*Corresponding Author's Email: maxvakul@yahoo.com
Proceedings of the World Conference on Foreign Language Education
Vol. 1, Issue. 1, 2024, pp. 1-11
DOI: https://doi.org/10.33422/worldfle.v1i1.545

ISSN: 3030-105X online

Diamond Scientific Publishing

Transliteration of Non-Latin Texts: From Everyday Practice to Linguistic Technologies

Maksym Vakulenko
Darmstadt University of Applied Sciences, Darmstadt, Germany

Abstract

This paper discusses various transcoding systems that convert non-Latin texts into Latin script. Particularly significant is the Romanization of Slavic languages. The Latinization systems generally fall into two classes: those based on the transliteration approach and those based on practical transcription. During transliteration, the main attention is paid to simple correspondence (mutual unambiguity) between original and converted characters that allows for text reversibility, i. e. the ability to restore the original text after re-transliteration. During practical transcription, the primary concern is the sound of words. In the last scenario, it is not always possible to restore the original text. The significance of transliteration extends to historical texts written in non-Latin scripts as well. Latinization systems are broadly utilized in multilingual Natural Language Processing tasks, resulting in their wider use and enhanced need.

Keywords: transliteration, practical transcription, alphabet, proper name, Cyrillic, orthography, standardization, simple correspondence, historical texts, Natural Language Processing.

1. Introduction

The idea of creating a universal transliteration system into the Latin alphabet was put forward centuries ago by the British linguist Richard Lepsius (1863). Over time, the countries whose languages are based on non-Latin alphabets have developed Romanization systems that are used in their international communication. As the necessity to render a word or text in another alphabetic system arises whenever languages with different alphabetic systems come into contact, the transliteration questions play an important role today. Such contacts are becoming more and more intense, so the issue of transliteration is urgent.

This paper aims to collect and summarize information about transliteration principles and how they are used in Slavic and other languages. To do this, we will consider the transliteration standards of different countries, as well as other Latinization systems and approaches. Particular attention will be given to the principles of transliteration, especially the requirement for simple correspondence (isomorphism) between the original and transliterated texts, which

ensures the possibility of recovering the original text after reverse conversion (re-transliteration). The analysis and comparison of diverse transliteration solutions takes into account etymological, phonetic, legal, and practical considerations.

2. Basic Definitions and General Considerations

Among various definitions of transliteration, the most precise and informative appears as the one describing transliteration as the "action, process, or result of converting one set of signs to another, usually involving at least one set of alphabetic letters" (McArthur 2018; see also ISO 3602:1989; ALA-LC Japanese 2022: 1).

This type of script conversion, known also as *scholarly* or *scientific transliteration*, is based on the principle of mutual unambiguity between the original graphemes and transliterated characters (Maslov 1987: 284). Mutual unambiguity is crucial for preserving the information contained in the text after re-transliteration (reverse transliteration). This relationship is exemplified by the *vukovica* (Cyrillic alphabet) and *gajica* (Latin alphabet) used in the languages of the former Jugoslavija, where words can be written interchangeably depending on the needs. This characteristic is fundamental for transliteration standards.

The core idea of transliteration is the conversion of written characters, with minimal consideration for their readout in other languages (Reformatskij 1960: 96; see also Superanskaja 2018: 29). Scholarly transliteration does not aim for the "most convenient" or "easiest" reading of the converted text in another language. This approach is common in applied linguistics.

In machine learning, however, "transliteration" often refers to "the process of converting terms written in one language into their approximate spelling or phonetic equivalents in another language" (Kirschenbaum and Wintner 2009: 433). This process is more accurately described as *Latinization* or *Romanization*, which includes both *transliteration* and *practical transcription* (see ISO 3602:1989). The practical transcription is a method of rendering words from one language using the conventional orthography of another language for practical purposes (see Superanskaja 2018: 3). When this process involves converting graphemes to phonemes, it is called *phonemization*, and when the graphemes are converted into notations of speech sounds (phones), it is nothing but *phonetization*.

The goal of transliteration is to represent the graphemes of the original text using the characters of the target script. It is essential for rendering proper names (especially *toponyms* and *anthroponyms*) and exoticisms from one language into another, as these often carry cultural and historical nuances that can be lost in translation. Typically, transliterated words enter the recipient language baser on of *orthographic transplantation* (see Superanskaja 2018: 29) which does not involve any grammatical post-adaptation. This post-adaptation may (but must not) happen at a later stage, when and if the recipient language adapts the loanword to its norms.

3. Transliteration of Cyrillic Texts in the Slavic and Related Languages

The Greek Latinization standard (ISO 843:1997) contains two tables. One of them provides transliteration and is recommended for bibliographic records, information technology use, directory and communication services, and other areas where the original text needs to be reconstructed. The second one provides practical transcription envisioned for passports, road signs, maps, and other cases where the correct pronunciation takes precedence over the reconstitution of the original form. However, this transcription is not oriented to any other language and meets the requirement of reversibility.

The transliteration standard ISO 9:1995 (ISO 9:1995) establishes, on the principle of one-to-one conversion, transliteration rules for Slavic and non-Slavic languages that use the Cyrillic alphabet. This standard offers two systems: System A, which uses Latin letters with diacritical marks, and System B, which employs combinations of basic Latin letters within the ASCII range of 0 – 127. For example, the Russian name *Сергей* is transliterated as *Sergej*, reflecting the Slavic written tradition.

This standard is widespread in Europe, and some countries have developed associated national standards. In particular, the Slavic names are rendered in German in accordance with the NF Z46-001:1995-06-01 standard (NF Z46-001) which is linked to ISO 9:1995. In Czech, a similar standard is employed for the transliteration of Cyrillic characters (ČSN ISO 9 2002), which introduces the norms of ISO 9:1995 as a National Czech standard.

However, some involved countries still use the non-isomorphic Latinization rules tailored to the English language. In Bulgaria, for instance, the "Transliteration Law" mandates practical transcription into English (Transliteratsiyata 2009), despite a strong Slavistic Romanization tradition in the country (Tilkov 1982: 238; Gešev 2016). These rules do not provide the bijective (isomorphic) transcoding, leading to difficulties in recovering the original text. For example, names like *Канчев* and *Кънчев* are both Romanized as *Kanchev*, making them indistinguishable (Gešev 2016).

In Ukraine, the ISO 9:1995 regulations are not common due to their limited accessibility. Instead, Resolution No. 55 of the Cabinet of Ministers of Ukraine establishes Ukrainian-English practical transcription as the Latinization norm (Kabinet Ministriv Ukrajiny 2010). This Romanization system, geared towards English, has typical drawbacks of such transcriptional approaches: it violates the requirement of bijectivity between the original and transliterated texts for letters <*г*>, <*і*>, <*ї*>, <*й*>, <*ц*>, iotated letters, and the softening sign <*ь*> (resulting from its ignorance); it neglects the Slavic Latin writing tradition; and it uses <h> both as a modifier and a separate letter which is not allowed in transliteration standards. These issues lead to systematic distortion and misidentification of names, such as *Возіанов – Возянов* (Vozianov), *Ліліана – Ліляна* (Liliana), *Ляльо – Ліалко* (Lialko), *Маріан – Мар'ян* (Marian), *медіана – медяна* (mediana), *Пії – Пій* (Pii), *Лар'їн – Ларін* (Larin), *Ільїн – Ілін* (Ilin), *Чорнобаї – Чорнобай* (Chornobai), *Юр'ївка – Юрівка* (Yurivka), *Воронько – Воронко* (Voronko), *Гальченко – Галченко* (Halchenko), *Паньківська – Панківська* (Pankivska), *Польова – Полова* (Polova), *Левитський – Левицький* (Levytskyi). As a result, the original form cannot be restored after reverse transliteration.

The correspondence <г> - <h> does not have a proper historical and phonetic basis. Since at least 1334, the Ukrainian letter <*г*> has predominantly been Romanized as the Latin <*h*> (Shevelov 1979: 451). However, this practice should not be considered as a scholarly approach to creating a comprehensive Ukrainian transliteration system that provides an equivalent Latin representation of the original text. At that time, the Ukrainian language did not even have an established Cyrillic alphabet. These were rather attempts to write Ukrainian names in a way that native speakers of Polish, German, Hungarian, and other target languages could be read them with sounds more or less approximating the original. The variation in rendering the Ukrainian <*г*> noted by Shevelov can be attributed to the fact that neither <*g*> nor <*h*> provided an accurate representation of its sound.

The letter <г> was inherited from Old (Church) Slavic, which in turn adopted it from the Greek alphabet. The Greek letter gamma <*γ*> is Romanized as <*g*> (ISO 843:1997), and is rendered as <*g*> in languages using Latin script: *geology* (English), *grafolog* (Czech), etc. Therefore, the absence of the element "g" in the transliteration of the Cyrillic <*г*> manifests a lack of historical awareness and sustainability. Furthermore, according to the International Phonetic

Alphabet characteristics, the sound [h] is much closer to the pronunciation of the Ukrainian <*х*> than to the readout of the Ukrainian <*г*> (Vakulenko 2022c).

The letter <*г*> was pronounced as a plosive sound [g] in all territorial variants of Old Slavic until at least the 12th century (Sobolevskij 1907: 43; Shevelov 1979: 451). The *spirantization* of the corresponding phoneme in *Ancient Ukrainian*, *Ancient Czech*, and *Ancient Slovak* occurred between 1157 and 1215 when the voiced pharyngeal fricative sound replaced the previous velar stop consonant [g] (Shevelov 1979: 450-454). The Ukrainian phonetic studies inform that this pharyngeal consonant persists in modern Ukrainian (Žovtobrjux and Kulyk 1965: 122; Bilodid 1969: 176). This sound is denoted in the *International Phonetic Alphabet* (IPA 2018) by the symbol "ʕ" (derived from the modified letter <*G*>). These phonetic facts support the conclusion that the element "g" should be retained in the transliteration of the Ukrainian <*г*> in subsequent historical periods.

Experimental phonetic data indicate that the palatalized variant of the contemporary Ukrainian phoneme /г/ is realized by the velar fricative consonant [ɣʲ] (Tocjka 1981: 83, 90). Given that palatalization, velarization, and pharyngealization are mutually exclusive (Chomsky and Halle 1968: 307), it is unlikely that glottal and pharyngeal allophones [ɦʲ] and [ʕʲ] will appear regularly. A similar situation exists in the Lithuanian language, where the phoneme /ɣ/ is sometimes realized as a glottal [ɦ]. Additionally, the corresponding letter was denoted by "h" in the Lithuanian Cyrillic alphabet used in 1864–1904 (Lithuanian Cyrillic 2023). However, since the palatalized variant of this phoneme is always velar, [ɣ] has an advantage over a [ɦ] (Pakerys 1995). Therefore, considering the softened allophones of the Ukrainian phoneme /г/ also supports using the element "g" in the transliteration of the corresponding Ukrainian letter.

In Ukrainian, a parallel alternation of voiced and voiceless phonemes occurs: /ɣ/ - /ʒ/ - /zʲ/ (*дорога* 'road' – *доріжка* 'path' – *на дорозі* 'on the road') and /х/ - /ʃ/ - /sʲ/ (*рух* 'movement' – *рушати* 'set off' – *у русі* 'in motion'). This indicates the phonemic nature of the corresponding allophone (Vakulenko 2019: 44), further supporting the idea of retaining the graphical element "g" when Romanizing the Ukrainian <*г*>.

The title of the Resolution No. 55 implies that its scope is limited to the Ukrainian alphabet, i. e. the ordered sequence of letters: <*а*>, <*б*>, ..., <*ю*>, <*я*>. As illustrated by the examples above, improper extension of its application to Ukrainian texts, particularly proper names, leads to significant issues in information transfer.

With the 2019 enactment of the Law of Ukraine "On Ensuring the Functioning of the Ukrainian Language as a State Language" (Ukrainian Language 2019), the application of this resolution also becomes legally problematic. This law mandates that the Latinization of Ukrainian proper names ensures their Ukrainian sound. However, the non-isomorphism of the conversion table underlying Resolution No. 55, causes considerable distortions in the pronunciation of Ukrainian words. For example, the representation of the Ukrainian Cyrillic letters <*і*>, <*ї*>, and <*й*> by the same Latin letter <*i*> contradicts the phonetic principle of the Ukrainian language, according to which these letters are always read out distinctly. Even the word *Україна* transcoded using this table is rendered with a Russian spelling and reading: *Ukraina*. This issue extends to other related names, such as *Новоукраїнка*, and *Леся Українка* becoming *Novoukrainka* and *Lesia Ukrainka*, respectively. Additionally, representing iotated sounds with combinations like "ya" (as in *Yakiv*), "yu" (as in *Yukhym*), and <ye> (as in *Yevpatoriia*) aligns with English, not Ukrainian, pronunciation.

On April 1, 2022, the national Ukrainian transliteration standard (DSTU 9112:2021) came into force. This standard implements the decision of the Transliteration Commission from November 16, 2000 (see Vakulenko 2015a: 234–254, 338–358; Vakulenko 2015b: 48),

provides transliteration for various periods of Ukrainian script while adhering to the Slavic written tradition. It ensures full reversibility of transliterated texts and is aligned with ISO 9:1995. The standard includes the complete letter register and bases its transliteration rules on a detailed analysis of the phonetic and phonological characteristics of the speech sounds in New Ukrainian, Old Ukrainian, Ancient Ukrainian, and Old Slavic (Vakulenko 2022a).

The system A (with diacritical marks) is recommended to be used in printed materials (maps, monographs, reference books, etc.), in the content of sites; on vehicles and other objects under the jurisdiction of Ukraine, in international information systems, in organization's sheets, for official documents (agreements, letters, protocols, acts, etc.), in postal stamps, in postal addresses, and on historical monuments.

The system B (without diacritical marks) may be used in telecommunication networks, in international information databases and multilingual text corpora, in address lines of sites and web pages, in code scripts, in the bibliography, in scientometric systems, in identification documents of a physical or a legal person, on signs and road boards (to indicate the names of streets, organizations, institutions, settlements, and other entities), for flight destinations in airports.

For example, the contemporary name of the Ukrainian capital *Київ* is transliterated by DSTU 9112:2021 as *Kyïv* (System A) and *Kyjiv* (System B).

4. Transliteration System for Slavic Linguistics

Similar transliteration regulations for both modern and ancient Slavic languages using the Cyrillic script (Church Slavonic, Old Russian, Russian, Belorussian, Ukrainian / Rusyn, Bulgarian, Macedonian, and Serbian / Montenegrin) are organized under the common authoritative transliteration system for Slavic linguistics (ESLL 2020). This system considers the Slavic written tradition more comprehensively than ISO 9:1995, particularly in its approach to transcoding iotated letters using "j". Another significant advantage is its high degree of reversibility, which allows the original text to be restored in most cases.

However, the name "Old Russian" which stems from Russian empire times, provides an allusion to the Russian language that is not correct both historically and linguistically. This term should be critically and prudently rethought based on historical facts, accounting for its etymology and inner form, and also for the necessity to provide similar periodization for all relevant languages. Another matter of contention arises concerning the emergence of the name "Rusyn" next and actually equated with "Ukrainian" (language). The authoritative linguists from Ukraine and other countries point to the unjustifiability of the Canadian historian Magocsi's attempts to single out the Rusyn language, as well as the political engagement and dubious character of the corresponding term itself (Čučka 2004). In fact, Rusyn is a dialect of Ukrainian, not a separate language.

Phonetic considerations should also be taken into account. The correspondence <*г*> - <*h*> suggested for the Belorussian and Ukrainian / Rusyn languages has, similarly to what has been explained above, little historical and phonetic motivation. Another phonetic and grammatical inconsistency is the transliteration of the softening sign <*ь*> by the acute accent (´) that is conventionally used to emphasize the vowel sounds. Additionally, this choice prevents the distinction between lowercase and uppercase letters.

The register of Church Slavic letters in the transliteration system for Slavic linguistics is not complete, missing characters <*є*>, <*ꙁ*>, <*ћ*>, <*ı*>, <*ї*>, <*ɴ*>, <*ȣ*>, <*ϖ*>, <*ѽ*>, and <*ҫ*> (cf. Istrina 1915: 22; Cejtlin, Večerka, and Blagova 1994: 56-57). Another inconsistency that results in the loss of reversibility is the identical Latinization of letters that denote the same

sound in different historical periods but have different appearances, such as <ѥ> and <є>, <ꙗ> and <я>. The specific form of these letters can provide important information about the period, place, and author of the document. Within the adopted approach, this information would be lost after Latinization.

Another concern is the presence of diacritical marks, which restricts the usability of this system. The widespread communicative bibliographic formats *MARC 21* (Fritz and Fritz 2003) and *UNIMARC* (Hopkinson 2008), as well as their derivatives, imply machine reading of bibliographic records that makes it necessary the use of basic Latin letters.

5. Transliteration in Chinese, Japanese, and Korean

Transliteration is extremely important in Asian languages. In China, for example, there are several systems for transliterating Chinese characters into the Latin alphabet, including the *Wade-Giles* system, the *Yale Romanization* system, and the *Hanyu pinyin* system. The last was adopted as the official Romanization system (Xing and Feng 2016: 100) and is now the most widely used system for transliterating Chinese characters (see also Wipperman 2023; HPRS 2023: 1; ISO 7098:2015). In 1982, the International Organization for Standardization adopted this system as an International standard ISO 7098 (Xing and Feng 2016: 101; ISO 7098:2015). Though it is stated that *Hanyu pinyin* is a phonetic transcription system, rather than a Chinese writing system (Xing and Feng 2016: 105), the phonetic bearings of correspondingly transliterated signs – in particular, <*b*>, <*p*>, <*d*>, <*t*>, <*g*>, <*k*>, <*h*>, <*c*>, <*ch*>, <*sh*>, <*r*>, <*j*>, <*q*>, <*x*>, <*e*>, <*er*> – are, in general, different from those in any other language that manifests a scholarly transliteration approach. Among the shortcomings of this system are the use of diacritical marks in the letter <*ü*> (which is often replaced by the digraph <*yu*>) and in tone notations, and exceedingly specific pronunciation of some letters, such as <*b*>, <*d*>, <*g*>, <*j*>, <*q*>, <*x*> (Xing and Feng 2016: 109-110).

The two main Latinization systems of Japanese are the *Hepburn* and *Kunrei* (Kent, Lancour, and Daily 1977: 155; Mair 2022). The first was invented by an American physician and missionary *James Curtis Hepburn* as a practical transcription into English where some syllables are Latinized in a way to render their actual pronunciation using English grammar. In particular, the syllable *tsu* when preceding the letters <*k*>, <*s*>, <*p*>, <*t*>, is omitted, while the following consonant is doubled (Hepburn 1872: xiv). The use of macron in the letters <ō> and <ū> also restricts the applicability of this system.

The *Kunrei* system is not geared to any foreign language and represents the Japanese phonetic structure (Kent, Lancour, and Daily 1977: 155) that indicates a pure transliteration approach. This system dates back to 1885 when, aiming to replace the *Hepburn* system, the physicist *Aikitsu Tanakadate* developed the *Nihon (Nippon)* system which was approved by the Japanese Cabinet as the official Romanization system on September 21, 1937. Among the Latinization schemes for Japanese, it is the most regular one and the closest to providing scholarly transliteration. Later it was slightly corrected to better reflect the modern Japanese pronunciation, and this modified version is known as the *Kunrei* system. For example, while the *Nihon* scheme renders the characters づ as <du> and ず as <*zu*>, respectively, the *Kunrei* system uses the same notation <*zu*> for both (Mair 2022). So, the advantage of the *Kunrei* system over the *Nihon* one is a better consideration of modern pronunciation, but the price for this gain is the loss of reversibility which is the basic requirement for transliteration. A certain inconsistency between *Nihon* and *Kunrei* is the use of diacritical marks in the letters <*ô*> and <*û*> that restricts their applicability. The *Kunrei* system was standardized by the International Organization for Standardization as ISO 3602 (ISO 3602:1989; see also Romanizing Japanese 2009: 1).

Though it is recommended to use the *Hepburn* scheme, especially in International communication (Romanizing Japanese 2009), the *Kunrei* system also receives today much attention, and in some cases, it even prevails (Mair 2022).

In Korea, there are over twenty different Latinization systems, where the most popular are the *McCune-Reischauer* Latinization system, the *Yale* system, and the *Revised Romanization of Korean* (Song 2005: 63-64). The first is an example of practical transcription, as it renders Korean characters differently depending on their sound. Besides, its application is restricted because it uses diacritics (the breve and the apostrophe). The second is oriented to render the writing and, hence, tends to be used in linguistic publications. The *Revised Romanization* system is characterized as an attempt to abandon diacritics and to move the *McCune–Reischauer* scheme closer to the *Yale* system to achieve more simplicity and consistency (Song 2005: 64). It was adopted as the official Romanization system by the South Korean government in 2000 and is now the most widely used system for transliterating Korean characters.

6. Rendering Exonyms

The transliteration rules apply to texts written in a source language, which may include *endonyms* (native variants of proper names) and *exoticisms*. However, *exonyms* (foreign variants of proper names) belong to the target language and follow its grammar, so they are not subject to transliteration. An exonym emerges when a name is commonly used in another language, adapting to a grammatical form that suits this other language. For example, the authentic geographical name *България* has an English exonym, *Bulgaria*, which is used as an exception to Bulgarian Romanization rules (Transliteratsiyata 2009). In general, different languages give rise to different exonyms for the same endonym. For example, the name *Україна* appears as *Ukraine* in English, French, and German texts, *Ucrania* in Spanish and Italian texts, *Ukrajina* in Czech and Croatian texts, and *Ουκρανία* in Greek texts. It is important that the transliterated form of an endonym may serve as a written basis to create various exonyms.

The diversity of exonyms may result also from the variance of names in the source language. For example, in 949, Constantine VII Porphyrogennitos transcribed the names of the modern capital of Ukraine, *Київ*, into Hellenic script as *Κίαβο* (*Kiavo*), *Κιοάβα* (*Kioava*), *Κίοβα* (*Kiova*) (Shevelov 1979). Various forms of this name and their derivatives have been documented in Old Ukrainian manuscripts, such as *Кіиóвїѧ* (*Kîyóvïä*), *Кієвскіѧ* (*Kîêvskîä*) (Bêrynda 1627: 3). These forms resulted in customary exonyms like *Kijów* (Polish), *Kiew* (German), *Kiev* (English, Spanish, French, Italian), and *Κίεβο* (Greek).

The two United Nations resolutions, IV/20 (1982) "Reduction of exonyms" and V/13 (1987) "Precedence of national official forms of geographical names", aim to promote the wider use of endonyms and their transliterated forms, such as *Beijing* (Chinese) and *Kyjiv* (Ukrainian).

It is important to understand that the Romanized names obtained according to Resolution No. 55 of the Cabinet of Ministers of Ukraine (Kabinet Ministriv Ukrajiny 2010) are neither endonyms nor exonyms. For example, the form "Kyiv" cannot be considered as an endonym because it is specially adapted to the conventions of English writing, disregarding the Ukrainian letter "ї". At the same time, the English exonym remains "Kiev".

7. Latinization for Linguistic Technologies

The use of Latin script has significantly expanded beyond the traditional linguistic task of rendering individual proper names and exoticisms. In machine learning, machine transliteration is regarded as a crucial component of information retrieval (Naji and Allan 2016).

Romanization enhances the similarity of non-Latin scientific texts to those of other languages written in Latin script because many international terms are of Latin or Greek origin and appear similarly in their Latinized forms across different languages. Additionally, a significant amount of terms are eponyms or include proper names that sound similar in different languages. Such similarity is essential for incorporating these languages into multilingual term and information retrieval tasks.

More broadly, many multilingual natural language processing scenarios – such as information search, machine translation, question answering, and entity recognition – strongly require the availability of Romanization for languages that use non-Latin scripts. Latinization has been extensively explored in the context of automatic transliteration tools for both ancient and modern languages. Notable examples include Arabic (Bogacz, Klingmann and Mara 2017), Sanskrit (Cai and Wang 2018), Javanese (Razak et al. 2018; Widiarti and Pulungan 2020), Punjabi (Abbas and Asif 2020), and Ukrainian (Vakulenko 2023).

8. Conclusion

An overview of transcoding systems for converting non-Latin texts into Latin script has been presented. The primary focus has been on Romanization approaches for Slavic and some Asian languages, where the specific characteristics of various Latinization systems have been scrutinized. In general, Romanization systems fall into two classes: the ones based on the transliteration and the ones based on the practical transcription. During transliteration, the main attention is paid to simple correspondence (mutual unambiguity) between original and converted characters, which allows for text reversibility, i.e. the ability to restore the original text after re-transliteration. During practical transcription, the primary focus is on pronouncing words. In this case, the original text cannot always be restored. The transliterated texts are used in various fields, such as printed materials, official documents, postal addresses, telecommunication networks, information databases and multilingual text corpora, multilingual natural language processing scenarios, address lines of websites and web pages, code scripts, bibliographic references, identification documents, signs and road boards, flight destinations in airports, and so on.

References

Abbas, M. R., & Asif, K. H. (2020). Punjabi to ISO 15919 and roman transliteration with phonetic rectification. *ACM transactions on Asian and low-resource language information processing* 19/2, 1–20. dl.acm.org/doi/10.1145/3359991

Bêrynda, P. (1627). *Lêkîkóno slavênorôsskïj, ý yměnŏ tlŏkovánïÊ [Slavic Rosj lexicon and explanation of names]*. Kîyóvïä (in Ukrainian). http://litopys.org.ua/index.html

Bilodid, I. (Ed.). (1969). *Sučasna ukrajinsjka literaturna mova: Fonetyka.* Kyïv.

Bogacz, B., Klingmann, M., & Marat, H. (2017). Automating transliteration of cuneiform from parallel lines with sparse data. *Proceedings of the international conference on document analysis and recognition (ICDAR)*, 615–620. Kyoto.

Cai, Z., & Wang, W. (2018). 2DLDA-based compound distance for similar online handwritten Tibetan transliteration of the Sanskrit character recognition. *Proceedings of the international conference on frontiers in handwriting recognition, ICFHR*, 223–228. Niagara Falls.

Cejtlin, R., Večerka, R., & Blagova, E. (Eds.). (1994). *Staroslavjanskij slovarj (po rukopisjam 10–11 vekov)*. Moscow.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. New York.

Fritz, R., & Fritz, D. (2003). *MARC 21 for everyone: A practical guide*. Chicago.

Čučka, P. (2004). Rusynsjka mova [Rusyn language]. In V. M. Rusanivsjkyj, O. O. Taranenko, M. P. Zjabljuk et al. (eds.), *Ukrajinsjka mova: encyklopedija [Ukrainian language: Encyclopedy]* (p. 565). Ukrajinsjka encyklopedija. Kyïv.

Gešev, V. (2016). Personal communication. Sofia.

Hepburn, J. C. ([2]1872). *A Japanese-English and English-Japanese dictionary*. Shanghai.

Hopkinson, A. ([3]2008). *UNIMARC manual: Bibliographic format*. London.

Istrina, E. (1915). *Rukovodstvo po istorii russkago jazyka*. Saint Petersburg.

Kent, A., Lancour, H., & Daily, J. E. (Eds.). (1977). Oriental literature and bibliography. In *Encyclopedia of library and information science*, vol. XXI. New York.

Kirschenbaum, A., & Wintner, S. (2009). Lightly supervised transliteration for machine translation. *12th conference of the European chapter of the Association for computational linguistics (EACL)*, 433–441. https://aclanthology.org/E09-1050.pdf

Lepsius, C. R. (1863). *Standard alphabet*. London.

Mair, V. (2022). Japanese Romanization: They still haven't decided. *Language log*. languagelog.ldc.upenn.edu/nll/?p=56461

Maslov, J. (2007). *Vvedenie v jazykoznanie*. Moscow.

McArthur, T. (Ed.). (2018). Transliteration. In *Concise Oxford companion to the English language*. Oxford. encyclopedia.com/humanities/encyclopedias-almanacs-transcripts-and-maps/transliteration

Moser, M. (2018). *Istorija ukrajinsjkoji movy*. Kyïv.

Naji, N., & Allan, J. (2016). On cross-script information retrieval. In Stefan Gradmann, Francesca Borri, Carlo Meghini, & Heiko Schuldt (eds.), *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics) 9626* (pp. 796–802). Berlin.

Nimčuk, V., & Purjaeva, N. (2004). *Istorija ukrajinsjkoho pravopysu XVI–XX stolittja: Xrestomatija*. Kyïv.

Pakerys, A. (1995). *Bendrinės lietuvių kalbos fonetika [Phonetics of the common Lithuanian language]*. Vilnius.

Razak, S. M. A., Seman, M. S. A., Mamat, W. A. W. Y. W., & Noor, N. H. N. M. (2018). Transliteration engine for union catalogue of Malay manuscripts in Malaysia: E-JAWI Version 3. *Proceedings of the international conference on information and communication technology for the Muslim world, ICT4M*, 58–63. Kuala Lumpur.

Reformatskij, A. (1960). Transliteracija russkix tekstov latinskimi bukvami. *Voprosy jazykoznanija* 5, 96–103.

Shevelov, G. (1979). *A historical phonology of the Ukrainian language*. Heidelberg.

Sobolevskij, A. (1907). *Lekcii po istorii russkago jazyka*. Moscow.

Song, J. J. (2005). *The Korean language: Structure, use and context*. London.

Superanskaja, A. ([2]2018). *Teoretičeskie osnovy praktičeskoj transkripcii*. Moscow.

Tilkov, D. (Ed.). (1982). Gramatika na săvremennija bălgarski knižoven ezik, vol. I: *Fonetika*. Sofia.

Tocjka, N. (1981). *Sučasna ukrajinsjka literaturna mova: Fonetyka, orfoepija, hrafika, orfohrafija*. Kyïv.

Vakulenko, M. (2015a). *Ukrajinsjka terminologhija: Kompleksnyj linghvistyčnyj analiz*. Ivano-Frankivsjk.

Vakulenko, M. (2015b). Practical transcription and transliteration: Eastern-Slavonic view. *Govor* 32/1, 35–56. hrcak.srce.hr/file/244983

Vakulenko, M. (2019). Ukrainian consonant phones in the IPA context with special reference to /v/ and /gh/. *Linguistica online* 22, 36–61. phil.muni.cz/linguistica/art/vakulenko/vak-001.pdf

Vakulenko, M. (2022a). Izomorfna latynizacija staroukrajinsjkyx i novoukrajinsjkyx tekstiv: Vid istoryčnoji pysemnoji spadščyny do sučasnyx linhvistyčnyx texnologij. *Balcania et Slavia* 2/1, 73–90. edizionicafoscari.unive.it/media/pdf/journals/balcania-et-slavia/2022/1/iss-2-1-2022_pAclewg.pdf

Vakulenko, M. (2023). Normalization of Ukrainian letters, numerals, & measures for natural language processing. *Digital scholarship in the humanities* 38/3, 1307–1321. 10.1093/llc/fqac090

Widiarti, A. R., & Pulungan, R. (2020). A method for solving scriptio continua in Javanese manuscript transliteration. *Heliyon* 6/4, e03827. 10.1016/j.heliyon.2020.e03827

Xing, H., & Feng, X. (2016). The Romanization of Chinese language. *Review of Asian and Pacific studies* 1, 99–111.

Žovtobrjux, M., & Kulyk, B. ([3]1965). *Kurs sučasnoji ukrajinsjkoji literaturnoji movy*. Kyïv.

Sources

ALA-LC Persian. (2012). loc.gov/catdir/cpso/Romanization/persian.df

Church Slavic Romanization table. (2022). ALA-LC Church Slavic. loc.gov/catdir/cpso/Romanization/churchsl.pdf

ČSN ISO 9. (2002). Informace a dokumentace – Transliterace cyrilice do latinky – slovanské a neslovanské jazyky. technicke-normy-csn.cz/csn-iso-9-010185-158060.html#

Cyrillic – ESLL transliteration table. (2024). brill-esll.stippweb.nl/projectdocuments/form/31951/8/

DSTU 9112.2021. (2022). Information and documentation. Cyrillic–Latin transliteration and Latin–Cyrillic re-transliteration of Ukrainian texts. Writing rules (in Ukrainian). zakon.rada.gov.ua/rada/show/v0260774-21#Text

Hanyu Pinyin Romanization System. assets.press.princeton.edu/chapters/i10279.pdf

IPA charts and subcharts in four fonts. (2018). internationalphoneticassociation.org/IPAcharts/inter_chart_2018/IPA_2018.html

ISO 3602.1989. Documentation**.** Romanization of Japanese (kana script). iso.org/standard/9029.html

ISO 7098. 2015. Information and documentation**.** Romanization of Chinese. iso.org/standard/61420.html

ISO 843.1997. Information and documentation. Conversion of Greek characters into Latin characters. iso.org/obp/ui/#iso.std.iso.843.ed-1.v2.en

ISO 9.1995. Information and documentation. Transliteration of Cyrillic characters into Latin characters. Slavic and non-Slavic languages. iso.org/standard/3589.html

Japanese Romanization table. (2022). ALA-LC Japanese. loc.gov/catdir/cpso/Romanization/japanese.pdf

Kabinet Ministriv Ukrajiny. (2010). Pro vporjadkuvannja transliteraciji ukrajinsjkoğo alfavitu latynyceju, January 27, 2010. Resolution No. 55 of the Cabinet of Ministers of Ukraine. zakon.rada.gov.ua/laws/show/55-2010-%D0%BF#Text

Korean Romanization and word division. (2009). ALA-LC Korean. loc.gov/catdir/cpso/Romanization/churchsl.pdf

Law of Ukraine. (2019). Pro zabezpečennja funkcionuvannja ukrajinsjkoji movy jak deržavnoji. zakon.rada.gov.ua/laws/show/2704-19#Text

Lithuanian Cyrillic. (2023). Omninglot. the online encyclopedia of writing systems & languages. omniglot.com/conscripts/lca.htm

NF Z46-001.1995-06-01. (1995). Information and documentation. Transliteration of Cyrillic characters into Latin characters. Slavic and non-Slavic languages. Deutsches Institut für Normung. din.de/en/wdc-beuth.din21.12351958

Recommended system for Romanizing Japanese. (2009). Department of English language / Komaba Organization for Educational Development. The University of Tokyo, Komaba. park.itc.u-tokyo.ac.jp/eigo/UT-Komaba-Romanization-of-Japanese-v1.pdf

Guidance. Romanization systems. Collection of Romanization systems and tables of correspondence approved by Permanent Committee on Geographical Names (PCGN) and the US

Zakon za transliteratsiyata. (2009). https.//www.mrrb.bg/bg/zakon-za-transliteraciyata/