence is unclear. This is demonstrated in *Table 2.*

Therefore, the end results are usually words with different meanings. For example, from "vesti" we derive "вешти" or "вести", from "dokazi" we derive "докази" or "докажи". There are a lot of other examples too. This instances are called "ambiguous transliteration".

Ambiguous transliteration

This paper is consisted of two smaller researches of ambiguous transliteration of full sentences from Latin to Cyrillic alphabet. The first research regards finding words that can have more than one meaning in the given alphabet. For that purpose we use a given base (dictionary) with

251460 words of Macedonian language including: verbs, names of locations and people, adjectives, nouns and other terms, and with the assistance of elaborated transliteration algorithm we obtained more than 5000 words with ambiguous transliteration.

Result

This means that more than 2 % of the words in the dictionary written in Latin alphabet have more than one meaning when transliterated in Cyrillic alphabet. This result of transliterated words has a great contribution to the structure of the language itself. When browsing those words in the given web service that makes difference between Cyrillic and Latin alphabet, the different meanings of the words is shown (for example забар, зелен, сок etc.). We will explain the procedure of transliteration of individual non-standard words with the following example. They should go through all these steps of transliteration demonstrated in diagram 1 and the obtained result shall be the ambiguous word.
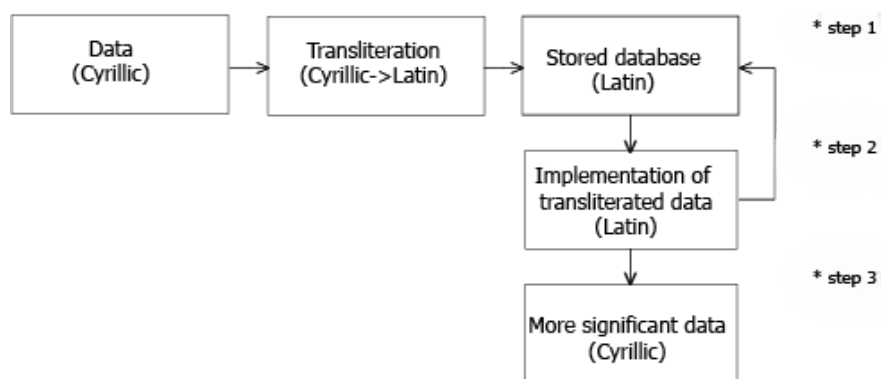
Diagram 1. Transliteration of individual words

Step 1: In the beginning, the transliteration of every data from the given dictionary commences. Let's take the noun {забар} that should be transliterated in Latin alphabet. This means that each character of the data is transliterated in the other alphabet {з,а,б,а,р -z,a,b,a,r) and it is saved in the database with its own unique key. This step is repeated for all words in the dictionary.

Step 2: Each transliterated word, transformed to Latin alphabet, is further selected and compared to the following transliterated word from the database

in order to check its ambiguity. It can be achieved by repeating the word {zabar} more than once.

Step 3: In this case the word {zabar} is placed in position of two nouns, which means that the end result has the following order {zabar}->{забар, жабар}. We can conclude that the final step enables a view of ambiguous term which basically is given in Cyrillic alphabet.

That is the part that we shall work on in order to properly transliterate full sentences and to obtain the real meaning of the full sentence. When transliterating only one word, the ambiguity of the words written on Latin alphabet cannot be resolved. However, when transliterating full sentences, a solution can be found.

Table 3. Overview of words with two meaning obtained by the steps of transliteration

Id	Ordinal number of the dictionary	Cyrillic	Latin	Words with two meaning
280	9279	бас	bas	bas(бас) - бас, баш
586	18295	важна	vazna	vazna(важна) - важна, вазна
667	20687	вести	vesti	vesti(вести) - вести, вешти
2110	42856	докази	dokazi	dokazi(докази) - докажи, докази
2355	53109	жаби	zabi	zabi(жаби) - жаби, заби
2472	54909	забар	zabar	zabar(забар) - жабар, забар
2836	69526	звуци	zvuci	zvuci(звуци) - звуци, звучи
2876	71706	знаци	znaci	znaci(знаци) - знаци, значи
3157	94910	каса	kasa	kasa(каса) - каса, каша
3262	97909	кожа	koza	koza(кожа) - кожа, коза
4304	175440	попусти	popusti	popusti(попусти) - попусти, попушти
4677	223624	сефот	sefot	sefot(сефот) - сефот, шефот

Solving ambiguous transliteration

We suggest one algorithm for intelligent solution of transliteration which can work under circumstances of full sentences transliteration. It is created in order to properly show the words with two meanings, obtained by transliteration under non-standard circumstances, in accordance with the grammar rules. In order to prove its practicability, another research was conducted related to transliteration of certain number of sentences downloaded from the Internet. These sentences contain different terms, adjectives, nouns, conjunctions etc. This is only part of the solution with particular steps required for resolving the existing problem i.e. the ambiguity of words. Firstly, a calculation of the K_i coefficient of words is conducted. Two formulas of ambiguity are used for that purpose. The first formula calculates the parameters which are not divided with 10%, whereas the second formula is with 10% The parameters are simple and are equally applied in the two formulas which are: the total number of sentences that contain one or more of the ambiguous words, the number of repetitions of all words in the sentence, the total number of all sentences that use ambiguous words. There are tests of 29 downloaded sentences conducted in the research. Those sentences contain the ambiguous word in Latin alphabet as "vesti" and "вести" or "вешти" in Cyrillic alphabet.

Formula 1. **Parameters without 10%**	*Formula 2.* **Parameters with 10%**
$K_i = P / S_1 + S_1 / S_2$	$K_i = (P / S_1 + S_1 / S_2) * 0,10$

In the following two table those 29 sentences are demonstrated, i.e. 16 sentences that contain the word "вешти" or 13 sentences that contain the word "вести". Using the given formulas we obtain results for the parameters of each word that are explained in details with the following 7 steps.

Table 4. Overview of 24 sentences that contain the word "вешти"

вешти	16	Repetitions	Sentences	Total sentences	Coefficient
вешти	на	12	0,75	0,55	1,30
вешти	минатата	1	0,06	0,55	0,61

вешти					
вешти	забава	1	0,06	0,55	0,61
вешти	во	10	0,63	0,55	1,18
вешти	Скопје	1	0,06	0,55	0,61
вешти	овие	2	0,13	0,55	0,68
вешти	жени	4	0,25	0,55	0,80
вешти	им	1	0,06	0,55	0,61
вешти	помагаа	1	0,06	0,55	0,61
вешти	болните	1	0,06	0,55	0,61
вешти	од	5	0,31	0,55	0,86
вешти	дневниот	1	0,06	0,55	0,61

Table 5. Overview of 24 sentences that contain the word "вести"

вести	13		Repetitions	Sentences	Total sentences	Coefficient
вести		најнови	3	0,23	0,448275862	0,68
вести		од	6	0,46	0,448275862	0,91
вести	Македонија		3	0,23	0,448275862	0,68
вести		и	5	0,38	0,448275862	0,83
вести		светот	4	0,31	0,448275862	0,76
вести		спортски	1	0,08	0,448275862	0,53
вести		забавни	1	0,08	0,448275862	0,53

вести	анализи	1	0,08	0,448275862	0,53
вести	интервјуа	1	0,08	0,448275862	0,53
вести	видео	1	0,08	0,448275862	0,53
вести	ги	1	0,08	0,448275862	0,53

This is followed by the steps for calculating K_i coefficient:

Step 1: Entering sentences in database, 29 sentences in this case.

Step 2: Sorting all words from all sentences where words can be repeated more than once.

Step 3: Calculation of **P -** parameter for repetition i.e. repetition of the words in the database.

Step 4: Calculation of only S_1 **-** number of sentences that contain the word "вешти" with previously appeared repetitions.

Step 5: Calculation of only S_2 **-** the total number of sentences that contain the words with two meanings "вешти" and "вести".

Step 6: Here the K_i **-** coefficient is calculated as summary of step 4 and step 5.

Step 7: Then, we write Macedonian sentences shown in Latin alphabet. In this case we have the sentence "Najnovi vesti od Makedonija i svetot". Each word of the sentences has its own coefficient. The summary of the coefficients gives the right word for transliteration.

All these steps described above regard table 4 and table 5, whereas the result of step 7 is obtained by these two tables.

Result

Out of the coefficients' sum for "вешти" we obtain the result 1,97 and for "вести" it is 3,86. This means that the sum with greater value represents the right word in some sentence or it gives the meaning of the whole sentence.

Table 6. Data for sentences of the word with two meanings

Najnovi vesti od Makedonija i svetot. (with sentences)

вешти	sentences (вешти)	вести	sentences (вести)
најнови	0	најнови	0,23
од	0,31	од	0,46
Македонија	0	Македонија	0,23
и	0,56	и	0,38
светот	0	светот	0,31
	0,87		1,61

Table 7. Data for coefficient of the word with two meanings in full sentences

Najnovi vesti od Makedonija i svetot. (with coefficient)

вешти	coefficient (вешти)	вести	coefficient (вести)
најнови	0	најнови	0,68
од	0,86	од	0,91
Македонија	0	Македонија	0,68
и	1,11	и	0,83
светот	0	светот	0,76
	1,97		3,86

For testing, a simple algorithm is elaborated which gives satisfactory results for determining the sense of the sentence and hence, it gives successful transliteration of full sentences. This is only one part of the solution which needs to be realized for transliteration of the words' meanings.

Conclusion

Based on the results obtained from conducted research, the following conclusions can be drawn.

The transliteration as reversible process enables individual transformation of letters from one alphabet to another. Due to the inability of several digital devices to use Cyrillic alphabet for writing of digital text, certain standards are used. Under some circumstances, there are words with two or

more meanings, and therefore, that produces obscurity in the meaning of the sentence. For that purpose, algorithm which can distinguish these words and calculate the coefficient for transliteration of words is used. This solution is possible if transliteration is applied for full sentences. So far, the calculation of the coefficient was not performed in digital form, for that reason the future job shall regard creating algorithm for transliteration of ambiguous words which can provide coefficient of words and with that the proper meaning of the full sentences as well. The application shall be the most useful in web services for transliteration of non-standard digital texts written in Latin to Cyrillic alphabet.

References

[1] Virga, Paola, and Sanjeev Khudanpur. "*Transliteration of proper names in cross-lingual information retrieval.*" Proceedings of the ACL 2003 workshop on Multilingual and mixed-language named entity recognition-Volume 15. Association for Computational Linguistics, 2003.

[2] Šimičević, Greta, and Ana Marija Boljanović. "*Transcription and Transliteration in a Computer Data Processing.*"

[3] *TRANSKRIPCIJA, TRANSLITERACIJA* I. "RASPRAVE ZJ, SV. 13 (1987), ZAGREB, 19-30 YU ISSN 0351—434x." Rasprave Zavoda za jezik IFF. 13 (1987): 19.

[4] Kuo, Jin-Shea, Haizhou Li, and Ying-Kuei Yang. "*Learning transliteration lexicons from the web.*" ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. Vol. 44. No. 2. 2006.

[5] Sherif, Tarek, and Grzegorz Kondrak. "*Substring-based transliteration.*" ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. Vol. 45. No. 1. 2007.

[6] Lin, Wei-Hao, and Hsin-Hsi Chen. "*Backward machine transliteration by learning phonetic similarity.*" proceedings of the 6th conference on Natural language learning-Volume 20. Association for Computational Linguistics, 2002.

[7] Antoniou, Grigoris, and Frank Van Harmelen. "*A semantic web primer.*" MIT press, 2004.

[8] Liyang, Yu. "*Introduction to the Semantic Web and Semantic Web Services.*" Taylor & Francis Group, 2007.

[9] Leon, Atkinson. "*Core PHP Programming.*" Prentice Hall PTR, ISBN: 0-13-089398-6, Second Edition, 2000.

[10] Vikram, Vaswani. "*PHP Programming Solutions.*" ISBN-10: 0-07-148745-X. McGraw-Hill, 2007.