Transcription and Transliteration in a Computer Data Processing

Greta Šimičević
Library, Faculty of Humanities and Social Sciences
Ivana Lučića 3, 10000 Zagreb, Croatia
gsimicev@ffzg.hr

Ana Marija Boljanović
Croatian Standards Institute
Ulica grada Vukovara 78, 10000 Zagreb, Croatia
ana.marija.boljanovic@hzn.hr

Summary

This paper shows different methods of transliteration of Cyrillic characters into Latin characters in bibliographic databases. The differences between transliteration and transcription have been presented. The paper draws attention to the problems that Latin databases are faced in the process of transliteration Cyrillic characters into Latin. Several examples of bibliographic databases, which apply different rules and standards – have been searched (the British Library, the Library of Congress, libraries in Croatia etc.). Searches were made according to certain criteria and the results have been presented. As the result of research there is a rate of application of certain existing rules and standards, which indicates that there is a need for standardization in this area and implementation of a unified system for transliteration at the international level.

Key words: conversion, transliteration, transcription, characters, digraphs, diphthongs, diacritical marks, standards, database

Introduction

In the course of the second half of 20th century computer catalogues, i.e. bibliographic databases replaced library catalogues on cards. Digitalization of library operations changed the way libraries function, which was most evident in the area of library material processing, i.e. data processing that it involves, its search possibilities and provision of user access to this data. During the said period little attention was paid to indexing and standardization, or to index languages and general data entry standards, as their purpose was barely understood. Databases were continually filled with data and became larger, clumsier and increasingly disorganized, resulting in maintenance problems, problems in data management and search problems. These increased even further once data-

bases became accessible to a large community via Internet. As the number of stored data continually increases, it has become essential to manage information systems in a precise manner, i.e. to select appropriate information language for storing and searching.

Amongst various other problems, bibliographic databases also faced problems of transcription and transliteration. Given that the subject of transfer of different scripts from one into another is both very broad and very demanding, in this paper we specifically cover the transfer of Russian Cyrillic characters into Latin, stipulating individual examples in practice of global bibliographic databases, with a particular overview of Croatian practice. We will try to specify problems that we discovered, possible methods for their solution through standardization, and certain discrepancies from international standards, as well as to explain the reasons for such practices and divergences.

Transcription and transliteration: possible procedures for transfer of one script into another

Transcription is a transfer of pronunciations and phonemes of one language into graphical system for phonetic recording of phonemes of another language, i.e. pronunciation of words in one language adapted to pronunciation in another language and to this other language's vocalization. Transcription respects phonetic characteristics of different languages and national variants, and need not necessarily involve transfer of one script into another, but may concern graphical transcript of words from one language into another even in cases when both language systems use the same script.[1] The transcription process is connected to a significantly narrower space than the global one, for it is often limited by a language system; in other words, by the rules (orthography) of the specific language system within which the process of transcription is being carried out. The most frequent differences between systems lie in the diverse phonetisation of certain graphemes that we transfer. For example: surname of the Russian author *Цветаева* looks in transcription of different language systems as follows: *Tsvetaeva* (Eng.), *Zwetajewa* (Ger.), *Cvetaeva* (Ita.), *Cvjetajeva* (Cro.), *Tswetaewa* (Pol.). The differences appear due to existence, or else lack of, specific graphemes and phonemes in different systems. For example, Latin Slavic languages have the diacritical characters *č, ž, š*, which other Latin language systems do not have so, they transfer Cyrillic characters for this phonetisation: *ч, ж, ш* into *ch, zh, sh*…etc. as well as Latin versions of these phonemes. Concurrently, Slavic language systems will transfer a diacritical character from another Slavic language as a diacritical character, for they both contain relevant

[1] Badurina, Lada, Ivan Makarović i Krešimir Mićanović. *Hrvatski pravopis.* Zagreb: Matica hrvatska, 2007., str. 221.

graphemes and phonemes as such. Similarly, German umlaut characters, such as for example *ü*, are transferred into Croatian language in a similar way, by applying phonetisation of the Croatian language system into *ue*, for umlaut characters as such do not exist in Croatian language system.

Russian surname *Щедрин* is another example, and in Croatian and Czech it is transferred as *Ščedrin*, in Polish it is *Szczedrin*, in English *Shchedrin*, in French *Chtchedrine*, in Dutch *Sjtsjedrin* and, in German *Schtschedrin*. This example shows that sometimes as much as seven letters of Latin alphabet are needed for the Cyrillic character *щ*, which makes it difficult to establish international catalogues or lists. This example also shows the issues concerning all four Russian Cyrillic diphthongs *я, ю, ё, щ* for which there are no graphemes in Latin alphabet. It is necessary to mention that a similar problem occurs in transfer of Latin diagraphs in other Latin language systems, or else into Cyrillic; such is the case for example with graphemes *dž, lj, nj, sch, ch*, as well as already mentioned diacritical characters and other special characters of Slavic and non-Slavic language systems that are not commonly accepted. Examples of such special characters in Russian Cyrillic are characters with strong or soft phonetisation, such as *э, ы, ь* and characters that existed throughout history of language; hereby noting that many Latin and other Cyrillic scripts abound with similar examples. There are recommendations on the global level that we may, but also need not accept. One of recommendations, for example, is to transfer the words from one language system into another in the same way that the language community of the former would transfer their words. However, this is primarily valid for idiomatic scripts.

From the above-mentioned it would arise that there would be as many transcription rules in the world as there are languages, so such transfer process from one script into another could not bring about uniformity on the global level.

Transliteration, on the other hand, is a transfer of characters (graphemes) of one script into characters of another script (e.g. from Glagolitic into Latin, from Cyrillic into Latin, etc.). This should occur almost automatically and both ways, so that the regress into original text should be possible. But, clearly – with 25 or 26 globally accepted characters in Latin script it is not possible to transfer 40 or 50 Cyrillic characters without occasional recourse to combinations of the usual Latin graphemes for the special characters.[2] The same symbols should not be used in transliteration of different characters in any language and, using two or more characters for one character is only acceptable when there is no better solution. As a possibility, transliteration has, on the global level, proven to be a much better procedure than transcription concerning harmonization of data entry into databases from other scripts into Latin, which brought about attempts at creating various international rules for transliteration.

[2] British standard BS 2979:1958. Transliteration of Cyrillic and Greek characters, BSI 1958.

Even though the definition of the procedure in itself should guarantee a simple and unambiguous solution of the problem of transfer in different scripts from one into another – given that the procedure itself should not be bound to any rules of various language traditions – this is not the case and in this procedure the issues of diversity in the use of graphemes and phonemes in language and script traditions becomes quite obvious.

Researching through various international bibliographic databases we found out numerous "inconsistencies" in application of different transliteration rules that were agreed on the global level. Examples of such "inconsistencies" in application of international transliteration rules are many but they generally boil down to accepted procedures linked to existence of large global language groups and their language or script traditions, which resulted in emergence of variants of international rules at the level of large language groups. This realization brings us inevitably to the application of transcription process within transliteration and, to the entire, already mentioned, body of issues that such practice brings about during transfer from one script to another, as well as to the hybridism of transcription and transliteration procedures. In transfer of Russian Cyrillic into Latin this is particularly manifest in the transfer of diphthongs *я, ю, ё, щ* and diacritical characters existent only in Slavic languages *ч, ж, ш* into Latin script of non-Slavic languages. There is also the problem of transfer of Russian graphemes such as *ц, х*, which are, for example, under a strong transcription tradition within databases of English speaking areas transferred as *ts, kh*, within German as *z, h/ch* (depending on the position of the character within a word)..., etc. Hence, we have – for example – transliteration procedure for one language group and transliteration procedure for another language group (e.g. transliteration procedure from Cyrillic into Latin for Slavic languages and transliteration procedure for non-Slavic languages).[3]

Standardization and other systems in the area of transliteration

In order to facilitate and improve communication and data and information exchange standards for transliteration of all international scripts into Latin script were developed by International Organization for Standardization (ISO) (see Table 1).

The fact that it has 162 member states speaks best about this international organization's significance. Experts from its member states contribute to the development of standards and standardization work and published standards are mostly adopted by the member states. As for international standard ISO 9 for transliteration of Cyrillic characters into Latin characters even back in 1954 the first issue of this standard was published. From the very beginning ISO 9 had the status of recommendation, established a provision within the text itself that

[3] As the result of research of the foreign bibliographic databases

the international standard may, in transliteration from Cyrillic into non-Slavic language, be amended or replaced with the national system accepted as usual practice. Nowadays, the third edition issued in 1995 is widely adopted by most European countries but with certain national modifications (e.g. Denmark, Deutschland, France, Italy, Poland, Russian Federation, Serbia, Sweden, Turkey, United Kingdom).

Table 1. List of valid international standards for transliteration

Document identifier	Title (English)	Publication date
ISO 9	Information and documentation – Transliteration of Cyrillic characters into Latin characters – Slavic and non-Slavic languages	1995-02-00
ISO 233	Documentation; Transliteration of Arabic characters into Latin characters	1984-12-00
ISO 233-2	Information and documentation; transliteration of Arabic characters into Latin characters; part 2: Arabic language; simplified transliteration	1993-08-00
ISO 233-3	Information and documentation – Transliteration of Arabic characters into Latin characters – Part 3: Persian language – Simplified transliteration	1999-01-00
ISO 259	Documentation; Transliteration of Hebrew characters into Latin characters	1984-10-00
ISO 259-2	Information and documentation – Transliteration of Hebrew characters into Latin characters – Part 2: Simplified transliteration / Note: Corrected and reprinted in 1995-07	1994-12-00
ISO 843	Information and documentation – Conversion of Greek characters into Latin characters / Note: Corrected and reprinted in 1999-05	1997-01-00
ISO 3602	Documentation; romanization of Japanese (kana script)	1989-09-00
ISO 7098	Information and documentation; romanization of Chinese	1991-12-00
ISO 9984	Information and documentation – Transliteration of Georgian characters into Latin characters	1996-12-00
ISO 9985	Information and documentation – Transliteration of Armenian characters into Latin characters	1996-12-00
ISO 11940	Information and documentation – Transliteration of Thai	1998-06-00
ISO 11940-2	Information and documentation – Transliteration of Thai characters into Latin characters – Part 2: Simplified transcription of Thai language	2007-05-00
ISO/TR 11941	Information and documentation – Transliteration of Korean script into Latin characters	1996-12-00
ISO 15919	Information and documentation – Transliteration of Devanagari and related Indic scripts into Latin characters	2001-10-00

Source: Perinorm International Database, British Standards Institute, 2009

Croatia had adopted almost all international standards for transliteration without any modifications. Even though ISO 9:1995 is adopted as the national standard in the Republic of Croatia, in practice of data entry into bibliographical data-

bases this standard is not applied consistently; entry practice is closer to ISO R 9:1968. Furthermore, the table which was given in ISO/R 9:1968 representing international system is in fact extended transliteration system for Serbian Cyrillic into Croatian Latin (see Table 2).

Table 2 Transliteration of the modern Russian alphabet (extracted from table ISO/R9)

TABLE 1. — Transliteration of the modern Russian alphabet

Russian					Transliteration	Examples		
Letter numbers	printed		written					
1	а	А	*а*	*А*	a	адрес	—	adres
2	б	Б	*б*	*Б*	b	баба	—	baba
3	в	В	*в*	*В*	v	вы	—	vy
4	г	Г	*г*	*Г*	g	голова	—	golova
5	д	Д	*д, д*	*Д*	d	да	—	da
6 [1])	е (ё)	Е (Ё)	*е (ё)*	*Е (Ё)*	e (ë)	ещё	—	eščë
7 [2])	ж	Ж	*ж*	*Ж*	ž	журнал	—	žurnal
8	з	З	*з, з*	*З*	z	звезда	—	zvezda
9	и	И	*и*	*И*	i	книга	—	kniga
10 [2])	й	Й	*й*	*Й*	j	первый	—	pervyj
11	к	К	*к*	*К*	k	как	—	kak
12	л	Л	*л*	*Л*	l	липа	—	lipa
13	м	М	*м*	*М*	m	муж	—	muž
14	н	Н	*н*	*Н*	n	нижний	—	nižnij
15	о	О	*о*	*О*	o	общество	—	obščestvo
16	п	П	*п*	*П*	p	пара	—	para
17	р	Р	*р*	*Р*	r	рыба	—	ryba
18	с	С	*с*	*С*	s	сестра	—	sestra
19	т	Т	*т, т, т*	*Т*	t	товарищ	—	tovariš
20	у	У	*у*	*У*	u	утро	—	utro

4

Source: ISO R/9:1968

In addition to standards there are also several global systems, which establish transliteration rules concerning practice of data entry into computer databases. Table 3 presents parallel transliteration systems by several different rules and/or recommendations, based upon Russian Cyrillic as example. Data entry into bibliographical databases in Croatia actually matches best to UN transliteration rules (see Table 3), with the exception of character *ë*, which is in current prac-

tice of Croatian bibliographical databases transliterated into *e*, and not into *ë* as required by the rule, probably due to the simple reason that this character is also often presented in Russian graphics with the grapheme *e*.[4] Such transliteration method can be traced back to the Rulebook and manual for preparation of alphabetical catalogues of Eva Verona from 1986, which was issued prior to the acceptance of international standards at the level of Republic of Croatia (see Table 3).

Table 3. Parallel overview of several transliteration rules and standards (extracted from Transliteration table)

Cyrillic	Scholarly	ISO/R 9:1968	GOST 1971	UN	ISO 9:1995; GOST 2002	ALA-LC	BGN/PCGN
А а	a	a	a	a	a	a	a
Б б	b	b	b	b	b	b	b
В в	v	v	v	v	v	v	v
Г г	g	g	g	g	g	g	g
Д д	d	d	d	d	d	d	d
Е е	e	e	e	e	e	e	e, ye †
Ё ё	ë	ë	yo	ë	ë	ë	ë, yë †
Ж ж	ž	ž	zh	ž	ž	zh	zh
З з	z	z	z	z	z	z	z
И и	i	i	i	i	i	i	i
Й й	j	j	j	j	j	ĭ	y
К к	k	k	k	k	k	k	k
Л л	l	l	l	l	l	l	l
М м	m	m	m	m	m	m	m
Н н	n	n	n	n	n	n	n
О о	o	o	o	o	o	o	o
П п	p	p	p	p	p	p	p
Р р	r	r	r	r	r	r	r
С с	s	s	s	s	s	s	s
Т т	t	t	t	t	t	t	t
У у	u	u	u	u	u	u	u
Ф ф	f	f	f	f	f	f	f
Х х	x	ch	x	h	h	kh	kh
Ц ц	c	c	cz, c	c	c	t͡s	ts
Ч ч	č	č	ch	č	č	ch	ch
Ш ш	š	š	sh	š	š	sh	sh
Щ щ	šč	šč	shh	šč	ŝ	shch	shch

[4] As the result of research of the Croatian bibliographic Databases

Table 3 cont.

Cyrillic	Scholarly	ISO/R 9:1968	GOST 1971	UN	ISO 9:1995; GOST 2002	ALA-LC	BGN/PCGN
Ъ ъ	"	"	"	"	"	"	"
Ы ы	y	y	y'	y	y	y	y
Ь ь	'	'	'	'	'	'	'
Э э	è	ė	eh	è	è	ė	e
Ю ю	ju	ju	yu	ju	û	i͡u	yu
Я я	ja	ja	ya	ja	â	i͡a	ya
Pre-1918 letters							
I i	i	i	i, i' **	ĭ	ì	ī	–
Ѳ ѳ	f	ḟ	fh	ḟ	f̀	ḟ	–
Ѣ ѣ	ě	ě	ye	ě	ě	i͡e	–
V v	i	ẏ	yh	ẏ	ỳ	ẏ	–

Source: http://en.wikipedia.org/wiki/Romanization_of_Russian#Transliteration_table

Computer systems and transliteration

The most important task in data processing is how to store all information contained in a unit of the material in such a way as to make it easily searched and successfully found at the request of a user. Computer systems differentiate various data by distinctiveness of characters. Thus it can happen that one and the same information entered into the computer system via both the transcription process and transliteration process would, in fact, signify two different pieces of information for the computer. If at the same time several different rules are applied for transcription and transliteration, we could from one semantically identical data item create, as far as computer is concerned, a multitude of different data items. This is particularly important for the entry of normative data, and for indexing, which in concrete terms of bibliographical databases represents data on authoring, subject, etc. Subsequent search results will depend exclusively on that, which rules have been applied to enter specific data into the computer system, and which rules have been applied to define the search. Researching through the largest and the most prominent bibliographic databases in the world and in the region we have established that in practice ISO international standard for transliteration from Russian Cyrillic into Latin is never completely and fully implemented. Most frequently this standard represents merely a basis upon which other rules generally linked with the transcription process are built, or which are imposed by certain large language groups and their language traditions. In this way, what may be called national variants of ISO standard emerged, which are then consistently applied in the majority of researched databases. Within narrow national levels such script transfer method functions very well, for it is familiar to this national group's users. But, the problem occurs on the global level, given that databases have now outgrown the national

level. Global user does not have access to the data transferred into Latin script through the transcription process; or even through transliteration process heavily influenced by the transcription tradition, especially if the information on the transfer method is lacking. Hence, a problem has been noted concerning search where transparent information on the rules based upon which transliteration has been implemented within certain database is lacking. As it were, this information does in fact exist, but it is hidden within encoded fields of entry and it is inaccessible for regular user.

There are numerous user oriented Internet pages covering issues of transliteration, and offering one-stop-shop information on various rules for transliteration of all scripts, Russian Cyrillic included. In addition to tabular overview of standards, many also have built-in software for automatic transliteration of Cyrillic characters, as well characters of other scripts into Latin (and vice versa) by different standards.[5]

Conclusion

Whilst trying to find an answer to the question why the practice on international level does not apply the single transliteration standard that exists and that has been adopted by consensus exactly for the purpose of bridging the recognized problems that arose from publication of bibliographic databases on the global level, we found that the answers are self-evident. One of the first facts is that bibliographic databases were established and were becoming larger and larger way before the problem of transfer of scripts on the global level became recognized. Subsequently adopted international standards became sort of an attack on language traditions of large groups. Additionally, it was very difficult to adopt a standard that would reconcile all language traditions. One of the larger problems is also the need to use a multitude of special characters whose perusal used to be far from simple. If today we tried to apply the unified standard, we would find numerous problems in translating the data that has already been entered. Naturally, all of this would not justify why all databases should not start applying single standard in the future, which would greatly facilitate search for the global user and would make many currently "unavailable" information accessible.

Transliteration process should serve as a unique technical aid in a transfer of characters of one script into characters of another script regardless of linguistic rules and traditions of any existing language system. Accordingly, this process should not produce a system for the original reading and writing, but the system for conversion of written sources in other written form, for its recording, storing and searching in another script, with the possibility of regress into the original

[5] Examples available through following web sites:
http://www.russki-mat.net/trans2.html
http://www.allmend-ru.de/etc/transliteration.html

system of characters. Tendency towards greater unification of standards on the global level continues to exist. Aspirations towards use of transcription process, which makes harmonization process on the wider level impossible, have been overcome and transliteration process, which is even with all transcription interventions on national levels still more homogenous than it would have been possible with transcription, has been completely accepted. This in itself facilitated user-friendly access to information that has been transferred from other scripts into Latin databases.

References

British standard BS 2979:1958. Transliteration of Cyrillic and Greek characters, BSI 1958

Badurina, Lada, Ivan Makarović i Krešimir Mićanović. *Hrvatski pravopis.* Zagreb: Matica hrvatska, 2007.

Deutches standard DIN 1460:1982. *Conversion of cyrillic alphabets of slavic languages,* DIN 1982.

International standard ISO/R 9:1954. *International system for the transliteration of Cyrillic characters,* ISO 1954.

International standard ISO/R 9:1968. *International system for the transliteration of Slavic Cyrillic characters*, ISO 1968.

Croatian standard HRN ISO 9:1995. *Information and documentation – Transliteration of Cyrillic characters into Latin characters – Slavic and non-Slavic languages,* ISO 1995

Katalog Nacionalna i sveučilišna knjižnice u Zagrebu http://katalog.nsk.hr/cgi- bin/Pwebrecon. cgi?DB=local&PAGE=First (2009-05-14)

Knjižnice iz sustava Cobiss Net http://www.cobiss.net/default-SR.asp (2009-06-08)

Library of Congress Online Catalog http://catalog.loc.gov/ (2009-06-17)

Parallel overview of several transliteration rules and standards, http://en.wikipedia.org/wiki/ Romanization_of_Russian#Transliteration_table (2008-02-12)

Perinorm International Database on CD ROM, British Standards Institute, 2009.

Russian standard GOST 7.79:2000 *System of standards on information, librarianship and publishing. Rules of transliteration of Cyrillic script by Latin alphabet,* GOST 2000.

Russki-mat.net. *Transliteracija russkogo alfavita: Transliteration and transcription using the latin alphabet*. http://www.russki-mat.net/trans.htm (2008-02-12)

Russki-mat.net. *Transliteracija russkogo alfavita: Automatic transliteration of Russian.* http://www.russki-mat.net/trans2.html (2008-02-12)

The European Library http://search.theeuropeanlibrary.org/portal/en/index.html (2009-05-20)

Verona, Eva. Pravilnik i priručnik za izradu abecednih kataloga. 1. dio, Odrednice i redalice. Zagreb : HBD, 1986.