

6. Mongolian

Mongolian is an Altaic language spoken in Mongolia, China and Russian Federation. Today about 8 million people in the world speak Mongolian. Most of that are approximately 2.7 million in Mongolia and 3.38 million in Inner Mongolia in China [37, 38]. Khalkha or Halha dialects of Mongolian is the national language of Mongolia [37].

6.1. Writing System

Mongolian has shown a varied history of writing. Early Mongolian was written in a script adapted from Old Sogdo-Uighur script in early thirteenth century. As Mongolian derived from Uighur script which originated from Aramaic script (of Semitic origin), it was initially written in a right-to-left direction. However, later the system was rotated by 90 degrees counter clockwise and currently the script is written in top down direction from left-top-right, a unique feature of this script [39]. However, over next two centuries Chinese, Arabic and Tibetan scripts were also used to write the language. In 1930's Cyrillic script was increasingly used, and on 1st Jan, 1946 it was formally adopted by the Mongolian government. It is still being used to write Mongolian language in Mongolia. There have been attempts to restore Traditional Mongolian script, e.g. by the government in 1994 [39]. Though currently both scripts are used in Mongolia, Cyrillic use is more widespread. Traditional Mongolian script is mostly being used in Inner Mongolia in China to write Mongolian. The Cyrillic and Traditional Mongolian scripts do not have clear correspondence. The current work is focused on the collation of Mongolian language using Cyrillic script.

6.1.1. Character Set

Cyrillic script has been derived from Greek script and has been traditionally used to write Slavic languages, including Russian. The Mongolian character set is slightly modified Cyrillic alphabet by adding two vowels such as (ө, ү). Each character has a capital and a small letter (shown in figure below) and it uses the numerals 0, 1, 2, 3..., 9.

**А Б В Г Д Е Ё Ж З И Й К Л М Н О П
Р С Т У У Ф Х Ц Ч Ш Щ Ъ Ы Ь Э Ю Я**

Capital Letters

**а б в г д е ё ж з и й к л м н о ө
р с т у ү ф х ц ч ш щ ъ ы ь э ю я**

Small Letters

Figure 6.1. Mongolian Character Set in Cyrillic Script

6.1.2. Script Details

Cyrillic is written from left to right. Words are separated by spaces and letters are cased as capital and small letters.

6.1.2.1. Case

In Mongolian, all the characters have upper and lower case variants, with the exception of Palochka [4]. For example Cyrillic letter Zhe has the upper case form Ж (U+0416) and the lower case form ж (U+0436). The characters with upper case are sorted before the ones with lower case.

6.2. Collation

Mongolian uses the conventional ordering of Cyrillic script and the three levels of collation associated with it. Numerals and letters are sorted at primary level, diacritics are sorted at secondary level, and case is handled at the tertiary level.

6.2.1. Text Processing

6.2.1.1. Normalization

Mongolian has few characters which can be encoded in multiple ways using Unicode. This is possible due to separate encoding of some marks in addition to encoding of composite forms. Some examples are shown in Table 6.1. below.

Table 6.1. Examples of Normalization in Mongolian using Cyrillic Script

Decomposed Form	Unicode of Decomposed Form	Equivalent Composed Form	Unicode of Composed Form
Е̂	0415 0308	Ё̂	0401
е̂	0435 0308	ё̂	0451
И̂	0418 0306	Й̂	0419

6.2.2. Unicode Collation Elements

Following collation elements give correct ordering of Mongolian script. The results are based on the order of words given in [40].

Table 6.2. Collation Elements for Mongolian Language Using Cyrillic Script

Glyph	Unicode	Collation Elements	Unicode Name
← Numerals →			
0	0030	00A0 0020 0002	DIGIT ZERO
1	0031	00A1 0020 0002	DIGIT ONE
2	0032	00A2 0020 0002	DIGIT TWO
3	0033	00A3 0020 0002	DIGIT THREE
4	0034	00A4 0020 0002	DIGIT FOUR
5	0035	00A5 0020 0002	DIGIT FIVE
6	0036	00A6 0020 0002	DIGIT SIX
7	0037	00A7 0020 0002	DIGIT SEVEN
8	0038	00A8 0020 0002	DIGIT EIGHT
9	0039	00A9 0020 0002	DIGIT NINE
←Consonants and Vowels→			
А	0410	0E29 0020 0002	CYRILLIC CAPITAL LETTER A
а	0430	0E29 0020 0008	CYRILLIC SMALL LETTER A
Б	0411	0E2A 0020 0002	CYRILLIC CAPITAL LETTER BE
б	0431	0E2A 0020 0008	CYRILLIC SMALL LETTER BE
В	0412	0E2B 0020 0002	CYRILLIC CAPITAL LETTER VE
в	0432	0E2B 0020 0008	CYRILLIC SMALL LETTER VE
Г	0413	0E2C 0020 0002	CYRILLIC CAPITAL LETTER GHE
г	0433	0E2C 0020 0008	CYRILLIC SMALL LETTER GHE
Д	0414	0E2D 0020 0002	CYRILLIC CAPITAL LETTER DE
д	0434	0E2D 0020 0008	CYRILLIC SMALL LETTER DE
Е	0415	0E2E 0020 0002	CYRILLIC CAPITAL LETTER IE
е	0435	0E2E 0020 0008	CYRILLIC SMALL LETTER IE
Ё	0401	0E2F 0020 0002	CYRILLIC CAPITAL LETTER IO
Е''	0415 0308	0E2F 0020 0002	CYRILLIC CAPITAL LETTER IO
ё	0451	0E2F 0020 0008	CYRILLIC SMALL LETTER IO
е''	0435 0308	0E2F 0020 0008	CYRILLIC SMALL LETTER IO
Ж	0416	0E30 0020 0002	CYRILLIC CAPITAL LETTER ZHE
ж	0436	0E30 0020 0008	CYRILLIC SMALL LETTER ZHE
З	0417	0E31 0020 0002	CYRILLIC CAPITAL LETTER ZE
з	0437	0E31 0020 0008	CYRILLIC SMALL LETTER ZE
И	0418	0E32 0020 0002	CYRILLIC CAPITAL LETTER I
и	0438	0E32 0020 0008	CYRILLIC SMALL LETTER I
Й	0419	0E33 0020 0002	CYRILLIC CAPITAL LETTER SHORT I
И'ч	0418 0306	0E33 0020 0002	CYRILLIC CAPITAL LETTER SHORT I
й	0439	0E33 0020 0008	CYRILLIC SMALL LETTER SHORT I
й'	0438 0306	0E33 0020 0008	CYRILLIC SMALL LETTER SHORT I
К	041A	0E34 0020 0002	CYRILLIC CAPITAL LETTER KA
к	043A	0E34 0020 0008	CYRILLIC SMALL LETTER KA
Л	041B	0E35 0020 0002	CYRILLIC CAPITAL LETTER EL
л	043B	0E35 0020 0008	CYRILLIC SMALL LETTER EL
М	041C	0E36 0020 0002	CYRILLIC CAPITAL LETTER EM
м	043C	0E36 0020 0008	CYRILLIC SMALL LETTER EM

Н	041D	0E37 0020 0002	CYRILLIC CAPITAL LETTER EN
н	043D	0E37 0020 0008	CYRILLIC SMALL LETTER EN
О	041E	0E38 0020 0002	CYRILLIC CAPITAL LETTER O
о	043E	0E38 0020 0008	CYRILLIC SMALL LETTER O
Ө	04E8	0E39 0020 0002	CYRILLIC CAPITAL LETTER BARRED O
ө	04E9	0E39 0020 0008	CYRILLIC SMALL LETTER BARRED O
П	041F	0E3A 0020 0002	CYRILLIC CAPITAL LETTER PE
п	043F	0E3A 0020 0008	CYRILLIC SMALL LETTER PE
Р	0420	0E3B 0020 0002	CYRILLIC CAPITAL LETTER ER
р	0440	0E3B 0020 0008	CYRILLIC SMALL LETTER ER
С	0421	0E3C 0020 0002	CYRILLIC CAPITAL LETTER ES
с	0441	0E3C 0020 0008	CYRILLIC SMALL LETTER ES
Т	0422	0E3E 0020 0002	CYRILLIC CAPITAL LETTER TE
т	0442	0E3E 0020 0008	CYRILLIC SMALL LETTER TE
У	0423	1350 0020 0002	CYRILLIC CAPITAL LETTER U
у	0443	1350 0020 0008	CYRILLIC SMALL LETTER U
Ү	04AE	1353 0020 0002	CYRILLIC CAPITAL LETTER STRAIGHT U
ү	04AF	1353 0020 0008	CYRILLIC SMALL LETTER STRAIGHT U
Ф	0424	1356 0020 0002	CYRILLIC CAPITAL LETTER EF
ф	0444	1356 0020 0008	CYRILLIC SMALL LETTER EF
Х	0425	1359 0020 0002	CYRILLIC CAPITAL LETTER HA
х	0445	1359 0020 0008	CYRILLIC SMALL LETTER HA
Ц	0426	135C 0020 0002	CYRILLIC CAPITAL LETTER TSE
ц	0446	135C 0020 0008	CYRILLIC SMALL LETTER TSE
Ч	0427	135F 0020 0002	CYRILLIC CAPITAL LETTER CHE
ч	0447	135F 0020 0008	CYRILLIC SMALL LETTER CHE
Ш	0428	1360 0020 0002	CYRILLIC CAPITAL LETTER SHA
ш	0448	1360 0020 0008	CYRILLIC SMALL LETTER SHA
Щ	0429	1363 0020 0002	CYRILLIC CAPITAL LETTER SHCHA
щ	0449	1363 0020 0008	CYRILLIC SMALL LETTER SHCHA
Ъ	042A	1366 0020 0002	CYRILLIC CAPITAL LETTER HARD SIGN
ъ	044A	1366 0020 0008	CYRILLIC SMALL LETTER HARD SIGN
Ы	042B	1369 0020 0002	CYRILLIC CAPITAL LETTER YERU
ы	044B	1369 0020 0008	CYRILLIC SMALL LETTER YERU
Ь	042C	136C 0020 0002	CYRILLIC CAPITAL LETTER SOFT SIGN
ь	044C	136C 0020 0008	CYRILLIC SMALL LETTER SOFT SIGN
Э	042D	136F 0020 0002	CYRILLIC CAPITAL LETTER E
э	044D	136F 0020 0008	CYRILLIC SMALL LETTER E
Ю	042E	1370 0020 0002	CYRILLIC CAPITAL LETTER YU
ю	044E	1370 0020 0008	CYRILLIC SMALL LETTER YU
Я	042F	1373 0020 0002	CYRILLIC CAPITAL LETTER YA
я	044F	1373 0020 0008	CYRILLIC SMALL LETTER YA

6.2.3. Results

Table 6.3. shows output obtained by sorting a sample input using the collation elements given in Table 6.2.

Table 6.3. Input and Corresponding Sorted Output for Mongolian

Input		Output	
Яион	Маяг	Аагим	ёслогч
ганц	Каир	аагим	ёст
бүч	Ёстой	Аагтай	Ёстой
Бэл	ааЖуу	Аагтай	Ёстой
бэл	ИГ	аагтай	ёстой
Аагим	аагтай	аагтай	ёстой
ганха	ИД	ааЖуу	ИГ
Тойн	егее	аажуу	ИД
год	Аагтай	Бэл	Кабин
ёстой	Цунх	бэл	Каир
дзздзх	Яри	бүч	Маяг
аагтай	Метр	ганха	Метр
дзэр	Тожгор	ганц	Тожгор
ёстой	аагим	год	Тойн
Кабин	аажуу	дзздзх	Цунх
Аагтай	еГЕЕ	дзэр	Цуца
ёслогч	Ёстой	еГЕЕ	Яион
ёслогч	Цуца	егее	Яри
ёст		ёслогч	

6.3. Conclusion

Mongolian is a simple case of collation. It is very similar to that of other Latin and Cyrillic based languages. Letters are sorted at primary level, marks are sorted at secondary level and case is sorted at tertiary level. There are no exceptions to this process. Some pre-processing is required before collation can be done to normalize multiple encodings.