INTERNATIONAL STANDARD

ISO 7098

Third edition
2015-12-15

Information and documentation — Romanization of Chinese

Information et documentation — Romanisation du chinois

Reference number
ISO 7098:2015(E)

International Organization for Standardization

COPYRIGHT PROTECTED DOCUMENT

ISO copyright office
Ch. de Blandonnet 8 • CP 401
CH-1214 Vernier, Geneva, Switzerland
Tel. +41 22 749 01 11
Fax +41 22 749 09 47
copyright@iso.org
www.iso.org

Contents

Page

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular the different approval criteria needed for the different types of ISO documents should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation on the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the WTO principles in the Technical Barriers to Trade (TBT) see the following URL: Foreword - Supplementary information

The committee responsible for this document is ISO/TC 46, *Information and documentation*.

This third edition cancels and replaces the second edition (ISO 7098:1991), which has been technically revised.

Annexes A, B and C form the integral parts of this International Standard.

Introduction

The first edition of ISO 7098 was published in 1982 after ISO/TC 46 recognized the need for an International Standard specifying the Chinese phonetic alphabet. The second edition was published in 1991.

This third edition is in response to new application needs, for instance to reflect current Chinese romanization practice and new developments in China and the rest of the world.

Information and documentation — Romanization of Chinese

1 Scope

This International Standard explains the principles of the Romanization of Modern Chinese Putonghua (Mandarin Chinese), the official language of the People's Republic of China as defined in the *Directives for the Promotion of Putonghua*, promulgated on 1956-02-06 by the State Council of China. This International Standard can be applied in documentation of bibliographies, catalogues, indices, toponymic lists, etc.

2 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

2.1
character
element of a writing system, whether or not alphabetical, that represents a phoneme, a syllable, a word or even prosodic characteristics of the language, by using graphical symbols (letters, diacritical marks, syllabic signs, punctuation marks, prosodic accents, etc.) or a combination of these signs (a letter having an accent or a diacritical mark)

EXAMPLE *a*, *B*, *ω* or *Γ* are, therefore, characters as well as basic letters.

2.2
alphabets
ordered character set, the order of which has been agreed upon

2.3
alphabetical characters
character set that contains *letters* (2.8)

2.4
alphanumeric characters
character set that contains both *letters* (2.8) and digits

2.5
graphic character
character that has a visual representation and is normally produced by writing, printing or displaying

2.6
ideophonographical character
graphic character (2.6) that represents an object or a concept and is associated with a sound element in a natural language

EXAMPLE Chinese hanzi 鹤(crane), Japanese kanji 戦(war) and Korean hanja 冊(book) are ideophonographical characters.

2.7
Chinese characters
ideophonographical character set for recording the Chinese language

Note 1 to entry: Chinese characters (hanzi) are also used in the writing systems of other languages.

2.8
letter
graphic character (2.6) that, when appearing alone or combined with others, is primarily used to represent a sound element of a spoken language

2.9
word segmentation
process of splitting text into a sequence of word segmentation unit

[SOURCE: ISO 24614-1:2010, 2.25]

3 General principles of conversion of writing systems

3.1 The words in a language, which are written according to a given script (the converted system), sometimes have to be rendered according to a different system (the conversion system), normally used for a different language.

This operation is often performed for historical or geographical texts, cartographical documents and, in particular, for bibliographical work in every case where it is necessary to write words supplied in various alphabets in a manner that allows intercalation with other words in a single alphabet so as to enable a uniform alphabetization to be made in bibliographies, catalogues, indices, toponymic lists, etc. It is indispensable in that it permits the univocal transmission of a written message between two countries using different writing systems or exchanging a message, the writing of which is different from their own. It, thereby, permits transmission by manual as well as mechanical or electronic means.

The two basic methods of conversion of a system of writing are transliteration and transcription.

3.2 Transliteration is the operation which consists of representing the characters of an entirely alphabetical character or alphanumeric character system of writing by the characters of the conversion alphabet.

In principle, this conversion should be made character by character: each character of the converted alphabet is rendered by one character, and one only of the conversion alphabet, to ensure the complete and unambiguous reversibility of the conversion alphabet into the converted alphabet.

When the number of characters used in the conversion system is smaller than the number of characters of the converted system, it is necessary to use digraphs or diacritical marks. In this case, one shall avoid as far as possible arbitrary choices and the use of purely conventional marks and try to maintain a certain phonetic logic in order to give the system a wide acceptance.

However, it shall be accepted that the graphism obtained may not always be correctly pronounced according to the phonetic habits of the language (or of all the languages) which usually use(s) the conversion alphabet. On the other hand, this graphism shall be such that the reader who knows the converted language may mentally restore unequivocally the original graphism and, thus, pronounce it correctly.

3.3 Retransliteration is the operation which consists of converting the characters of a conversion alphabet to those of the converted alphabet.

This operation is the exact opposite of transliteration; it is carried out by applying the rules of a system of transliteration in reverse order so as to reconstitute the transliterated word to its original form.

3.4 Transcription is the operation which consists of representing the characters of a language, whatever the original system of writing, by the phonetic system of letters or signs of the conversion language.

A transcription system is of necessity based on the orthographical conventions of a conversion language and its alphabet. The users of a transcription system shall, therefore, have a knowledge of the conversion language to be able to pronounce the characters correctly. Transcription is not strictly reversible.

Transcription may be used for the conversion of all writing systems. It is the only method that can be used for systems that are not entirely alphabetical and for all ideophonographic writing systems (Chinese, Japanese, etc.).

3.5 Romanization is the conversion of non-Latin writing systems to the Latin alphabet by means of transliteration or transcription.

To carry out Romanization, it is possible to use either transliteration or transcription or a combination of these two methods, according to the nature of the converted system.

3.6 A conversion system proposed for international use may call for compromise and the sacrifice of certain national customs.

It is, therefore, necessary for each national community of users to accept concessions, fully abstaining in every case from imposing as a matter of course solutions that are actually justified only by national practice (for example, regarding pronunciation, orthography, etc.). However, these concessions would obviously not relate to the use that a country makes of its national writing system: when this national system is not converted, the characters constituting it shall be accepted in the form in which they are written in the national language.

When a country uses two systems univocally, converting one into the other to write its own language, the system of transliteration thus implemented shall be taken a priori as a basis for the international standardized system, as far as it is compatible with the other principles mentioned hereafter.

3.7 Where necessary, the conversion systems should specify an equivalent for each character, not only the letters but also the punctuation marks, numbers, etc.

They should similarly take into account the arrangement of the sequence of characters that make up the text, for example, the direction of the script, and specify the way of distinguishing words and of using separation signs and capital letters, following as closely as possible the customs of the language(s) which use the converted writing system.

4 Principles for converting ideophonographic characters

4.1 The structure of ideophonographic characters, where conveyance of meaning is of greater importance than that of pronunciation, entails the existence of a large number of characters (more than 60 000 in the case of Chinese), thus, making sign by sign transliteration impossible and resulting in the need to devise a system of transcription.

Each character shall, therefore, be transcribed by one or more Latin letters standing for the pronunciation or pronunciations of the character in question. This means that the transcriber shall be familiar with the reading or readings of the text to be transcribed.

4.2 In as much as the transcription of ideophonographic characters is merely a matter of phonetic notation in Latin letters of characters of the languages which use them, identical characters will require different transcriptions depending on whether they are found in Chinese, Japanese or Korean texts.

4.3 On the other hand, the same character within the same language shall always be transcribed in the same way regardless of the type of graphic representation utilized (traditional form or simplified form of a Chinese character), except where a single character has more than one pronunciation.

4.4 Reversibility of Romanization systems of ideophonographic characters is impossible due to the following factors:

— the disparity in pronunciation of a given character in two different languages or within a single language;

— the high frequency of homophones within the same language (see Annex C);

— the possible coexistence of several writing systems within a given text.

4.5 In the case of those languages which use, even within the same text, more than one kind of script (for example Kana and Chinese characters in Japanese, Hangul and Chinese characters in Korean), both the transcription of the ideophonographic characters and the conversion of the other types of characters (for example Kana/Hangul) should yield a consistent and homogeneous system of Romanization.

4.6 Although, as a rule, spacing between syllables of Chinese is regular, it is usual to transcribe the different characters (or syllables) forming a single word by linking them together, in order to separate the different words by the space.

The principles and rules for formation of words (orthography) shall be standardized to the language concerned.

4.7 Although there are no capital letters in ideophonographic characters, it is usual when romanizing to capitalize some words, following the national uses.

5 Pinyin

The Scheme of the Chinese Phonetic Alphabet (*Hanyu Pinyin Fang'an* or *Pinyin Fang'an*), which was officially adopted on 1958-02-11 by the National People's Congress of the People's Republic of China, is used to transcribe Chinese. The transcriber writes down the pronunciation of Chinese characters according to their readings in Standard Chinese (*Putonghua*).

6 Syllabic forms

6.1 Each Chinese character generally represents one syllable. One word may consist of one or more syllables.

6.2 A Chinese syllable can be divided into two parts: initial and final.

6.2.1 Initial

— Bilabial: *b p m*;

— Labio-dental: *f*;

— Dorso-prepalatal: *d t n l*;

— Dorso-velar: *g k h*;

— Apico-alveolar: *z c s*;

— Apico-postalveolar: *zh ch sh r*;

— Dorso-palatal: *j q x*;

— Zero initial: nothing before the far left of the final.

6.2.2 Final

— Articulation A: Articulation with *a*, *o*, *e* as medial or main vowel. For example, *a*, *o*, *e*, *ei*, *ao*, *ou*, *an*, *ang*, *en*, *eng*, *ong, er,* and with *i* in *zi*, *ci*, *si*, *zhi*, *chi*, *shi*, *ri* as main vowel.

— Articulation B: Articulation with *u* as medial or main vowel. For example, *u*, *ua*, *uo*, *uai*, *ui*, *uan*, *uang*, *un*, *ueng*.

— Articulation C: Articulation with *i* as medial or main vowel. For example, *i*, *ia*, *ie*, *iao*, *iu*, *ian*, *iang*, *in*, *ing*, *iong*.

— Articulation D: Articulation with *ü* as medial or main vowel. For example, *ü*, *üe*, *üan*, *ün*. Hanyu Pinyin simplifies the spellings of syllables with *ü* by using the *u* form instead in cases where no ambiguity could result.

6.3 Table of syllabic forms

The table of Chinese syllabic forms is given in Annex A. This table covers all syllables of Chinese *Putonghua* except syllable ê and retroflexion syllable.

6.4 Reference dictionaries

Among reference books of modern Chinese are the following dictionaries.

— 中国社会科学院语言研究所词典编辑室编.《现代汉语词典》(第6版). 北京: 商务印书馆, 2012.

Dictionary Compilation Division, Institute of Linguistics, Chinese Academy of Social Sciences, *The Contemporary Chinese Dictionary* (6th Edition). Beijing: The Commercial Press, 2012.

This dictionary gives the transcriptions in Pinyin of more than 69 000 words.

— 《现代汉语词典(汉英双语)》. 北京: 外语教学与研究出版社, 2002.

The Contemporary Chinese Dictionary (*Chinese-English*). Beijing: Foreign Language Teaching and Research Press, 2002.

This dictionary includes equivalent English explanations for Chinese words.

— 德范克主编. 《ABC 汉英大词典》. 夏威夷: 夏威夷大学出版社, 2003.

John DeFrancis. *ABC Chinese-English Comprehensive Dictionary*. Hawai'i: University of Hawai'i Press, 2003.

This dictionary includes 71 344 words, arranged in Pinyin alphabet order. It is easy to check by Pinyin.

— 《新华字典》(第11版). 北京: 商务印书馆,2011.

Xinhua Zidian (11th Edition). Beijing: The Commercial Press, 2011.

This dictionary includes the transcriptions in Pinyin of more than 10 000 characters.

These dictionaries can be complemented by the following list of Chinese characters.

— 中华人民共和国国务院.《通用规范汉字表》. 北京:语文出版社, 2013.

State Council of People's Republic of China. *List of Standard Chinese Characters for General Use.* Beijing: Language and Culture Press, 2013.

This list includes 8 105 commonly-used Chinese characters. In addition, it has a concordance table of simplified characters and non-simplified characters.

7 Tones

7.1 Chinese is a tonal language.

This means that the tone affects meaning. The same sound pronounced in different tones can mean very different concepts.

Each syllable may have one of four tones or may be toneless. The four tones are marked by the following diacritic signs (every diacritic sign has a special hexadecimal code):

— 1st tone (high and level tone) ¯ (hex: 0304);

— 2nd tone (rising tone) ´ (hex: 0301);

— 3rd tone (falling-rising tone) ˇ (hex: 030C);

— 4th tone (falling tone) ` (hex: 0300).

Here is a graphical representation of the four tones.

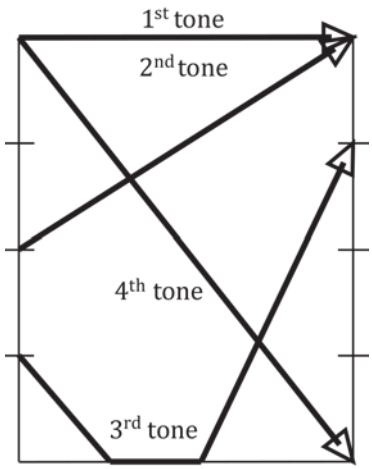
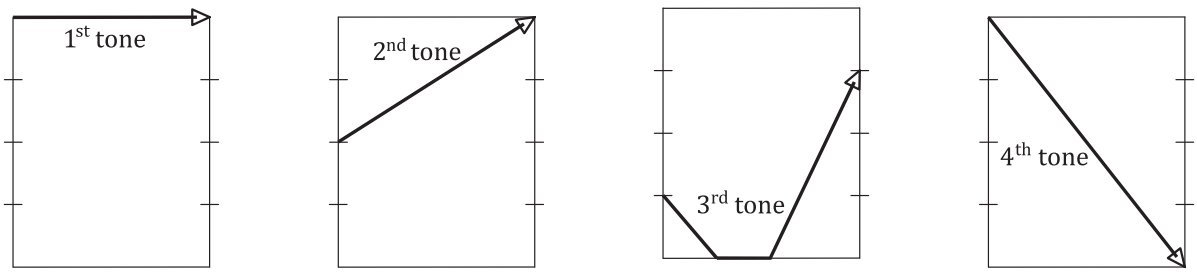


Figure 1 — Graphical representation of the four tones of Putonghua (superposed)



a) 1st tone **b) 2nd tone** **c) 3rd tone** **d) 4th tone**

Figure 2 — Graphical representation of the four tones of Putonghua (separate)

7.2 In the table of Chinese syllabic forms (see Annex A), the syllables do not carry tone marks. But in the text, it is usual to indicate the tone of a syllable by placing the diacritic sign on a vowel.

EXAMPLE ē, é, ě, è.

The diacritic sign for tone is placed on the main vowel in the final part of a syllable.

EXAMPLE /béi/, /què/.

In the final part /éi/ of syllable /béi/, /é/ is the main vowel; and in the final part /uè/ of syllable /què/, /è/ is the main vowel. Therefore, the tone sign is placed on /é/ and /è/, accordingly.

In this case, Latin script letters have to be extended to indicate the Chinese vowels with different tones. The hexadecimal codes of these extended Latin script letters are given in Annex B.

7.3 Neutral tone (atony) is indicated by the lack of a diacritic sign.

7.4 Changes of tone induced by the tone of the next syllable in a word are not shown.

7.5 For practical or technical reason, tones can also be expressed by numbers or letters.

For example, Arabic numbers 1, 2, 3, 4 and 5 usually express respectively 1st tone, 2nd tone, 3rd tone, 4th tone and atony of Chinese.

7.6 Tone marks may be written as a learning tool; however, they can be omitted for convenience.

8 Punctuation

Punctuation marks similar to those existing in the sets of Latin characters are transcribed as their Latin counterparts. Chinese specific punctuation marks are transcribed as follows.

Table 1 — Romanization of Chinese punctuation marks

Chinese mark		Latin mark		Note
。	hex: 3002	.	hex: 002E	full stop
、	hex: 3001	,	hex: 002C	special comma used to set off a short pause in the series
•	hex: 2022	Space	hex: 0020	disconnect mark
……	hex: 2026 2026	…	hex: 2026	horizontal ellipsis

9 Numerals

Numerals written in Chinese characters are transcribed in Pinyin. Numerals written in Arabic or Roman characters are kept as such.

10 Chinese Pinyin Orthography

10.1 Most of commonly used Chinese words are polysyllabic words.

In international documentation and information, it is reasonable to link different Pinyin monosyllables to form a polysyllabic Chinese word (see Annex C).

10.2 Before the Middle Age, the Greeks and Romans always knew what a word was, and they were able to identify words even if the texts were written without spaces between neighbouring words at that time.

Afterwards, the spaces between words were invented in Europe. The use of spaces implies the concept of word, it has become the standard for all modes of writing alphabetical languages to insert spaces between words, and the publishers and librarians in the world apply this common standard.

10.3 In Chinese Pinyin, it is also necessary to use the spaces to separate words, not syllables.

The word segmentation is a very good tradition of world civilization. In the Romanization of Chinese, it is beneficial to respect this good tradition.

10.4 In Chinese Pinyin, monosyllable is ambiguous.

One syllable can represent several Chinese characters. Therefore, Pinyin syllable is ambiguous in representation of Chinese characters. In Chinese Pinyin, the ambiguity index of monosyllables is big. In average, one Chinese syllable has to represent more than 20 Chinese characters for general use.

However, if different Chinese monosyllables are linked to form the polysyllabic Chinese word, the ambiguity index of Pinyin syllable will be reduced. In order to disambiguate Pinyin syllables, it is necessary to link different monosyllables to form a polysyllabic Chinese word.

10.5 The further description of the ambiguity index for Chinese syllables is given in Annex C.

10.6 *Basic Rules for Chinese Pinyin Orthography* (GB/T 16159-2012, Chinese Standard, 2012) contains rules for separating or joining syllables to form a word: rules for spelling common words (nouns, verbs, adjectives, pronouns, etc.), rules for spelling fused phrase expressions, rules for spelling personal names and place names, rules for representing tones, rules for hyphenation at the end of line, etc.

10.7 At present, in Chinese linguistics, there is no clear common definition of a Chinese word yet, so it is difficult to decide the boundary (dividing line) of a common Chinese word sometime, and, of course, it poses difficulty to link the monosyllables to form a common polysyllabic Chinese word.

However, the boundary of a Chinese proper noun is relatively clear. It is not so difficult to link different monosyllables to form a Chinese polysyllabic proper noun (the named entity as personal name, geographic name, language name, ethnic name, tribal name, religion name, etc.), because the boundary of a Chinese polysyllabic named entity is easy to decide according to the standards or regulations of Chinese. In international documentation and information, it is necessary and possible to link different Pinyin monosyllables to form a Chinese polysyllabic named entity in order to avoid ambiguity.

11 Transcription rules for named entities

11.1 Chinese personal names are to be written separately with the surname first, followed by the given name written as one word, with the initial letters of both capitalized.

EXAMPLE 1 Li Hua (李华).

EXAMPLE 2 Wang Jianguo (王建国).

The traditional compound surnames are to be written together.

EXAMPLE 3 Zhuge Kongming (诸葛孔明).

The two-character or multi-character double surnames without traditional permanence are to be written separately with the initial letters of both capitalized.

EXAMPLE 4 Zhang Wang Shufang (张王淑芳).

EXAMPLE 5 Xiang Situ Wenliang (项司徒文良).

EXAMPLE 6 Ouyang Meng Xiang (欧阳孟翔).

The pen names and other aliases are to be treated in the same manner.

EXAMPLE 7 Lu Xun (鲁迅).

EXAMPLE 8 Mao Dun (茅盾).

EXAMPLE 9 Zhang San (张三).

EXAMPLE 10 Wang Pangzi (王胖子).

11.2 A surname, given name or seniority order after the adjuncts “xiao”, “lao”, “da” and “a” is to be written separately and with the initial letter of the last name capitalized.

Adjuncts such as “xiao”, “lao”, “da” and “a” should not be capitalized unless they appear at the beginning of a sentence.

EXAMPLE 1 xiao Liu (小刘, younger Liu).

EXAMPLE 2 lao Qian (老钱, older Qian).

EXAMPLE 3 da Li (大李, older Li).

EXAMPLE 4 a Gui (阿贵, Mr. Gui).

If the character “xiao”, “lao”, “da” and “a” is part of a given name, follow the same practice for given names.

EXAMPLE 5 Wang Xiaojuan (王小娟).

EXAMPLE 6 Zhao Laoshan (赵老山).

EXAMPLE 7 Li Daqin (李大勤).

EXAMPLE 8 Lou Ashu (娄阿鼠).

11.3 Certain Chinese personal names and titles have already fused traditionally and are written as one word with the initial letter capitalized.

EXAMPLE 1 Kongzi (孔子, Master Confucius).

EXAMPLE 2 Baogong (包公, Duke Bao).

EXAMPLE 3 Xishi (西施,acme of beauty, 5th cent. B.C.).

11.4 In Chinese place names, a geographical proper name should be separated from the name of jurisdiction or the geographical feature name.

The multi-character geographical proper name, the name of jurisdiction or the geographical feature name should be separately written together as one word. The first letters of each element should be capitalized.

EXAMPLE 1 Beijing Shi (北京市, Beijing Municipality).

EXAMPLE 2 Hebei Sheng (河北省, Hebei Province).

EXAMPLE 3 Xikou Zhen (溪口镇, Xikou Town).

EXAMPLE 4 Shenzhen Tequ (深圳特区, Shenzhen Special Economic Zone).

EXAMPLE 5 Qujiatun Cun (瞿家屯村, Qujiatun Village).

EXAMPLE 6 Yalu Jiang (鸭绿江, Yalu River).

EXAMPLE 7 Tai Shan (泰山, Tai Shan Mountain).

EXAMPLE 8 Dongting Hu (洞庭湖, Dongting Lake).

11.5 If a geographical proper name or geographical feature name has a monosyllabic adjunct, they should be written together as one word.

EXAMPLE 1 Jingshan Houjie (景山后街, Jingshan Back Street where monosyllabic adjunct “hou” in the geographical feature name “Houjie” is written together with “jie” as one word).

EXAMPLE 2 Chaoyangmennei Nanxiaojie (朝阳门内南小街, South Street inside Chaoyangmen Gate where monosyllabic adjunct “nei” in the geographical proper name “Chaoyangmennei” is written together with “Chaoyangmen” as one word).

EXAMPLE 3 Dongsi Shitiao (东四十条, Dongsi tenth Street where monosyllabic adjunct “si” in the geographical feature name “Dongsi” is written together with “Dong” as one word).

11.6 If a Chinese place name does not contain syllable(s) of geographical feature or jurisdiction, write it together as one word.

EXAMPLE 1 Zhoukoudian (周口店, a historical site).

EXAMPLE 2 Wudaokou (五道口, a street name).

11.7 If a Chinese place name in which the syllable(s) of geographic feature or jurisdiction has become part of the proper name, write it together as one word. If in doubt, write separately as instructed in 10.4.

EXAMPLE 1 Wangcun (王村[镇], Wangcun Town, syllable *cun* has become a part of the proper name *Wangcun*).

EXAMPLE 2 Jingdezhen (景德镇[市], Jingdezhen City where syllable *zhen* has become a part of proper name *Jingdezhen*).

EXAMPLE 3 Heilongjiang (黑龙江[省],Heilongjiang Province where syllable *jiang* has become a part of proper name *Heilongjiang*).

11.8 Chinese transcription of non-Chinese personal names and place names are to be spelled according to their Chinese pronunciation as instructed in 11.1 and 11.3.

Multi-character surnames and given names are to be grouped separately. For reference, the original name or commonly known Roman (Latin) spelling, if known, may be noted in parentheses after the Chinese transcription.

EXAMPLE 1 Makesi or Makesi (Marx) for 马克思.

EXAMPLE 2 Daerwen or Daerwen (Darwin) for 达尔文.

EXAMPLE 3 Niudun or Niudun (Newton) for 牛顿.

EXAMPLE 4 Aiyinsitan or Aiyinsitan (Einstein) for 爱因斯坦.

EXAMPLE 5 Jiechuan Longzhijie or Jiechuan Longzhijie (Akutagawa Ryunosuke) for 芥川龙之介.

EXAMPLE 6 Lichade Nikesong or Lichade Nikesong (Richard Nixon) for 理查德•尼克松.

EXAMPLE 7 Apei Awangjinmei or Apei Awangjinmei (Ngapoi Ngawang Jigme) for 阿沛•阿旺晋美.

EXAMPLE 8 Wulanfu or Wulanfu (Ulanhu) for 乌兰夫.

EXAMPLE 9 Bali or Bali (Paris) for 巴黎.

EXAMPLE 10 Niuyue or Niuyue (New York) for 纽约.

EXAMPLE 11 Dongjing or Dongjing (Tokyo) for 东京.

EXAMPLE 12 Wulumuqi or Wulumuqi (Ürümqi) for 乌鲁木齐.

11.9 Transcribed non-Chinese place names which have already fused as one word are to be spelled according to their Chinese pronunciation.

EXAMPLE 1 Feizhou (非洲, Africa).

EXAMPLE 2 Nanmei (南美, South America).

EXAMPLE 3 Deguo (德国, Deutschland).

EXAMPLE 4 Dongnanya (东南亚, Southeast Asia).

11.10 The detailed spelling rules of personal names and geographical names should be alphabetized according to the following references:

— GB/T 28039-2011 中国人名汉语拼音字母拼写规则. 北京:中国标准出版社,2012.

GB/T 28039-2011 *Chinese Pinyin Spelling Rules for Chinese Personal Names*, Beijing: Standards Press of China, 2012.

— 中国地名汉语拼音字母拼写规则 (汉语地名部分). 《中国语言文字规范和标准选编》. 北京:中国标准出版社,1997, 454-455.

Chinese Pinyin Spelling Rules for Chinese Geographical Place Names (the part of Chinese Geographical Names), in *Selections of Norms and Standards for Language and Script of China*, Beijing: Standards Press of China, 1997, P454-455.

— GB/T 16159-2012 汉语拼音正词法基本规则. 北京:中国标准出版社,2012.

GB/T 16159-2012 *Basic rules for Chinese Pinyin Orthography, Beijing: Standards Press of China*, 2012.

11.11 Each language name is written as one word with the initial letter capitalized if the name includes country, tribe, race or ethnicity name before the language.

EXAMPLE 1 Hanyu 汉语 (Chinese).

EXAMPLE 2 Yingyu 英语 (English).

EXAMPLE 3 Deyu 德语 (German).

EXAMPLE 4 Fayu 法语 (French).

EXAMPLE 5 Xibanyayu 西班牙语 (Spanish).

EXAMPLE 6 Zhongwen 中文 (Chinese).

EXAMPLE 7 Hanwen 韩文 (Korean).

EXAMPLE 8 Kejiahua 客家话 (Hakka).

11.12 Each ethnic name or tribal name is written as one word with the initial letter capitalized.

EXAMPLE 1 Hanzu 汉族(Chinese ethnic group).

EXAMPLE 2 Maolizu 毛利族 (Maori tribe).

EXAMPLE 3 Maonanzu 毛难族 (Maonan ethnic group).

EXAMPLE 4 Weiwuerzu 维吾尔族 (Uyghur ethnic group).

11.13 Each religion name or its follower is written as one word with the initial letter capitalized.

EXAMPLE 1 Fojiao 佛教 (Buddhism).

EXAMPLE 2 Jidujiao 基督教 (Christianity).

EXAMPLE 3 Tianzhujiao 天主教 (Catholicism).

EXAMPLE 4 Yisilanjiao 伊斯兰教 (Islamism).

EXAMPLE 5 Huijiaotu 回教徒 (Muslim).

EXAMPLE 6 Jidutu 基督徒 (Christian).

11.14 In case of ambiguity, the apostrophe will be used to separate the syllables.

EXAMPLE 1 Xi'an vs. Xian for 西安 (Xi'an city).

EXAMPLE 2 Tian'anmen Guangchang vs. Tiananmen Guangchang for 天安门广场 (Tian'anmen Square).

In general documentation, these transcription rules for the named entities can be used manually.

12 Automatic transcription for named entities

In the computer-assisted documentation, there are two approaches to automatic transcription for named entities, namely:

— fully automatic syllable transcription;

— rule-based and semi-automatic word transcription.

12.1 Fully automatic syllable transcription

Fully automatic transcription procedures generating single syllables separated from each other can be used by any application or environment in which the results are regarded appropriate, especially in those that store the Latin transcription together with the original Chinese characters.

Using this approach, the place name 北京市 (Beijing Municipality) will be fully automatically transcribed into syllables /bei/, /jing/ and /shi/. The transcription procedure is as follows:

a) 北京市

b) bei jing shi

The fully automatic approach can be easily realized by computer routines.

12.2 Rule-based and semi-automatic word transcription

In language-related research and industry, the word is a fundamental and necessary concept. In translation, word count is the principal method for calculating the cost of a translation. Word segmentation is a standard function in translation memory systems and computer-assisted translation tools. Word segmentation is performed by term extraction tools, which are sometimes provided in terminology management systems and computer-assisted translation tools. Most content management systems and databases allow for searching by individual words. The content being searched has to be segmented to permit matching with a search word. Furthermore, search functions require knowledge of the boundaries of words. Text-to-speech systems generate speech based on words and, therefore, require word segmentation for lexicon lookup, stress assignment, prosodic pattern assignment, etc. Various natural language processing systems shall segment text into words in order to perform their functions. Lexical resources are often evaluated by size, usually by referring to the number of words.

ISO 24614-1 presents the basic concepts and general principles of word segmentation in natural language processing, and provides language-independent guidelines to enable written texts to be automatically segmented, in a reliable and reproducible manner, into word segmentation units.

ISO 24614-2 specifies rules for delineating word segmentation units for Chinese, Japanese and Korean. Some rules are common to all three languages, though each language also has its own distinct rules for identifying word segmentation units.

Therefore, in Romanization of Chinese, transcribing the named entities in Chinese characters into words in Pinyin will be reasonable. One word may consist of one or more syllables, and the boundary of a Chinese word is not so clear, full automatic word transcription cannot reach the good quality in present technique condition, so the rule-based and semi-automatic word transcription should be used.

The rule-based and semi-automatic word transcription of named entities can benefit from the following resources.

— A set of transcription rules: The rules in Clause 11 cover the general regulations for transcription of named entities. These rules can be used as the resource for semi-automatic transcription of named entities.

— A relevant transcription lexicon: *Chinese Lexicon with Pinyin Proper Nouns*[1)] includes the transcriptions in Pinyin for the most of named entities. This lexicon can be used as another resource for semi-automatic transcription of named entities.

The procedure of rule-based and semi-automatic transcription for named entities is as follows.

— First step: To transcribe a polysyllabic named entity in Chinese character to several monosyllables in Pinyin, list the possible different linking results.

— Second step: Based on the rule of Clause 11, combine the different monosyllables into a polysyllabic named entity, select the most suitable transcription result as the best transcription.

Using this approach, the transcription procedure of place name 北京市 (Beijing Municipality) is as follows:

a) 北京市;

b) bei jing shi;

c) beijing shi;

d) Beijing shi;

e) Beijing Shi.

According to rule 11.4, the geographical proper name /beijing/ should be separated from the jurisdiction name /shi/. The first letters of each element should be capitalized, thus /Beijing Shi/.

If there are any ambiguities or problems in word transcription, the manual post-edit may be done based on the transcription lexicon, the suitable named entity will be found by human-computer interaction (HCI). Therefore, this approach is semi-automatic.

1) 《汉语拼音词汇（专名部分）》，上海：上海辞书出版社，2015 (*The Chinese Lexicon with Pinyin Proper Nouns*，Shanghai: Shanghai Lexicographical Publishing House, 2015).

Annex A
(normative)

Table of Chinese syllable forms

The Chinese syllable forms can be summarized as the following table.

	b	p	m	f	d	t	n	l	g	k	h	z	c	s	zh	ch	sh	r	j	q	x	(Null)
a	ba	pa	ma	fa	da	ta	na	la	ga	ka	ha	za	ca	sa	zha	cha	sha					a
o	bo	po	mo	fo																		o
e			me		de	te	ne	le	ge	ke	he	ze	ce	se	zhe	che	she	re				e
ai	bai	pai	mai		dai	tai	nai	lai	gai	kai	hai	zai	cai	sai	zhai	chai	shai					ai
ei	bei	pei	mei	fei	dei	tei	nei	lei	gei	kei	hei	zei			zhei		shei					ei
ao	bao	pao	mao		dao	tao	nao	lao	gao	kao	hao	zao	cao	sao	zhao	chao	shao	rao				ao
ou		pou	mou	fou	dou	tou	nou	lou	gou	kou	hou	zou	cou	sou	zhou	chou	shou	rou				ou
an	ban	pan	man	fan	dan	tan	nan	lan	gan	kan	han	zan	can	san	zhan	chan	shan	ran				an
ang	bang	pang	mang	fang	dang	tang	nang	lang	gang	kang	hang	zang	cang	sang	zhang	chang	shang	rang				ang
en	ben	pen	men	fen	den		nen		gen	ken	hen	zen	cen	sen	zhen	chen	shen	ren				en
eng	beng	peng	meng	feng	deng	teng	neng	leng	geng	keng	heng	zeng	ceng	seng	zheng	cheng	sheng	reng				eng
ong					dong	tong	nong	long	gong	kong	hong	zong	cong	song	zhong	chong		rong				
er																						er
u	bu	pu	mu	fu	du	tu	nu	lu	gu	ku	hu	zu	cu	su	zhu	chu	shu	ru				wu *
ua									gua	kua	hua				zhua	chua	shua	rua				wa *
uo					duo	tuo	nuo	luo	guo	kuo	huo	zuo	cuo	suo	zhuo	chuo	shuo	ruo				wo *
uai									guai	kuai	huai				zhuai	chuai	shuai					wai *
ui					dui	tui			gui	kui	hui	zui	cui	sui	zhui	chui	shui	rui				wei * [1]
uan					duan	tuan	nuan	luan	guan	kuan	huan	zuan	cuan	suan	zhuan	chuan	shuan	ruan				wan *
uang									guang	kuang	huang				zhuang	chuang	shuang					wang *
un					dun	tun	nun	lun	gun	kun	hun	zun	cun	sun	zhun	chun	shun	run				wen * [2]
ueng																						weng *
i	bi	pi	mi		di	ti	ni	li				zi **	ci **	si **	zhi ++	chi ++	shi ++	ri ++	ji	qi	xi	yi +
ia					dia			lia											jia	qia	xia	ya +
ie	bie	pie	mie		die	tie	nie	lie											jie	qie	xie	ye +
iao	biao	piao	miao		diao	tiao	niao	liao											jiao	qiao	xiao	yao +
iu			miu		diu		niu	liu											jiu	qiu	xiu	you + [3]
ian	bian	pian	mian		dian	tian	nian	lian											jian	qian	xian	yan +
iang							niang	liang											jiang	qiang	xiang	yang +
in	bin	pin	min				nin	lin											jin	qin	xin	yin +
ing	bing	ping	ming		ding	ting	ning	ling											jing	qing	xing	ying +
iong																			jiong	qiong	xiong	yong +
ü							nü	lü											ju #	qu #	xu #	yu #
üe							nüe	lüe											jue #	que #	xue #	yue #
üan																			juan #	quan #	xuan #	yuan #
ün																			jun #	qun #	xun #	yun #

NOTE 1 (Null) Represents a zero initial (i.e. where nothing comes before the final sound in the far left column).

NOTE 2 * Whenever *u* comes at the beginning of a syllable, it is written as *w*. However, *w* shall not appear without an additional vowel, so *u* as a complete syllable is not written as *w* by itself but as *wu*.

NOTE 3 ** The *i* in **zi**, **ci**, **si** is different from most other uses of *i*. It is represented in IPA by ɿ, and it belongs to Articulation A, not to Articulation C..

NOTE 4 ++ The *i* in **zhi**, **chi**, **shi**, **ri** is different from most other uses of *i*. It is represented in IPA by ʅ, and it belongs to Articulation A, not to Articulation C. .

NOTE 5 + Whenever *i* comes at the beginning of a syllable, it is written as *y*. Thus, y, however, shall not appear without an additional vowel, so not *y*, *yn*, *yng* but *yi*, *yin*, *ying*.

NOTE 6 # Hanyu Pinyin simplifies the spellings of syllables with *ü* by using the *u* form instead in cases where no ambiguity could result. This is merely a spelling convention; the *u*'s here are still pronounced as *ü*.

NOTE 7 [1] **wei**: *ui* is actually an abbreviation of *uei*. This is why Hanyu Pinyin uses, for example, *shui*, not *shuei*, and *dui*, not *duei*.

NOTE 8 [2] **wen**: *un* is actually an abbreviation of *uen*.

NOTE 9 [3] **you**: *iu* is actually an abbreviation of *iou*. Thus, since *i* is written as *y* at the beginning of a syllable, the spelling becomes *you* instead of *yu* (which would be misleading).

NOTE 10 Syllable *ê* and retroflexed syllable have been omitted from this table.

NOTE 11 Syllable *er* (it is different from the retroflexed syllable) belongs to Articulation A.

Annex B
(normative)

Table of hexadecimal codes of Chinese vowels with tones

Chinese vowels	1st tone		2nd tone		3rd tone		4th tone	
A	ā	hex: 0101	á	hex: 00E1	ǎ	hex: 01CE	à	hex: 00E0
E	ē	hex: 0113	é	hex: 00E9	ě	hex: 011B	è	hex: 00E8
I	ī	hex: 012B	í	hex: 00ED	ǐ	hex: 01D0	ì	hex: 00EC
O	ō	hex: 014D	ó	hex: 00F3	ǒ	hex: 01D2	ò	hex: 00F2
U	ū	hex: 016B	ú	hex: 00FA	ǔ	hex: 01D4	ù	hex: 00F9
Ü	ǖ	hex::01D6	ǘ	hex: 01D8	ǚ	hex: 01DA	ǜ	hex: 01DC

Chinese vowels	1st tone		2nd tone		3rd tone		4th tone	
A	Ā	hex: 0100	Á	hex: 00C1	Ǎ	hex: 01CD	À	hex: 00C0
E	Ē	hex: 0112	É	hex: 00C9	Ě	hex: 011A	È	hex: 00C8
I	Ī	hex: 012A	Í	hex: 00CD	Ǐ	hex: 01CF	Ì	hex: 00CC
O	Ō	hex: 014C	Ó	hex: 00D3	Ǒ	hex: 01D1	Ò	hex: 00D2
U	Ū	hex: 016A	Ú	hex: 00DA	Ǔ	hex: 01D3	Ǔ	hex: 00D9
Ü	Ǖ	hex: 01D5	Ǘ	hex: 01D7	Ǚ	hex: 01D9	Ǜ	hex: 01DB

Annex C
(normative)

Ambiguity index for Chinese syllables

The number of basic Chinese syllables is only 405. These 405 Chinese syllables can represent the pronunciation of all Chinese characters. *List of Standard Chinese Characters for General use* (2012) includes 8 105 commonly-used Chinese characters. In this case, one Chinese syllable has to represent in average more than 20 Chinese characters (8 105/405 = 20,01) for the general use.

EXAMPLE 1 In *List of Standard Chinese Characters for General Use* (2013), the *Pinyin* syllable /bei/ can represent the following 31 Chinese characters:

北 杯 卑 背 椑 悲 碑 鹎 贝 孛 邶 狈 备 钡 倍 悖 被 琲 棓 辈 惫 焙 蓓 碚 鞁 褙 糒 鞴 鐾 呗 臂

EXAMPLE 2 In *List of Standard Chinese Characters for General Use* (2013), the *Pinyin* syllable /jing/ can represent the following 49 Chinese characters:

京 茎 泾 经 獍 荆 菁 旌 惊 晶 腈 鹊 睛 粳 兢 精 鲸 麠 鼱 井 阱 汫 刭 胫 颈 景 儆 憬 璥 璟 警 劲 径 净 迳 胫 俜 痉 竞 竟 踁 婧 靓 敬 靖 静 境 獍 镜

This means that the Pinyin syllable is ambiguous in representation of Chinese characters.

The ambiguity index is a mathematical description of the degree of ambiguity of Pinyin syllables.

The ambiguity index of the Pinyin syllable (I) equals the number of Chinese linguistic units (characters or words) represented by this Pinyin syllable (N) minus 1. Formula (C.1) is as follows:

$$I = N - 1 \qquad \text{(C.1)}$$

Formula (C.1) means that if the Pinyin syllable can represent N Chinese characters (or words), its ambiguity index (I) equals N – 1.

If one Pinyin syllable can only represent one Chinese character, its ambiguity index is zero. If one Pinyin syllable can represent two Chinese characters, its ambiguity index is 2 – 1 = 1. If one Pinyin syllable can represent three Chinese characters, its ambiguity index is 3 – 1 = 2. ..., etc.

In example 1, the Pinyin syllable /bei/ can represent 31 Chinese characters, its ambiguity index is 31 – 1 = 30; in example 2, the Pinyin syllable /jing/ can represent 49 Chinese characters, its ambiguity index is 49 – 1 = 48.

However, if we combine these two monosyllables /bei/ and /jing/ to form a bi-syllabic word /beijing/, the ambiguity index will be reduced, because /beijing/ can only represent three Chinese bi-syllabic words:

EXAMPLE 3 北京, 背景, 背静.

The ambiguity index of /beijing/ is reduced to 3 – 1 = 2.

If we capitalize the first letter of /beijing/ as /Beijing/, the ambiguity index will be reduced to 1 – 1 = 0. It means that /Beijing/ is a Pinyin word without ambiguity, its sense number is only 1. The sense of /Beijing/ is exactly the name of the capital of China:

EXAMPLE 4 北京.

Therefore, if we link different Pinyin monosyllables to form a polysyllabic Chinese word, the ambiguity index of a Pinyin syllable will be reduced. It is an advantage to link different monosyllables to form one polysyllabic Chinese word.

Bibliography

[1] ISO 3602, *Documentation — Romanization of Japanese (kana script)*

[2] ISO 24614-1, *Language resource management — Word segmentation of written texts — Part 1: Basic concepts and general principles*

[3] ISO 24614-2, *Language resource management — Word segmentation of written texts — Part 2: Word segmentation for Chinese, Japanese and Korean*

[4] ISO/TR 11941,[2)] *Information and documentation — Transliteration of Korean script into Latin characters*

[5] Scheme of Chinese phonetic alphabet, Selections of Norms and Standards for Language and Script of China. Standards Press of China, Beijing, 1997, pp. 441

[6] Directives for the promotion of Putonghua, promulgated by the State Council of China, *Selections of Norms and Standards for Language and Script of China*, Beijing: Standards Press of China, 1997, P439-440

[7] ROMANIZATION A.L.A.-L.C. Chinese, Rules of Application, http://www.loc.gov/catdir/cpso/romanization/chinese.pdf

[8] LIBRARY OF CONGRESS. Pinyin Conversion Project, New Chinese Romanization Guidelines, http://www.loc.gov/catdir/pinyin/romcover.html, 1998

[9] FENG Z. Chinese Romanization and Its Application in HCI, *Human-Computer Interaction, Advanced Interaction Modalities and Techniques, Proceedings of 16th International Conference HCI International, Part II, Lecture Notes in Computer Science (LNCS)*, Springer, 2014, p 406-416

2) Withdrawn.

ICS 01.140.10

Price based on 18 pages