***Sabina V. Mammadzadeh***
Institute of Information Technology of ANAS, Baku, Azerbaijan
sabinamammadzadeh@gmail.com

DOI: 10.25045/jpis.v09.i1.11

# TRANSLITERATION AND TRANSCRIPTION IN DATA PROCESSING

*This article highlights various transliterations of non-Latin graphics into Latin script in the bibliographic databases. The differences between the transliteration and transcription are explained. The bibliographic database examples which apply different conversion rules and standards are presented. Application level of existing guidelines and standards is determined based on conducted research. Relevance of acceptance of standardization system and a single transliteration system on international level is emphasized.*

***Keywords****: conversion, transliteration, transcription, characters, diacritics, diphthong, standard, databases.*

## Introduction

In the second half of the twentieth century, electronic catalogs and bibliographic databases began to replace traditional library catalogs. Digitization of library activities has completely changed their functionality, including resource search and access ways. Databases were then accessible to broad public via the Internet. With the increasing number of stored data, there was a need for more efficient management of information system, namely selection of appropriate information language for storage and search. The diversity of information and databases entries in different languages necessitate the use of single transcription and transliteration tables. Given that the conversion of various graphs is a very broad and significant field, this article shows transliteration of non-Latin alphabets into the Latin alphabet separately illustrating global bibliographic databases. In addition, existing problems in the field of transliteration are identified, and possible solutions are provided through their standardization.

## Transcription and transliteration: possible converting procedures

Transcription is the transformation of pronunciation and phonemes of one language into a graphic system of another according to its phonetics. In other words, transcription is the adaptation of the words in one language into the pronunciation and vocalization of another. Transcription should take into account phonetic characteristics of different languages and national variations, and even if the graphics are the same in both languages only, the graphics conversion should be implemented [1].

Transcription process is often limited to a language system, i.e. orthography of specific language for which transcription process is carried out. The main difference between these systems is that certain graphics are sounded in different ways. For instance, Russian last name *Цветаева* is converted differently in transcription of various language systems such as *Tsvetaeva* (eng.), *Zwetajewa* (ger.), *Cvetaeva* (ita.), *Cvjetajeva* (hung.), *Tswetaewa* (pol.).

These differences occur due to certain graphics and phonemes that present in separate systems. For example, there are some diacritics Azerbaijani, Serbian and Turkish languages using the Latin alphabet such as "ç", "ş", "ğ", "č", "š", "ž", and therefore, they should be phonetically converted (e.g., ch, sh, gh, zh, etc.). The letters with "umlaut" ("sound alteration") in the German language are similarly transcribed. For example, the German letter "ü" is transcribed as "ue" into the Croatian language.

Another example is Russian last name *Щедрин*. It is transcribed as *Şedrin* into the Azerbaijani language, as *Ščedrin* – into the Swedish, Czech and Croatian, as *Szczedrin* - in Polish, *Shchedrin* - in English, *Chtchedrine* - in French, *Sjtsjedrin* - in Danish, and *Schtschedrin* - in German. It is observed that only Cyrillic letter "*щ*" is written in 7 different ways in Latin graphics. This, in turn, complicates the adoption of international catalogs or conversion tables.

In addition, some languages have soft signs such as "*ь*", "*ñ*" and their transcription are also challenging.

Given the abovementioned, it can be concluded that the number of transcription rules may be appropriate to the number of the languages available in the world. Thus, international uniform conversion for transcription from one alphabet to another is impossible to achieve.

Unlike transcription, transliteration is the conversion of letters (graphemes) of one scripts into the letters of another. This process is almost automatic, and reverse transliteration is also possible. Naturally, it is impossible to transliterate commonly accepted 25 or 26 Latin letters into 40 or 50 Cyrillic letters without the use of ordinary combination of Latin graphemes. In this case, one character can be replaced by two or more characters [2]. Generally, transliteration is considered to be more appropriate procedure rather than transcription in terms of harmonization. Therefore, different international rules are adopted in the field of transliteration. Nevertheless, a simple and adequate solution should be found for the conversion of one script into another.

Study of various international bibliographic databases shows that some contradictions occur in the application of separate transliteration rules. These contradictions are largely associated with broader global language groups and their alphabetic traditions, which have resulted in the adoption of international transliteration rules for large language groups. All these processes generate the hybrid of transliteration and transcription procedures. This is mostly observed in the transliteration of Cyrillic diphthongs, such as “*я*”, “*ю*”, “*ё*”, “*ц*”, including the Azerbaijani diacritics “*ç*”, “*ş*”, “*j*” into the Latin script. For example, the letters “*ц*” and “*х*” are transliterated as "*ts*" and "*kh*" in the English-language databases, and as "*z*", "*h*" or "*ch*"in German-language databases, depending on their position within the word.

## Standardization in the field of transliteration and other systems

International Organization for Standardization (ISO) adopts standards for the transliteration of all international alphabets into the Latin alphabet in order to facilitate and improve communication and information exchange (Table 1) [4].

The importance of this organization is particularly emphasized by 162 ISO member countries, and the standardization is verified by those states. In accordance with the international standard *ISO 9*, the transliteration standard from Cyrillic into the Latin alphabet was published in 1986 [5]. From the very beginning of its establishment, documentation of the organization sets out that it can be modified or replaced by a national system in line with international standards in the field of transliteration. At present, the standard for the transliteration from Cyrillic into Latin, originally published in 1995, is adopted by many European countries (Denmark, Germany, France, Italy, Poland, Russian Federation, Serbia, Sweden, Turkey, Great Britain, etc.). For example, Croatia adopted almost all international standards for transliteration without any modification.

In addition to the standards, various global transliteration systems are applied to the database for data entry. Table 2 shows examples of different parallel translation systems.

Former Soviet Union member countries also used ISO 9 standard (Russia, Azerbaijan, Armenia, Belarus, Kazakhstan, Kyrgyzstan, Tajikistan, Turkmenistan, Uzbekistan (*GOST 7.79*)) [6].

Russia, the United States and Great Britain have a great experience in the field of transliteration. The first standard for Cyrillic graphic in this area was developed by the General Directorate of Geodesy and Cartography under the USSR Council of Ministers [7]. Other transliteration systems currently supported by the Russian language include *BGN / PCGN, ALA-LC, GOST (1983) / UN (1987), ISO 9* and *scholarly* transliteration systems.

Table 1

International standards for transliteration adopted by ISO [5]

| Standardst and progects | stages | TC |
|---|---|---|
| ISO 9:1995<br>Information and documentation -- Transliteration of Cyrillic characters into Latin characters -- Slavic and non-Slavic languages | 90.60 | ISO/TC 46 |
| ISO 233:1984<br>Documentation -- Transliteration of Arabic characters into Latin characters | 90.60 | ISO/TC 46 |
| ISO 259:1984<br>Documentation -- Transliteration of Hebrew characters into Latin characters | 90.60 | ISO/TC 46 |
| ISO 843:1997<br>Information and documentation -- Conversion of Greek characters into Latin characters | 90.60 | ISO/TC 46 |
| ISO 7098:2015<br>Information and documentation -- Romanization of Chinese | 60.60 | ISO/TC 46 |
| ISO 9984:1996<br>Information and documentation -- Transliteration of Georgian characters into Latin characters | 90.60 | ISO/TC 46 |
| ISO 11940:1998<br>Information and documentation -- Transliteration of Thai | 90.60 | ISO/TC 46 |
| ISO 15919:2001<br>Information and documentation -- Transliteration of Devanagari and related Indic scripts into Latin characters | 90.60 | ISO/TC 46 |

*BGN / PCGN* System - is the transliteration standard for the conversion of geographical names of Russia adopted by the United States Board on Geographic Names (BGN) in 1944 and by the Permanent Committee on Geographical Names (PCGN) of the British Government in 1947 [8].

*ALA-LC* - Romanization standard for other writing systems using Latin graphics. This system was adopted by the *American Library Association (ALA)* and the *Library of Congress (LC)* and was last updated in 1997 [9]. The system is used to provide accurate writing of bibliographic names in both North American and British Libraries, including English-language publications.

*GOST (1983) / UN (1987)* - recommended by the United Nations and based on the official system adopted by the 1984 General Directorate of Geodesy and Cartography under the USSR Council of Ministers and the United Nations Declaration of 1987 (*V/18* ) [10].

In addition, *Doc 9303* standard [11] approved by the *International Civil Aviation Organization (ICAO)*, which is the United Nations body, and *BS 2979* standard approved by *Oxford University Press* and the *British Library* in 1958 [12] were adopted for the transliteration of international telegraph standards, passports and driving licenses.

| Common systems for Romanizing Russian | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Cyrillic | | Scholarly | ISO/R9:1968; GOST 1971(1); UNGEGN (1987) | GOST 1971(2) | ISO9:1995; GOST 2002(A) | GOST 2002(B) | 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)[9] |
| **Ё** | **ё** | ë | ë | jo | ë | yo | ë | ë (yë)[9] |
| **Ж** | **ж** | ž | ž | zh | ž | zh | zh | zh |
| **З** | **з** | z | z | z | z | z | z | z |
| **И** | **и** | i | i | i | i | i | i | i |
| **Й** | **й** | j | j | jj | 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 (h) | h (ch)[2] | kh | h | x | kh | kh |
| **Ц** | **ц** | c | c | c | c | cz (c)[3] | t͡s | ts |
| **Ч** | **ч** | č | č | ch | č | ch | ch | ch |
| **Ш** | **ш** | š | š | sh | š | sh | sh | sh |
| **Щ** | **щ** | šč | ŝ (šč)[2] | shh | ŝ | shh | shch | shch |
| **Ъ** | **ъ** | ″ | ″ | ″ | ″ | ″ | ″[6] | ” |
| **Ы** | **ы** | y | y | y | y | y' | y | y |
| **Ь** | **ь** | ′ | ′ | ′ | ′ | ′ | ′ | ’ |
| **Э** | **э** | è | ė (è)[2] | eh | è | e' | ė | e |
| **Ю** | **ю** | ju | ju | ju | û | yu | i͡u | yu |
| **Я** | **я** | ja | ja | ja | â | ya | i͡a | ya |
| **І** | **і** | i | i | – | ì | i (ih, i’)[4] | ī | – |
| **Ѳ** | **ѳ** | f (th)[1] | ḟ | – | f̀ | fh | ḟ | – |
| **Ѣ** | **ѣ** | ě | ě | – | ě | ye | i͡e | – |
| **Ѵ** | **ѵ** | i (ü)[1] | ẏ | – | ỳ | yh | ẏ | – |
| **Ѕ** | **ѕ** | dz (ʒ)[1] | – | – | ẑ | js | – | – |
| **Ѯ** | **ѯ** | ks | – | – | – | – | – | – |
| **Ѱ** | **ѱ** | ps | – | – | – | – | – | – |
| **Ꙍ** | **ѡ** | ô (o)[1] | – | – | – | – | – | – |
| **Ѫ** | **ѫ** | ǫ (u)[1] | – | – | ǎ | – | – | – |
| **Ѧ** | **ѧ** | ę (ja)[1] | – | – | – | – | – | – |
| **Ѭ** | **ѭ** | jǫ (ju)[1] | – | – | – | – | – | – |
| **Ѩ** | **ѩ** | ję (ja)[1] | – | – | – | – | – | – |

## Transliteration Information Systems

One of the problems encountered in data processing is improper storage of materials in the database, which complicates the search for the relevant results. Computer systems distinguish different data based on its letters. Thus, the same information entered the computer system through both transcription and transliteration can be presented as two different information by computer. Simultaneously, the same semantic information can be distinguished with different rules for transcription and transliteration. This is particularly important for normative data entries and indexing, i.e., providing accurate information about authors in the bibliographic database. Search relevance also depends on the rules, according to which specific data is included in the computer, and on the search guidelines. The study of large and most popular bibliographic databases shows that *ISO* standards for multilingual transliteration have not been fully implemented. This is largely due to transcription processes and some extensive language groups and their linguistic traditions. In this regard, various national standards of ISO standards have been adopted. These methods are widely used in ethnic minorities.

In addition, web-pages that offer transliteration tables and automatic transliteration software for all languages are also available. Since the 1990s, various transliteration sites have been developed with the expanded use of transliteration on the Internet. These online converting sites may include *http://softario.com/typus.html* (current version supports Arabic, Armenian, Estonian, Georgian, Greek, Hebrew, Kazakh, Russian and Ukrainian languages), *http://cesty.in/transliteration* (Cyrillic), *http: //transliterate.com/* (Greek and Hebrew), *http://translate.malerkotla.co.* (Hindi and Urdu), *http://transliterations.info* (Greek, Hebrew, Thai, Arabic, etc.), and *http://vikku.info* (Punjabi, Hindi, Tamil, Sanskrit, etc.). One of the steps taken in this field in Azerbaijan is the website *http://transliterasiya.az*, developed by the researchers of the Institute of Information Technology of ANAS in 2011. Transliteration of words or texts and website (page) with the entry of URL from Azerbaijani into five scripts (Russian, English, German, French and Persian) is available here. Transliteration tables used in the site have been compiled together with experts of relevant fields [13].

On such Internet pages, transliteration tables of the Azerbaijani scripts are also found. One of them is the transliteration table adopted by the American Congress Library for the transliteration of the Azerbaijani script into Latin script [14]. Moreover, another transliteration table for the Azerbaijani script is adopted by the European Parliament Terminology Coordination [15]. It should be noted that the European Parliament Terminology Coordination’s main role is to assist translators with their day-to-day tasks and facilitate terminology research and terminology management in the translation units, and to increase the EP’s contribution to the EU terminology database IATE [16].

## Recommendation regarding the adoption and application of the transliteration standard of the Azerbaijani script into the Latin script

Recommendations on the adoption and application of the transliteration standard of the Azerbaijani script into the Latin script must be jointly provided by the Institute of Linguistics and the Institute of Information Technology of ANAS. This initiative should be motivated by the following facts:

1. The Latin alphabet is mainly used for messaging and communication via the Internet.
2. There are some misunderstandings as a common transliteration table for the Azerbaijani language hasn’t been adopted yet. People who communicate and speak in Azerbaijani, especially when they communicate via the Internet, type Azerbaijani words in the Latin script in several versions.

   Such problems are not available in countries such as Georgia, Russia, Israel, China, or Japan that have adopted a single system for using the Latin script. In these countries, a common method for transliteration is used.

3. The use of the Latin alphabet does not mean abandoning the Azerbaijani alphabet. The application of a single standard is intended to facilitate writing and reading in the Latin script. Thus, this standard will simplify the work of people using the Latin script for technical reasons.
4. The application of this standard may also provide transliteration spell check.
5. The letters on foreign keyboard are not based on the state standard for transliteration, thus the use of the Azerbaijani keyboard is complicated.

With the help of this standard, communication with people whose computers do not support the Azerbaijani script or those who do not know the Azerbaijani script can be comforted. Introduction of a single standard and automatic algorithms for transliteration of the Azerbaijani script into the Latin and vice versa should be developed.

A number of issues need to be addressed to adopt transliteration standard and solve the above-mentioned problems. It should be noted that the language should be taken into account when designing the transliteration standards. Therefore, it has to be discussed and used in practice for a period of time before adoption, and several comprehensive surveys should be conducted. The results of the survey need to be discoursed in forums, recommendations and suggestions on possible solutions of the problem should be put forward.

## Conclusion

The ever-expanding bibliographic databases necessitate the international settlement of the transliteration issue. Internationally adopted standards have a negative impact on the linguistic tradition of broader language groups. Furthermore, adopting a standard that is appropriate for all language traditions is a very complicated issue. Naturally, the future implementation of transliteration process for all databases, based on a single standard, can facilitate the process of search and, ultimately, prevent the spread of different forms of the same data.

The article highlighted the importance of storage and search of any resource of one script in other scripts, taking into account that the transliteration is the process conversion of letters regardless of linguistic rules and traditions of any existing language system. In this regard, the article studied different versions of transliteration from non-Latin scripts into Latin, and provided examples of bibliographic databases adopted different transliteration guidelines and standards. The standardization of transliteration in Azerbaijan and adoption of a common transliteration system at the international level were substantiated.

## References

1. Badurina, Lada, Ivan Makarović i Krešimir Mićanović. Hrvatski pravopis. Zagreb: Matica hrvatska, 2007., str. 221.
2. British standard BS 2979:1958. Transliteration of Cyrillic and Greek characters, BSI 1958, https://www.shop.bsigroup.com/ProductDetail/?pid=000000000000090651
3. US Library of Congress Romanization Tables, https://www.loc.gov/catdir/cpso/romanization/azerbaij.pdf
4. https://www.iso.org/ics/01.140.10/x/
5. https://www.iso.org/standard/3589.html
6. https://en.wikipedia.org/wiki/Romanization_of_Russian
7. Technical reference manual for the standardization of geographical names United Nations Group of Experts on Geographical Names
8. http://www.translitteration.com/transliteration/en/russian/bgn-pcgn/
9. http://www.loc.gov/aba/about/
10. http://www.unstats.un.org/UNSD/geoinfo/UNGEGN/docs/_data_ICAcourses/_HtmlModule/_Selfstudy/S11/S11_01.html
11. http://www.icao.int/Security/mrtd/pages/Document9303.aspx

12. Shiloh A. The plague of print. New Scientist (No.284) 1962, p.179
13. http://www.transliterasiya.az
14. https://www.loc.gov/catdir/cpso/romanization/azerbaij.pdf
15. http://www.transliteration.eki.ee/pdf/Azerbaijani.pdf
16. http://www.termcoord.eu