United Nations E/CONF.101/141

Economic and Social Council

Distr.: General
26 June 2012
Original: English

Tenth United Nations Conference on the Standardization of Geographical Names
New York, 31 July – 9 August 2012
Item 13(a) of the provisional agenda*
Writing systems and pronunciation: Romanization

Reversibility of Romanization systems

Submitted by Estonia **

* E/CONF.101/1.
** Prepared by Peeter Päll (Estonia).

Reversibility of romanization systems[1]

Introduction

Romanization systems, i.e. systems for rendering non-Roman elements in Roman script, are created for multiple purposes and therefore it would be impossible to work out universal requirements. In bibliographic records, e.g., it would be important to render all the nuances of the original script, and therefore stringent transliterations, approved by ISO, are used. Romanization systems for names, however, whether personal or geographical names, tend to be more oriented towards practical needs, contending with more or less satisfactory pronunciation of names. In non-Roman scripts where one and the same sound is spelled with different characters, these characters are often romanized with the same Roman equivalent as well.

It seems that there are two important questions that we need to ask when evaluating every romanization system:

1) how well does the system represent phonological features of the language? (phonological appropriateness) and

2) how well does the system represent graphical features of the language? (reversibility criterion).

Most of the United Nations approved romanization systems for geographical names represent quite adequately the phonological structure of the source language, omitting only features that are difficult to reproduce in other languages context (for example, tones in Chinese or Thai names). But it would be difficult to measure exactly the phonological appropriateness for each of the systems, as there needs to be an agreed phonological description for each of the languages as a prerequisite. Besides, several original scripts themselves are deficient in terms of representing the phonological features of a language, and as romanization systems are reflections of the written form, this deficiency is copied into the romanized equivalents, without the "fault" of the system itself. So one can only characterize in general terms the phonological adequacy of romanization systems.

But what about reversibility? As this is an attribute of a writing system, we could think of some measurable ways of assessing reversibility of individual romanization systems.

The following is an attempt to devise one of such methods and to assess the currently approved (and some proposed) systems on the basis of this method.

Measuring reversibility

Reversibility, according to the glossary (GT 2002: 22), is a "characteristic of transliteration that permits a written item to be converted from one script or writing system into another, and

[1] Prepared by Peeter Päll (Estonia).

subsequently to be reconverted back into the source script, the result being identical with the original".

As a general remark one has to note that writing systems are not equal in their number of graphic elements and thus it would in some cases be technically very difficult to devise a romanization system which is reversible. Roman script has 26 basic letters a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z. Thai script has 44 signs for consonants and 57 combinations for vowels, so already the difference in size would explain the difficulty of compiling a reversible system for the romanization of Thai. Not to speak of Chinese script that contains thousands of characters.

Reversibility can be measured in several ways, here one can think of two methods. *Actual reversibility* can be measured if we take a representative sample of romanized names, apply reconversion rules to the names, and then take the percentage of cases where reconverted names match with the original spellings. This would give us an idea on how many names might be distorted in practical terms, taking into account the frequency of characters. Some non-Roman characters might be romanized ambiguously in terms of reversibility but if they are very rare, it would not matter that much.

Mechanical reversibility can be measured if we take the number of graphemes, look at their romanized equivalents and calculate the percentage of graphemes that are romanized unambiguously. This method is surely easier to determine.

Here one should redefine the term *grapheme*. According to the glossary, grapheme is a "graphic representation of a phoneme in a particular language. Examples: *j* is the grapheme for the /dʒ/ phoneme in English and for /ʒ/ in French." (GT 2002: 13) This definition ties a grapheme to a phoneme but in our case, it would be practical to think of graphemes as the smallest meaningful elements of a script, i.e. elements that you would typically find in an ABC-book or a primer for the script. For example, tone marks ◌့ and ◌း or end-of-syllable mark in Burmese would count as graphemes. In order to simplify further the task of counting graphemes in each script, one can take the number of characters listed in the *Report on the Current Status of United Nations Romanization Systems for Geographical Names* under each language/script as a basis.

An ideal reversible romanization would have only unique single-letter romanization equivalents for each of the characters in the source script. Among the UN-approved systems none is such. A frequent second option is to have digraph romanization equivalents that sometimes give a more natural-looking result, e.g. Russian Cyrillic ю and я are romanized *ju* and *ja*, respectively (in the ISO standard they are romanized *û* and *â*). Digraph equivalents are unambiguous if there is no possibility that romanization of two non-Roman characters would yield the same result. E.g. ю → *ju* and я → *ja* are unambiguous if йу → *ju* and йа → *ja* are impossible. In reality, combinations йу and йа do exist in Russian, especially in the transcription of foreign names (Гайана 'Guyana') or within compound names (Казенойам), although they are rare. In such cases the romanization equivalents might be termed as half-

ambiguous. A third case is when a non-Roman character is omitted in romanization. In these cases one can perhaps speak of doubly ambiguous romanization equivalents, because the absence of a non-Roman character cannot be even seen in the romanized form.

In a simplified method, mechanical reversibility is counted by subtracting the number of ambiguous characters from the total of graphemes. If a number of graphemes is, e.g., 27 and the system contains one ambiguous character (-1), one half-ambiguous character (-0.5) and one doubly ambiguous character (-2), then the result would be 27 – (1 + 0.5 + 2) / 27 = 87 %.

Reversibility of individual romanization systems[2]

In the following account, character numbers refer to the tables in the *Report on the Current Status of United Nations Romanization Systems for Geographical Names*.

Amharic. Amharic uses a syllabic script. The total number of graphemes in the UN table is 281, these stem from 34 basic characters. Of these characters, row 1 (ሀ), row 3 (ሐ), row 13 (ኀ) and row 18 (ኸ) are romanized with the same consonant *h* – altogether 30 characters; row 5 (ሠ) and row 7 (ሰ) are romanized with *s* – altogether 14 characters; row 16 (አ) and row 20 (ዐ) are romanized without an initial consonant – altogether 14 characters; row 30 (ጸ) and row 31 (ፀ) are romanized with *ts'* – altogether 14 characters. Reversibility: 281 – (30 + 14 + 14 + 14) / 281 = 74 %.

Arabic. Number of graphemes: 42 (from the second table, *aw* and *ay* are excluded, as these are really combinations in the original script). Of these characters, five are romanized with digraphs: 5 ث → *th*, 8 خ → *kh*, 10 ذ → *dh*, 14 ش → *sh*, 20 غ → *gh*. Four of these can be termed half-ambiguous, as there can be combinations *t+h, k+h, d+h, s+h* in the original script (*gh* is unambiguous, as *g* is never alone). Reversibility: 42 – 2 / 42 = 95 %.

Bulgarian. Number of graphemes: 30. Characters 10 й → *j* and 28 ь → *j* are romanized with the same letter but these cases are not ambiguous, as й occurs after vowels (Михайловград → *Mihajlovgrad*) and ь occurs after consonants (Мальовица → *Maljovica*). Three characters are romanized with digraphs: 26 щ → *št*, 29 ю → *ju*, 30 я → *ja*. It is not known whether combinations шт, йу and йа are possible in Bulgarian, but assuming they are, these characters might be termed half-ambiguous. Reversibility: 30 – 1.5 / 30 = 95 %.

Chinese. The number of all possible graphemes is estimated to be more than 100,000. It is difficult to estimate how many of these characters have a unique romanization equivalent (syllable) but considering the overall number of graphemes, reversibility must be close to zero.

Greek[3]. Number of graphemes: 25 (including the accent mark). Four characters are ambiguous: 7 η → *i*, 9 ι → *i*, and 15 ο → *o*, 24 ω → *o*. Character 14 ν → *n*, but *n* may also

[2] This section is intended to explain how exactly the reversibility rate for each of the systems was calculated. Some readers may wish to skip this and jump to conclusions on page 7.

[3] Here only transcription variant is considered, as this is used for geographical names.

represent γ in combinations γγ, γξ, γχ. Character 2 β → *v*, but *v* may also represent υ in combinations αυ, ευ, ηυ. Character 21 φ → *f*, but *f* may also represent υ in combinations αυ, ευ, ηυ. These three characters might be termed half-ambiguous. Three characters are romanized with digraphs: 8 θ → *th*, 22 χ → *ch*, 23 ψ → *ps*. Two of these are unambiguous, as the letter *h* is never used alone, the third might be half-ambiguous, if in Greek πσ would be possible. Assuming it would, we get reversibility 25 – (4 + 1.5 + 0.5) / 25 = 76 %. If the accent mark is not romanized, the reversibility would fall to 68 % (actual reversibility would then be much lower, as almost every Greek word has an accent mark).

Hebrew. In this case one should distinguish between the two modes of script: the fully pointed (used also for geographical names in maps until recently) and the regular mode (without vowel points). In the first instance the number of graphemes would be 41. Of these, the following characters are ambiguous: consonants 1 א → ', 18 ע → '; 3 ב → *v*, 7 ו → *v*; 10 ט → *t*, 26 ת → *t*; 17 ס → *s*, 25 שׂ → *s*; all 15 vowel characters. Three characters are romanized with digraphs: 13 כ → *kh*, 21 צ → *ts*, 24 שׁ → *sh*. It is not known if these combinations could occur also as a result of romanizing two consecutive Hebrew consonants *k+h, t+s, s+h*, but assuming they could, these characters can be termed as half-ambiguous. Reversibility in the fully pointed mode: 41 – (8 + 15 + 1.5) / 41 = 40 %. However, in everyday communication (the regular mode) points are not used. The number of graphemes in this case would be 22 (the consonants without points, including signs used for consonants/vowels depending on context: ו and י). The ambiguous characters would still be the eight consonants mentioned above + 11 י → *y, i*. The latter is ambiguous, because *i* in Hebrew script could be represented either with a full character or with an "invisible" point. Reversibility in the regular mode: 21 – (9 + 1.5) / 21 = 50 %.

Hindi[4]. The number of graphemes is 60 (13 + 10 + 4 + 33). In the version described in the status report all the characters are unambiguous, as digraph equivalents (2 ख → *kha*, 4 घ → *gha*, etc.) do never seem to occur as a result of romanizing two consecutive Hindi characters joined with *halant* (e.g. क्ह → *k+h*). If this assumption is true, reversibility would be 100 %. However, the use of *halant* may cause ambiguity in some cases. Although the "official version" tells us to romanize कानपुर → *Kānapur*, in practice many non-initial syllables with *a* are often not pronounced (the name in question is usually romanized *Kānpur*) and therefore ambiguity is created: consonant combinations expressed with *halant* would give the same result (so *Kānpur* could theoretically be written कान्पुर). Such cases are quite frequent, therefore the 100 % reversibility is only theoretical.

Khmer. The number of graphemes is 109 (33 + 32 + 12 + 16 + 11 + 5 diacritical signs: ◌៓, ◌៑, ◌៎, ◌៏ and ◌័), including combinations. In the case of Khmer the list of

[4] Romanization systems for the languages of the Indian group have never been implemented, Hindi is taken here only as an example of these systems.

unambiguous romanization equivalents will be shorter than the list of ambiguous ones: the consonants 5 ᠩ → *ngô*, 10 ᠨᠶ → *nhô*, 25 ᠮ → *mô*, 26 ᠶ → *yô*, 27 ᠷ → *rô*, 29 ᠸ → *vô*, 30 ᠰ → *sô*, 31 ᠬ → *hô*, 33 ᠠ → *'â*[5]; the same characters as subscripts; all 12 independent vowel characters; all 16 vocalic nuclei and 9 shortened syllables and vocalic nuclei (excl. 2 ᠺᠠ᠋ → *ă–oă,eă*, 3 ᠺᠡ᠋ᠺ → *ă–oă,eă*). Altogether, 55 graphemes. 2 of the diacritical marks are not expressed in romanization, which means that they are doubly ambiguous. Reversibility: 53 / 109 = 49 %.

Macedonian Cyrillic. The number of graphemes is 32. Of these, four characters have positional ambiguities (4 г → *g*, 6 ѓ → *đ, g*; 13 к → *k*, 24 ќ → *ć, k*) which allow to term them as half-ambiguous. Four characters are romanized with digraphs: 10 ѕ → *dz*, 15 љ → *lj*, 18 њ → *nj*, 30 џ → *dž*. These are half-ambiguous as the combinations may rarely represent also romanization of Cyrillic character combinations (надзор, Билјана, надживее). Reversibility: 32 – 4 / 32 = 88 %.

Mongolian (in China). The number of graphemes is subject to interpretations, the UN table lists 27 romanization equivalents, but some of the Mongolian characters in the table coincide. The romanization of Mongolian names in China is based on their pronunciation, not on spelling. As the relation between the spelling and pronunciation is very complex (e.g. the name of the capital of the Nei Mongol autonomous region is romanized *Hohhot*, but written ᠬᠥᠬᠡᠬᠣᠲᠠ which may be transliterated as *kökeqota*), it would not make sense to calculate the reversibility of the romanization system for Mongolian in China.

Persian. The number of graphemes is 44 (similarly to Arabic, combinations for *ey* and *ow* are excluded from the account). 5 characters are romanized with digraphs (7 چ → *ch*, 9 خ → *kh*, 14 ژ → *zh*, 16 ش → *sh*, 22 غ → *gh*). Of these, 4 are half-ambiguous as they may also represent character combinations *k+h, z+h, s+h, g+h* in Persian. Reversibility: 44 – 2 / 44 = 95 %.[6]

Russian. The number of graphemes is 33. Of these, three are romanized with digraphs (27 щ → *šč*, 32 ю → *ju*, 33 я → *ja*). They are half-ambiguous, as they may also in some cases represent character combinations *š+č, j+u, j+a* in Russian (e.g.Ашагы-Кушчу, Байуба, Батыйаул). Reversibility: 33 – 1.5 / 33 = 95 %.

Serbian. The number of graphemes is 30. Of these, three are romanized with digraphs (14 љ → *lj*, 17 њ → *nj*, 29 џ → *dž*). It is not known if character combinations лј, нј, дж are possible

[5] Note that in conjunction with the *â*-series or *ô*-series of vowels some consonant characters might become unambiguous, e.g. the syllable *kĕ* can be written only ᠺᡇ, and *kĭ* can be written ᠺᡳ.

[6] A source of confusion which is difficult to express numerically is the habit to romanize some names as one word, although in Persian they would be written separately, e.g. خرم آباد → *Khorramābād*.

in Serbian, but if they are, these characters must be termed as half-ambiguous. Reversibility: 30 – 1.5 / 30 = 95 %.

Thai. The number of graphemes is 109 (44 + 57 + 8 other characters and diacritical marks: ๆ, ฯ, ◌่, ◌้, ◌๊, ◌๋, ◌์, ◌็), including combinations. There are only 5 non-ambiguous characters: 7 ง → *ng*, 33 ม → *m*, 35 ร → *r*, 37 ว → *w*; vocalic nuclei 48 กิว → *io*. Several diacritical marks, including tone marks, are not expressed in romanization, which make them doubly ambiguous. In result we get reversibility close to zero.[7]

Tibetan. The number of graphemes is 40 (30 consonants, 4 subscript characters, 5 vowels, 1 other symbol). Like Mongolian, the romanization of Tibetan is based on pronunciation. As the relation between the spelling and pronunciation is very complex (e.g. the geographical name romanized as *Cowargarzê* but written མཆོད་དར་དཀར་རྩེ in Tibetan, may be transliterated as *mtsho-bar-dkar-rtse*), it would not make sense to calculate the reversibility of the romanization system for Tibetan.

Uighur[8]. The number of graphemes is 32. Of these, 14 characters are ambiguous: 1 ا → *a*, 2 ە → *a, e*, 30 ي → *e*; 6 ج → *j*, 12 ژ → *y, j*, 32 ې → *y*; 8 خ → *h*, 29 ھ → *h*; 15 غ → *g*, 19 گ → *g*; 17 ق → *k*, 18 ك → *k*; 25 و → *o*, 27 ۇ → *o*. One character (20 ڭ → *ng*) is romanized with a digraph, so this is half-ambiguous. Reversibility: 32 – 14.5 / 32 = 55 %.[9]

Let us also examine the new proposed systems of romanization.

Belarusian. The number of graphemes is 33. Of these, five characters are romanized with digraphs: 6 е → *je, ie*, 7 ё → *jo, io*, 24 х → *ch*, 32 ю → *ju, iu*, 33 я → *ja, ia*. Considering the very phonetic nature of the Belarusian orthography, combinations like йе, іэ, йо or іо seem impossible, so the only remaining combination could be цг → *ch*. The apostrophe is not romanized in Belarusian but it is rarely used. In many cases, e.g. between a consonant and е, ё, ю, я its use is predictable: п’я → *pja*, п’е → *pje*, only before *i* its use is not predictable: e.g Мар’іна Горка → *Marina Horka*, so it is ambiguous. Reversibility: 33 – 1.5 / 33 = 95 %.

Bulgarian. The number of graphemes is 30. Two characters are ambiguous: 1 а → *a*, 27 ъ → *a*. Other two similarly romanized characters 10 й → *y* and 28 ь → *y*, are not ambiguous (see the current UN system earlier). 7 characters are romanized by di- or trigraphs: 7 ж → *zh*,

[7] Thai script is very much based on pronunciation, omitting characters from romanization that are not pronounced. Representation of phonological features is systematic with the exception that vowel length and syllable tones are not marked.

[8] Here the version meant for general use (incl. geographical names) is considered.

[9] Vowels word-initially begin with a *hamzeh*. As a complication, word-initial forms of vowels may also be used in a compound name: جالالئاۋات → *Jalalawat*.

23 ц → *ts*, 24 ч → *ch*, 25 ш → *sh*, 26 щ → *sht*, 29 ю → *yu*, 30 я → *ya*. These characters are half-ambiguous (cf. also из<u>х</u>од, бра<u>тс</u>тво, <u>сх</u>ема). Reversibility: 30 – (2 + 3.5) / 30 = 85 %.[10]

Persian. The number of graphemes is 44. The following 17 characters are ambiguous: 4 ت → t, 19 ط → t; 5 ث → s, 15 س → s, 17 ص → s; 8 ح → h, 31 ه → h; 11 ذ → z, 13 ز → z, 18 ض → z, 20 ظ → z; 22 غ → q, 24 ق → q; consonant 21 ع → ’ and diacritical mark 12 ء → ’; vowels 4 بٰا → ā, 5 بٰی → ā. Reversibility: 44 – 17 / 44 = 61 %.[11]

Ukrainian. The number of graphemes is 34. Of these, 3 characters are ambiguous: 12 і → *i*, 13 ї → *i, yi*, 14 й → *i, y* (character 11 и → *y* is not mixing with 14, as character 14 occurs only word-initially which is not possible for и). 10 characters are romanized with digraphs: 4 г → *h, gh*, 8 є → *ie, ye*, 9 ж → *zh*, 26 х → *kh*, 27 ц → *ts*, 28 ч → *ch*, 29 ш → *sh*, 30 щ → *shch* (tetragraph), 31 ю → *iu, yu*, 32 я → *ia, ya*. These characters may be termed as half-ambiguous (with the exception of Nos. 4 and 9). Two graphemes, 33 ь and 34 ’, are not romanized at all, these are doubly ambiguous. Reversibility: 34 – (3 + 4 + 4) / 34 = 68 %.

Conclusion

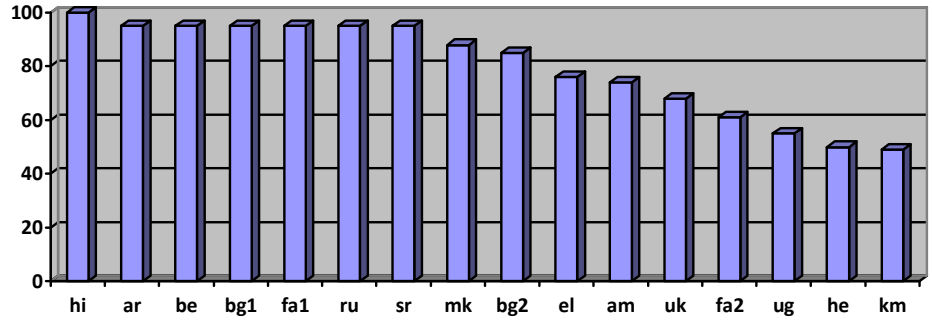
Because of the lack of exact data, many of the above calculations are only approximate, based on assumptions and sometimes subjective assessments, errors not excluded. In several cases, the reversibility is more likely to be higher, as for digraphs the likelyhood of having ambiguous combinations in non-Roman script is perhaps overestimated. Still, the overall conclusion is that there are no fully reversible romanization systems, and the rate of reversibility over 90 % should be considered as very good.

Survey results in a table and a diagram will look as follows:

1. Hindi – 100 %
2. Arabic – 95 %
3. Belarusian (proposed) – 95 %
4. Bulgarian (current) – 95 %
5. Persian (current) – 95 %
6. Russian – 95 %
7. Serbian – 95 %
8. Macedonian Cyrillic – 88 %
9. Bulgarian (proposed) – 85 %
10. Greek – 76 %
11. Amharic – 74 %
12. Ukrainian (proposed) – 68 %
13. Persian (proposed) – 61 %
14. Uighur – 55 %
15. Hebrew – 50 %
16. Khmer – 49 %
17. Chinese – n/a
18. Mongolian (in China) – n/a
19. Thai – n/a
20. Tibetan – n/a

[10] The ending ия is romanized *ia* (София → *Sofia*), which may add further complication, should there also be names ending in иа.

[11] Persian new romanization is very much based on pronunciation which explains why so many consonants occurring mostly in Arabic loanwords are romanized in the same way. There seem to be also cases, where the pronunciation will not have one-to-one relationship with the spelling.



There are several reasons why some languages/scripts prefer non-reversible romanization systems, some of these are explained above. If we compare the data above with how well the systems have been implemented (exactly measuring that would again be difficult), we could even say that some of the most fully implemented systems (Chinese, Thai) are the least reversible which is perhaps something that we find difficult to admit.

While reversibility should remain an essential requirement in approving romanization systems for international use, one has to acknowledge that many of the UN-approved systems are not reversible, although they function properly even without this requirement. Reversibility cannot be mandatory, therefore.

What should perhaps be an unconditional minimum, however, is that romanization systems should systematically represent the phonological structure of the language in question. This would be an important aspect in everyday communication, when there is a need to memorize, write and pronounce the names. Romanized names could also serve as faithful starting points, should there be a need to convert the names into other scripts than Roman (Arabic, Cyrillic, Greek, etc.).

References

GT 2002 = Glossary of Terms for the Standardization of Geographical Names. Edited by Naftali Kadmon. United Nations, New York 2002.

Report on the Current Status of United Nations Romanization Systems for Geographical Names. *http://www.eki.ee/wgrs/.*

Summary

The document addresses the problems of reversibility of romanization systems. It proposes to measure reversibility of romanization systems by counting the graphemes of the source script and calculating the percentage of graphemes romanized unambiguously. Individual romanization systems are assessed, based on this methodology. As a result, there are no systems that are fully reversible, the rate of reversibility over 90 % must be considered as very good. Consequently, reversibility of romanization systems can only be recommended, but not made mandatory.