1 Q − 2

Sequencial Talks on The Multilingual Text Processing(2): Codeset Designs for Mongolian and Its Related Scripts

Tomoko Inagawa Kataoka*, Kazutomo Uezono†, Hidejiro Daikokuya†,
Dawa Yidemucau†, Toshio Oya†, Yutaka Kataoka* and Hiroyoshi Ohara†

* Media Network Center, Waseda University † School of Science and Engineering, Waseda University

1. Introduction

Internatinalization (18N) is a simultaneous mixing of any number of scripts, which may be varied from one another in writing direction or internal structure of a syllabic. Practically, mixing of the scripts written horizontally and those written vertically should be realized without any inconsistencies. Such truely I18Nized world of scripts was once realized in the age of the Mongolian Empire. Not only their first Uighur and its reformed classic Mongolian scripts, written vertically, but also Paspa, Soyombo, Tibetan, Devanagari, and Chinese Hantsu, of course, were mixed especially for writing religeous texts or monuments.

Also today, Mongolian language is transcribed by the descendants of the Classic Mongolian Script, *phonemic* and *position-dependent* one like Perso-Arabic.

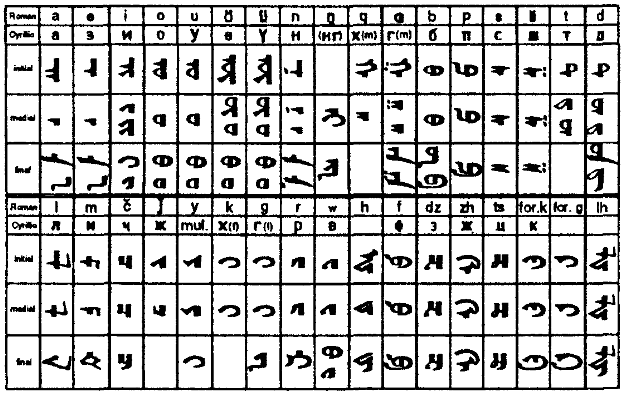

Roman
Cyrillic
initial
medial
final
a e i o u ö ü n ŋ q ɣ b p s š t d
l m č ǰ y k g r w h f dz zh ts for.k for.g lh

Figure 1: Scripts for Mongolia

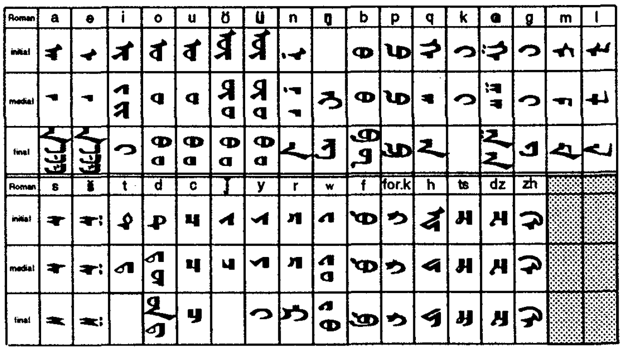

Roman
initial
medial
final
a e i o u ö ü n ŋ b p q k ɣ g m l
s š t d c ǰ y r w f for.k h ts dz zh

Figure 2: Scripts for Inner Mongolia

These two involve glyph-sound ambiguities but differ in the number of the letters and in some writing conventions. Todo, which means "exact" in Mongolian, script keeps a one-letter-to-one-sound principle. It is evident that the principled ways to define both of a character and a glyph are required to manipulate such scripts.

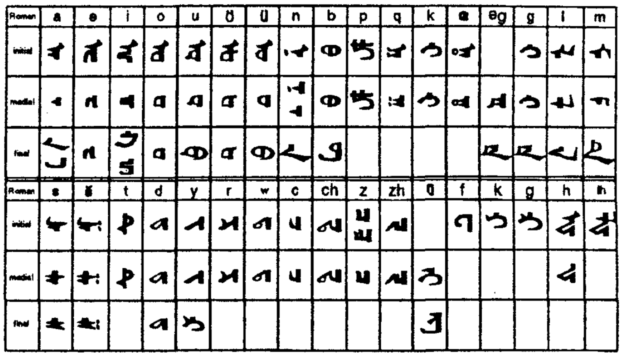

Roman
initial
medial
final
a e i o u ö ü n b p q k ɣ ŋg g l m
s š t d y r w c ch z zh ŋ f k g h ħ

Figure 3: Todo (Oirat) Scripts

2. Codeset Designs

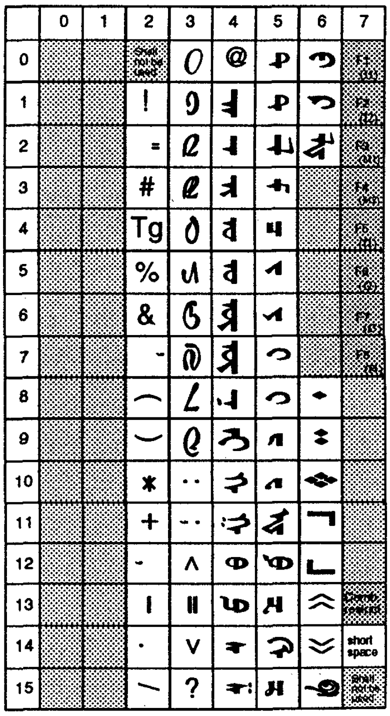

0 1 2 3 4 5 6 7
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Shall not be used
!
#
Tg
%
&
*
+
?
@
F1 F2 F3 F4 F5 F6 F7 F8
Comb. restrict
short space
Shall not be used

Figure 4: Codeset for Mongolia

The proposed codeset is a Character definable one for text manipulation, quite different from GB

8045:87 for Mongolian scripts. The character 'a' and 'e', for example, are given distinct codepoints, although they happen to have the same glyph in the medial position. The first byte specifies a character, while the second Style variation selectors can identify the forms for each character.

3. The Cyrillic Scripts

The *Cyrillic script* has been used in Mongolia since 1946. It does not have enough number of characters to transcribe Khalkha-Mongolian. Two characters are supplied for it, and there are two other Cyrillic systems used for dialects of Mongolian language: Kalmuck (Oirat) and Buriat. The numbers of characters, and even some sound values of characters differ among the three systems: thus, it is impossible to unify these system, as in the cases of three Mongolian scripts.

Mongolia: 35 characters

А а a	Б б b	В в b,(w)	Г г g, ɢ	Д д d	Е е je, jö	Ё ё jo	Ж ж dž	З з dz	И и i
Й й ĭ	К к (k)	Л л l	М м m	Н н n, ŋ	О о o	Ө ө ö	П п p	Р р r	С с s
Т т t	У у u	Ү ү ü	Ф ф (f)	Х х x	Ц ц ts	Ч ч tš	Ш ш š	Щ щ (štš)	Ъ ъ
Ы ы i	Ь ь	Э э e	Ю ю ju, jü	Я я ja					

Buriat: 36 characters

А а a	Б б b	В в (v)	Г г g	Д д d	Е е je	Ё ё jo	Ж ж ž	З з z	И и i
Й й ĭ	К к (k)	Л л l	М м m	Н н n	О о o	Ө ө ö	П п p	Р р r	С с s
Т т t	У у u	Ү ү ü	Ф ф (f)	Х х x	Һ һ h	Ц ц (ts)	Ч ч tš	Ш ш š	Щ щ (štš)
Ъ ъ	Ы ы ī	Ь ь	Э э e	Ю ю ju, jü	Я я ja				

Kalmuck (Oirat): 39 characters

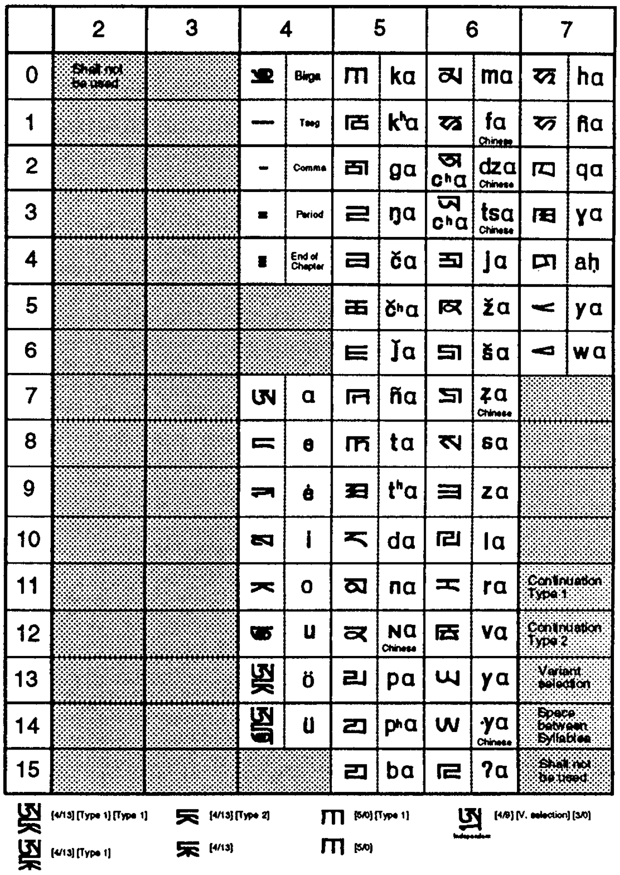
А а a	Ә ә ä	Б б b	В в w	Г г g	Һ һ γ	Д д d	Е е e	Ё ё (jo)	Ж ж (ž)
Җ җ dž	З з z	И и i	Й й j	К к k	Л л l	М м m	Н н n	Ң ң ŋ	О о o
Ө ө ö	П п p	Р р r	С с s	Т т t	У у u	Ү ү ü	Ф ф (f)	Х х x	Ц ц ts
Ч ч tš	Ш ш š	Щ щ (štš)	Ъ ъ	Ы ы ī	Ь ь	Э э e	Ю ю ju	Я я ja	

characters enclosed by '(' and ')' are for foreign words.

Figure 5: Three Cyrillic Script Systems

4. The Paspa Script

The Paspa script was designed by a Tibetan, which was intended to be the international phonemic script. It has enough number of characters to represent the sounds and syllable structures even of the other languages than Mongolian. The Tibetan script shows the syllable boundaries by Tseg's, while Paspa by the connections of characters. Thus, it has the codepoint to define which variant to select.



	2	3	4	5	6	7
0	Shall not be used		Birga	ka	ma	ha
1			Tseg	kʰa	fa Chinese	ɦa
2			Comma	ga	cʰa dza Chinese	qa
3			Period	ŋa	cʰa tsa Chinese	ya
4			End of Chapter	ča	ja	aḥ
5				čʰa	ža	ya
6				ǰa	ša	wa
7			a	ña	ƶa Chinese	
8			e	ta	sa	
9			ė	tʰa	za	
10			i	da	la	
11			o	na	ra	Continuation Type 1
12			u	Na Chinese	va	Continuation Type 2
13			ö	pa	ya	Variant selection
14			ü	pʰa	·ya Chinese	Space between Syllables
15				ba	ʔa	Shall not be used

[4/13] [Type 1] [Type 1]
[4/13] [Type 2]
[5/0] [Type 1]
[4/9] [V. selection] [3/0]
Independent
[4/13] [Type 1]
[4/13]
[5/0]

Figure 6: Codeset for Paspa

5. Summary and Further Remarks

All Mongolian related scripts were researched in the historical order and were shown to be encoded into character codesets. The Waseda I18N & multilingual System (System 1) reinforced by the researches on the historical scripts is quite valid for manipulating the enormous databases of the museums and libraries, ICAI applications, etc.

References

[1] Kataoka, T. I. et al., 1996. Definition of the Mongolian Character Codesets Enabling Multilingual Text Manipulation, IPSJ SIG Notes, Computer and Humanities, 96-ch-29, pp. 61-66.