American Journal of Applied Scientific Research
2019; 5(2): 35-40
http://www.sciencepublishinggroup.com/j/ajasr
doi: 10.11648/j.ajasr.20190502.11
ISSN: 2471-9722 (Print); ISSN: 2471-9730 (Online)

A Multilingual Unified Character Encoding Scheme Based on a Normative Character Set

Fu Yong*, Guo Gong

College of Mathematics & System Sciences, Xinjiang University, Urumqi, China

Email address:
1481848326@qq.com (Fu Yong)

*Corresponding author

To cite this article:
Fu Yong, Guo Gong. A Multilingual Unified Character Encoding Scheme Based on a Normative Character Set. *American Journal of Applied Scientific Research.* Vol. 5, No. 2, 2019, pp. 35-40. doi: 10.11648/j.ajasr.20190502.11

Received: April 25, 2019; **Accepted**: June 3, 2019; **Published**: July 2, 2019

Abstract: The scheme is a unified character coding set of Sibe, Manchu, Mongolian and Todo languages, and the letters in the set are arranged in Latin alphabetical order. Compared with the system based on the nominal character encoding, this set does no need free variation selectors in Sibe and Manchu, and greatly reduce the use of the selectors in other languages. Thus, it is more in line with the general user's habits, and improve the information query, search, and transmission.

Keywords: Normative Characters, Character Encoding, Multilingual

1. Introduction

Traditional Mongolian, Todo, Sibe and Manchu are similar in that they all have four different forms of letters: isolate, initial, medial and final forms. Each form of letter has zero to more than one normative form (usually considered to be a character in the alphabet). In addition, some letters have other presentation forms in certain contexts, making them the world's most complex coherent writing system. For example, the character of the vowel U in Manchu has one isolate form, one initial form, three medial forms and four final forms.

The coded character set given in the National Recommendation Standard GB/T 26226-2010 [1] of China Mongolian Information Technology Working Group is based on the traditional Mongolian language. Each character in this coded character set is said to be a representative of several character forms of a letter, and a unique code point is specified. The selected character form is called the nominal character of the corresponding letter, and the character set is named the nominal character set [2, 3].

The nominal characters in Todo, Sibe, Manchu and in Aligali [4-6] (In essence, the Aligali are phonetic symbols created for Sanskrit in Mongolian and Manchu) are the acceptance or supplement in the nominal character set. That is to say, if a nominal character already exists in the nominal character set, the nominal character is recognized as the nominal character of the language. Otherwise a new nominal character is added. In addition to the nominal character, the other character forms of a letter are represented as a deformed manifestation of the nominal character, without additional coding points.

Although the nominal character system reduces the size of coded character set, it brings about many other problems [7], among which the most typical ones are as follows:

(1) Some characters in Sibe and Manchu have the same character form with the traditional Mongolian nominal characters, but they cannot be recognized. For example, the nominal character of the Mongolian vowel letter I is the isolate form character "ᠢ" (encoded as U+1822). In Sibe and Manchu, the vowel letter I also has isolate form "ᡳ", and its initial, medial and final form are also same or similar with Mongolian, but the form is not recognized by the nominal character of Mongolian. Instead, the nominal character is supplemented by the form"ᢎ"in Sibe (encoded as U+185E) and the form"ᢏ"in Manchu (encoded as U+1873).

There are many such cases in GB/T 26226-2010. The reason, is that there are differences in the spelling rules of characters in different languages, and it is difficult to realize various changes of characters in different contexts by using the same nominal character. In other words, there is the conflict problem of the deformation of nominal characters in different languages.

(2) As there are too many presentation forms corresponding to one nominal form, the transformation relationship between

the nominal form and the presentation form becomes very complex, resulting in selection conflict. That is, the presentation form of some letters cannot be simply determined by the position of the letter in the word automatically. To solve this problem, GB/T 26226-2010 had to use manual intervention to select the appropriate character form by manual input control operators.

GB/T 26226-2010 uses seven operators: the operators in the nominal character list has four, namely the free variation selector one (FVS1: U+180B), free variation selector two (FVS2: U+180C), free variation selector three (FVS3: U+180D) and the Mongolian vowel separator (MVS: U+180E). The narrow no-break space (NNBSP: U+202F) and zero width joiner (ZWJ: U+200D) and zero width non-joiner (ZWNJ: U+200C) are also used.

Table 1 is a typical example of using control operators given in GB/T 26226 - 2010. The use of control operators is not only difficult to remember and easy to forget but more importantly, it will cause doubts and confusion to ordinary users, especially at the cost of changing users' operation habits.

Table 1. *Examples of using control operators in table 13 of GB/T 26226—2010 (p 24).*

Presentation form of using ZWNJ	Input character sequences	Presentation form of not using ZWNJ	Input character sequences
			
			
			
			
			
			
			
			
			

A simple example: in editing texts, whether in Chinese or in other languages, people often find that for some purpose, every word or letter in the text is separated by spaces: "中国国家标准" - > "中国国家标准", "Standard" - > "S t a n d a r d". For another example, word is often split and combined in education, as shown in Figure 1.

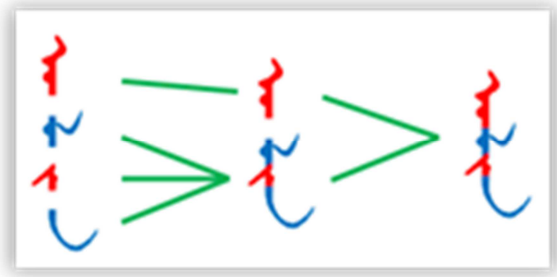

Figure 1. *Split and combination of the Sibe and Manchu word ALIN.*

Figure 2 shows the comparison of the disassembly of four groups of words directly inserted into space: the first group is the expected effect, and the following three groups are the disassembly of words using nominal characters specified in GB/T 26226-2010. Marked 2 is in the Microsoft font Mongolian Baiti, marked 3 is in the YiWenTong (易文通) font SMBT1, and marked 4 is in the MenkSoft (蒙科立) font Menk Qaganti.

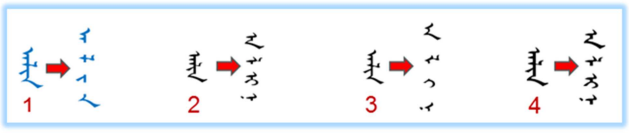

Figure 2. *Comparisons of letters of four fonts inserted into spaces.*

The font marked 1 (using normative characters) has no change in the shape of each character after inserting spaces, while the characters in the subsequent three fonts have significant changes. The reason for this is that these three groups of words are nominal characters. When a space is inserted into a word, the nominal character loses its base of distortion and can only be displayed as the default nominal character. To use these three fonts to achieve the first font effect, in addition to inserting spaces, two control operators must be inserted on both sides of each space.

The use of nominal characters not only causes problems in inserting spaces but also causes confusion and suspicion in common text operations such as finding, retrieving and replacing characters.

(3) The nominal character set specified in GB/T 26226-2010 is first arranged in the order of compilation of some dictionaries in traditional Mongolian, and then the nominal characters in other languages are identified or supplemented successively. For traditional Mongolian it is ordered, but for other languages it is disordered. Although the international standard does not require the order of characters, it does not exclude the ordered sequence of characters. If the characters of traditional Mongolian, Todo, Sibe, Manchu and Aligali are in a unified sequence, and their order is like that of Latin letters, it will be of great benefit to the information processing and application of these languages.

To solve the above problems, this paper proposes a unified character coding scheme based on normative characters for traditional Mongolian, Todo, Sibe and Manchu languages.

2. Definitions of Several Terms

2.1. Definition of GB/T 26226-2010 [1, p 2]

2.1.1. Nominal Form, or Nominal Character

The main form of Mongolian letter. It is applies to the

representation, transmission, exchange, processing, storage, input and output of Mongolian written characters and symbols.

2.1.2. Presentation Form, or Variant Presentation Character

A presentation form is either the nominal form of the letter or the optional form of a sequence of characters in other graphical character areas used in a particular context, depending on the position of the character relative to other characters. Generally, the presentation form is not used to replace the nominal form of graphic characters specified in this coded character set.

2.1.3. Free Variation Selector

A combination character immediately after a specific nominal character, used to distinguish different variants of the same nominal character under the same condition.

2.2. Definition of the Scheme [8, p 2]

2.2.1. Normative Form or Normative Character

Normative forms are the basic expressions of letters in many languages, such as Mongolian, Todo, Sibe and Manchu. It contains the basic forms of the isolates, initials, medials and finals of each letter.

2.2.2. Presentation Form, or Variant Presentation Character

A presentation form is a variant form of normative character. Under certain context conditions, its form is similar to the corresponding normative form and has subtle deformation. The variant presentation characters are not included in the normative character set.

2.3. Description of Definitions

In general, the normative forms can be thought of as all the form of characters in a language alphabet (In fact, the character form in the alphabet is usually incomplete). Normative characters include all kinds of basic forms of letters, which are common, and most of them have obvious differences for the same letter, mainly reflecting the differences of letters in different positions in a word. For example, the letter U in Sibe and Manchu, whose normative forms are shown in Table 2.

Table 2. *Nine Normative Forms of Manchu Letter U.*

Letter	Isolate form	Initial form	Medial form	Final form
U				

As can be seen from Table 2, the differences between the normative character forms of the U letters are still quite obvious. On the contrary, the difference between the presentation form and the normative form is not very obvious. For example, (BU) in spelling, B is changed from a normative form to a presentation form and U is changed from a normative form to a presentation form . Such deformation is subtle. When spaces are inserted between characters, the presentation forms are restored to normative forms, which will not cause people to be surprised and confused.

3. Establishment of Multilingual Unified Character Encoding Scheme Based on Normative Characters

3.1. Identity Based on the Combination of Phonetics, Shapes and Functions

Mongolian, Todo, Sibe and Manchu are all vertically and cohesively written characters, and the shape and function of the characters (or the purpose of using characters) are similar. Therefore, the recognition of these characters in this scheme is based on the phonetics, shapes and functions of the normalized characters in different languages. For example, for vowel A, there are corresponding normative characters in several languages. The normative characters of these languages can be integrated together, as shown in Table 3.

Table 3. *The Normative characters of Mongolian, Todo, Sibe and Manchu vowel A.*

A	Isolate form		Final form		Initial form	Medial form		
Coding sequence	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Mongolian Todo Sibe Manchu								

It can be seen from Table 3 that the first isolate form (Coding sequence is 1), the Final form (3, 4), the initial form (5) and the first Medial form (6) of the letter A are identical with each other in four languages; the second isolate form (2) and the second medial form (7) of the letter A are identical with the traditional Mongolian and Todo language, while the third medial form (8) belongs only to the traditional Mongolian.

3.2. Integration of Multilingual Normative Character Sets

Table 3 shows a sequence of characters in four languages, which is the multilingual normative character set of the letter A. In the same way, other letters can form their own multilingual normative character sets. The normative character set of all letters is integrated in a certain order to form a whole multilingual normative character set. The normative character set of each letter is a subset of the whole multilingual normative character set.

3.3. Some Need to Be Clarified

The normative character set formed by the above way will produce many normative characters with similar or even identical shapes. Taking Sibe and Manchu as examples, the

form in the first medial form of vowel letter A, the form in the second medial form of vowel letter E and the form in the medial form of consonant letter N used for syllable tail are all of the same shape, they all appear as form ᠨ. Some peers may think this is not allowed. In fact, although these characters have the same shape, they have different pronunciations, are used for different purposes and have different character names, and they are distributed in different subsets of the normative character set.

Section 6.3.2 of the International Standards ISO/IEC 10646-2014 describes graphic characters as follows: “The same graphic character shall not be allocated to more than one code point. There are graphic characters with similar shapes in the coded character set; they are used for different purposes and have different character names.” [10, p 11]

And an example of this is given in section 13 of the International Standards: “Graphic characters specified in this International Standard are uniquely identified by their names. This does not imply that the graphic symbols by which they are commonly imaged are always different. Examples of graphic characters with similar graphic symbols are LATIN CAPITAL LETTER A, GREEK CAPITAL LETTER ALPHA and CYRILLIC CAPITAL LETTER A.” [10, p 20] They are all A and have the same shape, but they are alphabetics of different languages, and are used for different purposes and have different character names. Therefore, in the international standard, although the shapes are the same, they have their own codes.

There is no problem that three characters ᠨ in the normative character set are used for different purposes and have different character names as members of the normative character set and can be coded separately. Similarly, this is true for two ᡴ and two ᡷ, and other similar character forms in the normative character set.

Similarly, many different forms of a letter can become a member of the coded character set as long as they are used for different purposes and have different character names. For example, the upper and lower case letters of Latin letters have different character encodings. For example, Arabic in the U+FB50 to U+FBFF code area contains different forms of deformed characters such as isolate, initial, medial and final forms of letters. Another example is that the Chinese characters "国" and "國" are simplified and traditional forms of the same word. Although they have the same meaning, they have different shapes and different purposes of use. International standards ISO/IEC 10646-2014 specify codes U+56FD and U+570B respectively [10, p 575]. Therefore, there is no reason to exclude characters other than nominal characters in the normative character set of Mongolian, Todo, Sibe and Manchu languages.

3.4. About Aligali

Aligali characters are the phonetic symbols of Sanskrit used in Mongolian and Manchu [5, 6]. Sanskrit has abundant pronunciation. In most cases, traditional Mongolian and Manchu can use existing characters to annotate Sanskrit, but some pronunciations can not be annotated, so traditional Mongolian and Manchu have created some new characters for these pronunciations. The Aligali haracters in GB/T 26226-2010 are alse the nominal characters of each Aligali letter.

In addition to having specific pronunciations, Aligali letters also have isolate, initial, medial and final forms. Therefore, the Aligali subset of the normative character set can also be established in the manner described above.

4. Ordering and Encoding of Characters in the Normative Character Set

4.1. Ordering of Characters in a Subset of the Normative Character Set

Order of the characters in a subset of the normative character set is the order of the normative characters of a letter. This order reflects the sorting relationships between different characters of the same letter when comparing strings. This problem has been very mature conclusion: different character forms of a letter can be arranged in any of the following ways: (1) isolate form, final form, initial form, and medial form. (2) isolate form, initial form, final form and medial form. The characters in Table 3 are arranged in way of (1). In this way, for identiti reasions, differtent languages have no priority.

4.2. Order of Subsets in the Normative Character Set

The order of subsets reflects the order of letters. At present, there are three main alphabetical orders of phonetic written language in China: (1) the order of Latin letters (a, b, c,...), (2) the order of phonetic symbols of letters in Chinese Pinyin (b, p, m, f,...) and (3) the traditional order of letters (e.g. the order of the Manchu alphabet appearing in the Twelve Manchu Syllabic Headings[1] [11] (满文十二字头)). Many scholars seem to prefer to adopt the traditional order out of their desire to inherit traditional culture.

The Twelve Syllabic Headings of Mongolian or Manchu divides the syllables into twelve divisions. Although most of letters have a order, there are still many letters without order, so the traditional sorting methods are different in order. For example, there are 17 kinds of order in Mongolian Dictionaries [12, 13], and similar situation in Manchu Dictionaries. Moreover, due to the different languages, the number of vowels and consonants is different, and their traditional ordering habits are also different. Therefore, in order to form a unified multilingual order, the traditional order is obviously not desirable.

At first, many Chinese dictionaries adopted the order of Chinese Pinyin or other Chinese Pinyin symbols. However, with the application of computer technology and the trend of internationalization, new editions of Chinese dictionaries such as Xinhua Dictionary (新华词典) and Chinese Dictionary (汉语字典) also adopt Latin alphabetical order.

Therefore, it should be an ideal choice to specify the order of subsets of the normative character set in Latin alphabetical order. As for some of the traditional Mongolian, Todo, Sibe,

1 The twelve divisions of the traditional Mongolian or Manchu syllabary.

Manchu and Aligali letters which can not correspond to 26 letters in English are expressed in other extended Latin letters and need to be solved by the International Standard Organizations (ISO/IEC) and relevant experts through consultation. Once it's solved, the order of the subset of the normative character set can be arranged in the inherent order of the Latin alphabet and the extended Latin alphabet.

4.3. Normative Character Set Encoding

Character encoding based on the normative character set can be done in two ways:

(1) Encoding in the order given by the characters of the normative character set. This requires relevant organizations and institutions to redefine or expand the code area of these languages. If the code area is redefined, the advantage of the normative character set can be brought into full play. If the code area is only expanded from the original base, the characters in the original code area U+1800 to U+18AF need to be redesigned and the characters sequence needs to be adjusted.

(2) The original characters in the code area from U+1800 to U+18AF are not adjusted, and the new code area only specifies the code points of other characters supplemented by the normative character set. In this way, the efficiency of coded character set will be reduced and the advantages in query, retrieval and sorting will be losed.

5. Conclusion

The number of coded characters increases with the adoption of the normative character set of this scheme. In GB/T 26226-2010, there are 123 nominal characters (excluding numbers and other symbols) in Mongolian, Todo, Sibe, Manchu and Aligali, while the number of characters with can be found from ref. [9] and GB/T 26226-2010 and can be added to normative character set, is about 370 to 380. that is to say, the number of normative characters is only 250 more than that of nominal characters. This seems to be a disadvantage in terms of quantity. However, the increase in this number is not very obvious, but the benefits are very significant.

First of all, the normative characters make the word processing process more in line with the user's general usage habits. They are consistent with operations of Latin and Chinese operations in inserting spaces, searching and replacing, and no longer confused and doubtful to users.

Secondly, the use of normative characters makes the logic of character morphing simple. Therefore, neither Sibe nor Manchu need to be controlled and changed by free variant selectors. Traditional Mongolian and other languages will greatly reduce the use of control operators.

Third, if the coding can be carried out according to the normative character sequence proposed in the previous way (1), the efficiency of searching, comparing and sorting text information will be greatly improved, and the cost of text information processing will be reduced.

Fourth, reducing the use of controllers will also reduce the storage space of text data, which is conducive to improving the efficiency of storage and transmission.

Fifth, the use of normative characters makes it easier to change a normative form of a letter in a text to other presentation form in a computer font automatically. Taking the SMBT1. TTF[2] font based on nominal characters as examples, among the font, there are 41 nominal characters and 250 presentation characters in Sibe and Manchu. Similarly, among the XM_YaBai. TTF[3] font based on normative characters, there are 132 normative characters and 75 presentation characters in Sibe and Manchu. Obviously, in SMBT1. TTF font, each nominal character needs to convert about 6.10 representation characters, while in XM_YaBai. TTF font, only about 0.56 representation characters need to be converted for each normative character. From this point of view, the rendering process of nominal characters is ten times more complex than that of normative characters.

For systems using nominal characters, the key on the keyboard usually bind one of the nominal characters. Complex character deformations will be handed over to fonts for completion. This is relatively simple for input method programming. However, there are too many presentation characters corresponding to a nominal character, and there are still many times when users need to use different controllers to select the appropriate presentation characters. Users must remember many free variable selectors and other operators, as well as how they are used. This increases the user's burden, but also changes the user's usual operating habits.

On the contrary, in a system using normative characters, most keys on the keyboard bind only one letter, while a few keys bind more than one letter, and each letter has more than one normative character, which of course makes the programming process of input method much more complicated. However, the job is for input method to control a key by software to select the right one among multiple normative characters. Therefore, for the general user, there is no worry about memorizing operators, and the text processing process is more simple and convenient. This is to give the difficulty to the programmer and the convenience to the user. The input method of Sibe and Manchu developed by Urumqi SoBetter Digital Technology Co., Ltd. can control the different changes of normative characters in different positions of words by means of "automatic selection" when input Sibe and Manchu texts. It does not need operators at all, which greatly facilitates use of users. This proves the superiority of using normative characters.

In a word, the scheme of multilingual normative character set is more conducive to the progress and development of information technology in these languages.

References

[1] Information technologi —Mongolian presentation forms character set and use rules of controlling characters. GB/T 26226-2010 [S]. China Standard Press. 2011.1.

2 Developed by Weifang Beida Jade Bird Huaguang Imagesetter Co., Ltd.

3 Developed by Urumqi SoBetter Digital Technology Co., Ltd.

[2] Aodenbala. Research and Implementation of Convertion for Mongolian Presentation Character to Basic Letter [D/OL]. Inner Mongolia University. June, 2010. http://www.docin.com/p-1398081892.html&isPay=1.

[3] WANG Zhen, LIU Huidan, WU Jian. Design and Implement of Mongolian Shaping Model under the New Standard [J]. Journal of Chinese Information Processing, Vol, 27, No. 1. Jan. 2013. pp 108-114.

[4] Quejingzhabu and his colleagues. Conversion Rules of Traditional Mongolian Nominal Characters to Presentation Characters (Augmented Set). [R/OL]. 2017.3. pp 35-40. https://www.docin.com/p-2025002658.html&isPay=1.

[5] NIE Hong-yin. Sanskrit and Mandarin Pronunciation in Tongwen Yuntong. MANCHU STUDIES [J]. Manchu Institute of Heilongjiang Province, China 2014 (1), pp 5-10.

[6] Tongwen Yuntong·Aligali·Read the Mantra Method [O]. Photocopy of the Qing Qianlong inner palace, Shanghai Hanfeng building, Eighteen years of the Republic of China.

[7] Fu Yong. Existing Problems and Solutions for Current Sibe and Manchu Coding [J]. China Electronics Standardization Institute. Information Technology & Standardization. 1-2, 2015. pp 58-61.

[8] Fu Yong, Guo Gong, Feng Hui. A Study on the Division and Spelling of Xibe Letters Based on Phoneme [J]. Journal of Jilin Normal University (Humanities & Social Science Edition). Vol 45, No. 5, Sep. 2017. pp 35-44.

[9] Urumqi SoBetter Digital Technology Co., Ltd. Information technology - Sibe and Manchu coded character set: Q/SBT002-2014 [S]. [2014-9-1]. pp 2.

[10] Information technology — Universal Coded Character Set (UCS). ISO/IEC 10646 [S], 2014.9. pp 11, 20, 575.

[11] Jerry Norman. A Comprehensive Manchu-English Dictionary. Harvard University Press, 2013. pp 222: Juwan juwe uju bithe, 386: uju (6).

[12] Mi Ji-sheng. The alphabetical order of the Mongolian alphabet and the Mongolian dictionary. Journal of Inner mongolia Normal University (philosophical & social science) [J]. Inner mongolia Normal University, China. 1982.9 (1), pp 17-19.

[13] E·Baoyinwuliji. Problems and Solutions in the Development of “The Sequencing Rules of Traditional Mongolian Words” [J]. China Electronics Standardization Institute. Information Technology & Standardization. 1-2, 2015. pp 40-42, 46.