Unicode Technical Report #2

Sinhala
Tibetan
Mongolian

Revision	
Authors	The Mongolian proposal was written by Joe Becker. The Sinhala proposal was written by Andy Daniels. The Tibetan proposal is a revision of the one in Unicode 1.0, Vol 1.
Date	1992
This Version	http://www.unicode.org/unicode/reports/tr2.html
Previous Version	
Latest Version	http://www.unicode.org/unicode/reports/tr2.html

Technical Reports contain material that has been approved by the Unicode Consortium for publication, but that is not necessarily considered part of the Unicode Standard. Often, technical reports are superseded by later standardization, or the informative material that they contain is incorporated into explanatory chapters in subsequent editions of The Unicode Standard. Sometimes minor updates to the Unicode Standard itself are published as Technical Reports. Some Technical Reports are not available for downloading.

Technical Report #2 - Sinhala, Tibetan, and Mongolian

ASCII plain text version without charts

This Technical Report is comprised of three preliminary draft proposals that the Unicode Technical Committee wishes to present for initial public review and commentary. These are: Sinhala, Mongolian, and Tibetan.

Status of this document

This document has been considered and approved by the Unicode Technical Committee for publication as a Technical Report. At the current time, the specifications in this technical report are provided as information and guidance to implementers of the Unicode Standard, but do not form part of the standard itself. The Unicode Technical Committee may decide to incorporate all or part of the material of this technical report into a future version of the Unicode Standard, either as informative or as normative specification. Please mail corrigenda and other comments to errata@unicode.org.

Document

Introduction

These proposals represent the committee's strong technical recommendation for the basic approach to these scripts, but some degree of further feedback is needed to complete them. At this time, the committee is interested in suggestions for improvement. The committee is recommending that these eventually be assigned to particular blocks of codepoints as follows:
Sinhala U+0D80 U+0DFF
Tibetan U+1000 U+105F
Mongolian U+1060 U+109F

Specific open issues for each of these are addressed in the respective draft block introductions.

Sinhala

Sinhala U+0D80 - U+0DFF

Sinhala (or Sinhalese) is used to write Sinhala, the majority language of Sri Lanka (formerly Ceylon). It is also used to write Pali and Sanskrit. The script is a descendant of Brahmi and resembles the scripts of South India in form and structure. Sinhala differs from the other Indo-Aryan languages in that is has a series of prenasalized stops which are distinguished from the combination of a nasal followed by a stop. In addition, Sinhala has signs for both a short and long low front vowel, similar to the one in the English word 'hat.' Because of these extra letters, the encoding for Sinhala does not follow the pattern established for the other Indic scripts (e.g., Devanagari), but does use the same ordering patterns, making use of phonetic order, matra-reordering, and use of the virama to indicate conjunct consonant clusters.

Neither the Sinhala numerals nor the Kundaliya are in general use today, having been replaced by Arabic digits and Western-style punctuation. They are included in Unicode for scholarly use.

Encoding Structure. The code assignments for Sinhala depart from the general Indic pattern established by Unicode Devanagari because of the abovementioned additional letters. The general layout of the block, however, retains the basic Indic layout.

U+0D80 to U+0D81 Vowel modifiers
U+0D82 to U+0D93 Independent Vowels
U+0D94 to U+0DBB Consonants
U+0DBC to U+0DCC Vowel Signs
U+0DCD Virama
U+0DCE to U+0DCF Unassigned
U+0DD0 to U+0DDA Numerals
U+0DDB Punctuation (kundaliya)
U+0DDC to U+0DFF Unassigned

Issues:

Colloquial Sinhala has no prenasalized palatal stop. Minimal pairs abound for the other prenasalized stops, (e.g. /aN.da/ "sound" vs. /a.n.da/ "egg," /siNdu/ "horse" vs. /sindu/ "ballads," /saNgara/ "ointment" vs. /sanggara/ "battle"), but examples have not been found for SINHALA LETTER NJA (U+0D9F). This is perhaps simply a ligature for NYA+JA, and so should not be given a code point of its own. If this is dropped, it is recommended that the code point remain unassigned.

A Sinhala avagraha has been attested. If required, it should be assigned to U+0DCF.

There is a standard extant for Sinhala described in A Standard Code for Information Interchange in Sinhalese by V.K. Samaranayake and S.T. Nandasara (ISO-IEC JTC1/SCL/WG2 N 673, Oct. 1990). The coding proposed in it was found to be an inadequate basis for a modern, computer-based interchange code, though it is adequate to handle the capabilities of a Sinhala typewriter for representing contemporary colloquial Sinhala. In addition, the document is ambiguous as to coding order -- presumably, given the graphic decomposition in the code set, the text stream is to be coded in visual, not phonetic order. An additional problem is that there is no provision to handle exceptional cases. For example, RA+U/UU is, as is common in the Brahmi family, written in a non-standard way. The vowel is represented by the matra that is normally used for the low front vowels, i.e, E/EE. RA+E/EE, however, is written with ligatures in which the matra is attached to the stem of the RA. In light of the above, the technical recommendatio

DRAFT 03 Nov 1992

SINHALA DRAFT CHARACTER NAMES

@ Vowel modifiers
0D80 SINHALA SIGN ANUSVARA
0D81 SINHALA SIGN VISARGA

@ Independent Vowels
0D82 SINHALA LETTER A
0D83 SINHALA LETTER AA
0D84 SINHALA LETTER E
0D85 SINHALA LETTER EE
0D86 SINHALA LETTER I
0D87 SINHALA LETTER II
0D88 SINHALA LETTER U
0D89 SINHALA LETTER UU
0D8A SINHALA LETTER RI
= SINHALA LETTER VOCALIC R
0D8B SINHALA LETTER RII
= SINHALA LETTER VOCALIC RR
0D8C SINHALA LETTER LU
= SINHALA LETTER VOCALIC L
0D8D SINHALA LETTER LUU

= SINHALA LETTER VOCALIC LL
0D8E SINHALA LETTER EY
0D8F SINHALA LETTER EEY
0D90 SINHALA LETTER AI
0D91 SINHALA LETTER O
0D92 SINHALA LETTER OO
0D93 SINHALA LETTER AU

@ Consonants
0D94 SINHALA LETTER KA
0D95 SINHALA LETTER KHA
0D96 SINHALA LETTER GA
0D97 SINHALA LETTER GHA
0D98 SINHALA LETTER NGA
0D99 SINHALA LETTER NGGA
prenasalized GA
0D9A SINHALA LETTER CA
0D9B SINHALA LETTER CHA
0D9C SINHALA LETTER JA
0D9D SINHALA LETTER JHA
0D9E SINHALA LETTER NYA
0D9F SINHALA LETTER NJA
prenasalized JA?
0DA0 SINHALA LETTER TTA
0DA1 SINHALA LETTER TTHA
0DA2 SINHALA LETTER DDA
0DA3 SINHALA LETTER DDHA
0DA4 SINHALA LETTER NNA
0DA5 SINHALA LETTER NDDA
prenasalized DDA
0DA6 SINHALA LETTER TA
0DA7 SINHALA LETTER THA
0DA8 SINHALA LETTER DA
0DA9 SINHALA LETTER DHA
0DAA SINHALA LETTER NA
0DAB SINHALA LETTER NDA
prenasalized DA
0DAC SINHALA LETTER PA
0DAD SINHALA LETTER PHA
0DAE SINHALA LETTER BA
0DAF SINHALA LETTER BHA
0DB0 SINHALA LETTER MA
0DB1 SINHALA LETTER MBA
prenasalized BA
0DB2 SINHALA LETTER YA
0DB3 SINHALA LETTER RA
0DB4 SINHALA LETTER LA
0DB5 SINHALA LETTER VA
0DB6 SINHALA LETTER SHA
0DB7 SINHALA LETTER SSA
0DB8 SINHALA LETTER SA

0DB9 SINHALA LETTER HA
0DBA SINHALA LETTER LLA
0DBB SINHALA LETTER FA

@ Vowel Signs
0DBC SINHALA VOWEL SIGN AA
0DBD SINHALA VOWEL SIGN E
0DBE SINHALA VOWEL SIGN EE
0DBF SINHALA VOWEL SIGN I
0DC0 SINHALA VOWEL SIGN II
0DC1 SINHALA VOWEL SIGN U
0DC2 SINHALA VOWEL SIGN UU
0DC3 SINHALA VOWEL SIGN RI
= SINHALA VOWEL SIGN VOCALIC R
0DC4 SINHALA VOWEL SIGN RII
= SINHALA VOWEL SIGN VOCALIC RR
0DC5 SINHALA VOWEL SIGN LU
= SINHALA VOWEL SIGN VOCALIC L
0DC6 SINHALA VOWEL SIGN LUU
= SINHALA VOWEL SIGN VOCALIC LL
0DC7 SINHALA VOWEL SIGN EY
0DC8 SINHALA VOWEL SIGN EEY
0DC9 SINHALA VOWEL SIGN AI
0DCA SINHALA VOWEL SIGN O
0DCB SINHALA VOWEL SIGN OO
0DCC SINHALA VOWEL SIGN AU

@ Virama
0DCD SINHALA SIGN VIRAMA

0DCE
0DCF

@ Numerals
0DD0 SINHALA DIGIT ONE
0DD1 SINHALA DIGIT TWO
0DD2 SINHALA DIGIT THREE
0DD3 SINHALA DIGIT FOUR
0DD4 SINHALA DIGIT FIVE
0DD5 SINHALA DIGIT SIX
0DD6 SINHALA DIGIT SEVEN
0DD7 SINHALA DIGIT EIGHT
0DD8 SINHALA DIGIT NINE
0DD9 SINHALA NUMBER TEN
0DDA SINHALA NUMBER ONE HUNDRED

@ Punctuation
0DDB SINHALA KUNDALIYA

0DDC
0DDD

0DDE
0DDF

Tibetan

Tibetan(U+1000 - U+105F)

The Tibetan script is used for writing the Tibetan language in Tibet proper and for Tibetan and related languages spoken elsewhere in the Himalayan region, including Bhutan, India, and Nepal. The Tibetan script is a member of the Indic family of scripts descended from Brahmi. The original Brahmi letter shapes can still be clearly discerned in Tibetan, but Tibetan removes the Brahmi voiced aspirates and adds letters for Tibetan sounds not found in Brahmi.

General Principles of the Script: As in all Indic scripts, each Tibetan letter is a consonant containing an inherent vowel sound. Tibetan letters each also contain an inherent tone related to the voicing or non-voicing of the original Brahmi letters, but this is not marked in the script. The inherent vowels are modified by means of floating marks associated with the base letter. Removal of the inherent vowel is not always marked in Tibetan words and must be determined from context. Consonant clusters are sometimes rendered as conjuncts formed by stacking letters along a vertical axis. Conjuncts are represented in the text stream by placing a conjunct marker (virama called srog-med in Tibetan) between letters to be conjoined. Three letters (YA, RA, WA) normally change shape when they are members of conjuncts.

Punctuation: Common Tibetan punctuation includes shad U+104B to mark phrases. Shad is doubled to mark full stops (U+104E). Tseg (U+104A) is a syllable delimiter normally occurring after each syllable. Automatic line wrapping processes can wrap after occurrences of tseg that are also word boundaries. There are no interword spaces in Tibetan, a zero-width space might be used to set off word boundaries for automatic line-wrap algorithms (in the worst case, a break after tseg is better than between letters). U+104C and U+104B are a decorative variants of shad sometimes used at the beginning or end of a text. The character U+104F is an honorific flourish, double (Swasti) and triple forms of which are used at the beginnings of texts. It normally joins with one or two more occurrences of the same character to form ligatures, and is almost never used alone usually being followed by a decorative or doubled shad.

Transcription of Sanskrit: The Sanskrit retroflex letters are retained in the Tibetan script. In this proposal, the voiced aspirates are represented by conjuncts formed of simple consonants placed above the letter HA U+1021 (they could also be included as precomposed entities). U+103B is the visarga (see Devanagari), and (U+1038) is the anusvara.

The letter AA (U+101A, called ah-chung) is used as a subscript below another consonant, with or without a vowel sign, to indicate a long vowel. The Ah-chung subscript is represented explicitly by U+1025. The long vowels

incorporating ah-chung in typical fonts are sometimes written in other ways (e.g., the Landtsa font style uses double vowel signs exclusively to indicate vowel length). Hence, these long vowels have been encoded atomically.

When the Tibetan script is used to write Sanskrit, consonants are frequently stacked vertically in ways that do not occur in native Tibetan words; this usually indicates deletion of one or more vowel sounds. This behaviour is indicated in Unicode by insertion of a virama U+1030 between the consonants to be stacked, in a manner similar to the way virama functions in Devanagari.

Unicode does not encode the superscript and subscript forms for the letters WA, RA, and YA, as these shape changes can be algorithmically determined from context by the typographical rules for Tibetan; these shape changes are signalled by presence of the virama. Examples of these modified forms are ra-ta (ra subjoined) and ra-go (ra head). (The normal rules for form changes in written Tibetan are contained in various grammatical treatises on the Tibetan language and cannot be covered here in detail.)

The reversed gigu and long reversed gigu (U+102A, U+102B) are used in conjunction with the consonants RA and LA to represent Sanskrit vocalic /r/ and /l/, respectively. If these vowels are in initial position, the consonant+{gigu} combination is used, e.g. U+101C, U+102A for syllable-initial short vocalic /r/. When the vowel is used in a syllable with an initial consonant, the RA or LA forms a conjunct with the consonant. E.g., the sequence U+1000, U+1030, U+101C, U+102A codes the syllable /kr/ (with vocalic /r/).

Stacking Behaviour and Other Issues: In some exceptional cases, especially with transliterations from Sanskrit in mantras, arbitrary stacks of letters may occur. YA, RA, and WA may appear in stacks without the normal shape changing. This usage requires some type of stacking code or ligature making code distinct from the virama. There is currently no such control code in Unicode, and Tibetan is one example in which it is necessary for correct plaintext rendering.

In Unicode (Unicode 1.0), the Tibetan block introduction indicated that a Zero Width Joiner (U+200D) could be used to induce this type of stacking, but this usage has been disallowed due to the narrower interpretation of the ZWJ character.

There may be some use for the ``lenition mark'' (a non-spacing mark) that appeared in Unicode 1.0. It has been removed here, and the two precomposed forms FA and VA added (these are apparently used for transcribing foreign words). It is not known at this time whether the diacritic is ever used with other letters. (Letters U+1014 through U+1016 are primordial, and hence not an issue; the issue is with letters beyond FA and VA.)

The punctuation marks U+1034 through U+1037 are new, taken from the British proposal of July 1992 presented in commentary to the UK vote on DIS 10646. Atomic encoding of the long vowels does leave open the possibility of alternate spellings for long vowels. The order given here, especially after

U+1022, is somewhat different than the Unicode 1.0 order.

Encoding Structure: The Tibetan script block is divided into the following ranges:
U+1000 to 1024 Consonant Letters
U+1025 to 102F Non-Spacing Vowel Signs
U+1030 to 1039 Other non-Spacing Marks
U+1040 to 1049 Digits
U+104A to 104F Symbols
U+1050 to 105F Reserved for Tibetan

DRAFT TIBETAN CHARACTER NAMES

1000 TIBETAN LETTER KA
1001 TIBETAN LETTER KHA
1002 TIBETAN LETTER GA
1003 TIBETAN LETTER NGA
1004 TIBETAN LETTER CA
1005 TIBETAN LETTER CHA
1006 TIBETAN LETTER JA
1007 TIBETAN LETTER NYA
1008 TIBETAN LETTER REVERSED TA
1009 TIBETAN LETTER REVERSED THA
100A TIBETAN LETTER REVERSED DA
100B TIBETAN LETTER REVERSED NA
100C TIBETAN LETTER TA
100D TIBETAN LETTER THA
100E TIBETAN LETTER DA
100F TIBETAN LETTER NA
1010 TIBETAN LETTER PA
1011 TIBETAN LETTER PHA
1012 TIBETAN LETTER BA
1013 TIBETAN LETTER MA
1014 TIBETAN LETTER TSA
1015 TIBETAN LETTER TSHA
1016 TIBETAN LETTER DZA
1017 TIBETAN LETTER WA
1018 TIBETAN LETTER ZHA
1019 TIBETAN LETTER ZA
101A TIBETAN LETTER AA
101B TIBETAN LETTER YA
101C TIBETAN LETTER RA
101D TIBETAN LETTER LA
101E TIBETAN LETTER SHA
101F TIBETAN LETTER REVERSED SHA
1020 TIBETAN LETTER SA
1021 TIBETAN LETTER HA
1022 TIBETAN LETTER A
1023 TIBETAN LETTER FA
1024 TIBETAN LETTER VA
1025 TIBETAN AH CHUNG SUBSCRIPT

vowel length mark
1026 TIBETAN VOWEL SIGN I
1027 TIBETAN VOWEL SIGN II
1028 TIBETAN VOWEL SIGN U
1029 TIBETAN VOWEL SIGN UU
102A TIBETAN VOWEL SIGN REVERSED GIGU
102B TIBETAN VOWEL SIGN LONG REVERSED GIGU
102C TIBETAN VOWEL SIGN E
102D TIBETAN VOWEL SIGN EE
102E TIBETAN VOWEL SIGN O
102F TIBETAN VOWEL SIGN AU
1030 TIBETAN VIRAMA
= srog med
1031 TIBETAN CANDRABINDU
= kladkor (lekhor)
1032 TIBETAN CANDRABINDU WITH ORNAMENT
= datsekthikley
1033 TIBETAN HONORIFIC UNDER RING
1034 TIBETAN LANGCHEN NYOBUM
1035 TIBETAN JNIM TWO
1036 TIBETAN JNIM ONE
1037 TIBETAN HONORIFIC PREFIX
1038 TIBETAN ANUSVARA
1039 TIBETAN UNDER RING
103A TIBETAN CHUCHENYIGE
103B TIBETAN VISARGA
= mambcad (namchey)
103C TIBETAN COMMA
= tertsek (also used as Tibetan visarga)
103D TIBETAN DITTO
= duyik
103E TIBETAN LEFT BRACE
103F TIBETAN RIGHT BRACE
1040 TIBETAN DIGIT ZERO
1041 TIBETAN DIGIT ONE
1042 TIBETAN DIGIT TWO
1043 TIBETAN DIGIT THREE
1044 TIBETAN DIGIT FOUR
1045 TIBETAN DIGIT FIVE
1046 TIBETAN DIGIT SIX
1047 TIBETAN DIGIT SEVEN
1048 TIBETAN DIGIT EIGHT
1049 TIBETAN DIGIT NINE
104A TIBETAN TSEG
104B TIBETAN SHAD
104C TIBETAN RINCHANPHUNGSHAD
= rinchen pung shey
104D TIBETAN RGYANSHAD
= druishey
104E TIBETAN DOUBLE SHAD
X Devanagari double danda -> 0965

104F TIBETAN SINGLE ORNAMENT
= nyizla
= goyik (honorific; marks beginning of texts)

Proposal for Mongolian Encoding

The Mongolian draft proposal consists of a draft chart, a draft character names list, and a discussion in the form of a draft block introduction highlighting unresolved questions/issues. The content is based on the document "General Information on Mongolian Characters" registered by China as ISO-IEC JTC1/SC2/WG2 N628, May 1990. A great deal of useful input and materials were supplied by Lloyd Anderson of Ecological Linguistics, Professor John Krueger of Indiana University, Professor John Street of University of Wisconsin, Mr. Ochir of Inner Mongolian University, and Wayne Richter of Western Washington University.

Meta-Issue: Although the basic Mongolian alphabet which forms the core of this draft proposal can be easily laid out, there are important reasons for waiting perhaps one more year before freezing it into an encoding standard:

* A few of the remaining open issues are not minor, but rather address the basic relationship between the encoding of the language and the representation of the script.

* The native and scholarly communities who will be most affected by a future encoding standard are actively in the midst of attempting to resolve these issues.

* To ensure both logical soundness and practical acceptability of a chosen encoding design, we need to insist on possession of at least one reference implementation of any proposed encoding system; to date no such implementation is in hand.

Draft January 13, 1993

Mongolian (U+1060 - U+109F)

Development of the Mongolian script began in the 12th century, along with the enormous spread of Mongolian influence under Genghis Khan. The script has been in continuous use in the area which is now the Mongolian Autonomous Region of China (Inner Mongolia), and is now being taught again in the Mongolian People's Republic of the former USSR (Outer Mongolia) where it had been supplanted by Cyrillic in the 1930's. The script is used to write classical Mongolian, it serves to represent modern dialects in the areas just mentioned, and it is extended with additional letters for the Manchu and related Sibo languages.

The Mongolian script originated ultimately from the Aramaic, a right-to-left Semitic script. At some point the whole page underwent a rotation through 90 degrees counterclockwise, with the result that Mongolian is traditionally

written vertically in columns advancing from left to right. (In recent usage in China, when Mongolian is to be integrated with left-to-right horizontal text, the Mongolian lines may be rotated a further 90 degrees.) Although there is a non-accidental resemblance between Mongolian script and rotated Arabic script, the Mongolian language is not related to Arabic.

Encoding Difficulties. The relation between language and script in Mongolian is not at all simple. It is in some ways similar to that of English, in that the script retains archaisms not reflected in modern dialects (viz. English spellings such as "knight"). Thus the elements of the script do not correlate well with modern pronunciation. Further, some elements of the Mongolian script apply to more than one phoneme, somewhat as the English letter "c" applies to the disparate sounds [k] and [s]. The situation is further complicated by the fact that Mongolian letters assume contextual forms according to a variety of rules. The visible glyph forms do not constitute the proper "Mongolian alphabet", yet neither do modern phonetic elements. The recent Cyrillic-based alphabet is a phonetic representation of a modern dialect and does not correspond cleanly with the traditional script. Finally, the Mongolians themselves often present the traditional script in the form of a syllabary, but this is not a fruitful

Standards Activity. Work has been done and is ongoing on Mongolian standardization, but no firm standard definitions of the script's elements and their encoding have yet emerged. In DP 10646 a Mongolian chart appeared, based on the Chinese standard GB 8045-1987, which attempts to fit both underlying letters and glyphic fragments into the 96-cell framework of a 7- and 8-bit encoding. The result is a mixture of letters and glyphs in no discernible order, with too many entities to be an alphabet and too few to be a usable glyph set. Later China withdrew that proposal, and submitted ISO WG2 N628 of May 1990, which lists a sound enumeration of the basic alphabet. That list is taken as the basis of the current proposal.

Language Coverage. The primary focus of standardization is a set of letters sufficient to cover modern Mongolian usage of the traditional script, including the representation of foreign words. A secondary focus is basic coverage of classical Mongolian. Left for later are ancillary features of classical Mongolian (e.g. the classical numerals) and additional letters required to extend the script to other languages, particularly Manchu.

* Question: What is the full set of letters and features required for classical Mongolian?

* Question: What is the precise set of extension letters required for Manchu? (We currently have two listings, which do not agree.)

Encoding Principles. Ultimately, analytical sources such as dictionaries and textbooks provide a near-concensus on an underlying "basic Mongolian alphabet". This alphabet could be regarded an approximation to the phonetic repertoire of classical Mongolian, or from a modern perspective it could be regarded as a somewhat arbitrary collection of elements. (Likewise, English spelling in the Latin alphabet could either be analyzed historically, or

simply considered as rather arbitrary from the modern viewpoint). In any case, as with the encoding of the Arabic script, the visible glyphic forms of Mongolian are *not* taken to be the basic encoding elements. Those entities given in ISO WG2 N628 and in DP 10646 that are merely presentation forms are considered to be resources for the rendering process, and not part of the encoding. It is intended that they be excluded from Unicode/10646BMP.

Basic Alphabet. A collation of two dozen reference sources reveals that the list of 29 letters supplied in ISO WG2 N628 can serve well as the basis for an alphabetic Mongolian encoding, both in its content and in its ordering. Therefore, the present proposal consists precisely of that list, plus the interposition of two variant forms (U+106D, U+106F) and three somewhat marginal additional letters (U+1080, U+1081, U+1082).

Encoding Structure. The Unicode block for the Mongolian script is divided into the following ranges:
U+1060 to U+1062 Punctuation
U+1062 to U+1068 Vowels
U+1069 to U+107A Basic consonants
U+107B to U+107F Consonants for foreign words
U+1080 to U+1083 Additional letters
U+1084 to U+109F Currently unassigned

The "Punctuation", "Vowels", and "Basic consonants" groups form the core alphabet used in representing the Mongolian language. The "Consonants for foreign words" group are modern additions used represent important foreign sources such as Chinese and Russian. Taken together, these two groups form the well-attested basic alphabet of the modern script. The following "Additional letters" group contains forms whose usage (or status as an independent letter) is less well-attested, but which nevertheless are listed in many sources as independent letters. It is intended that the "Currently unassigned" space will accommodate later additions for classical Mongolian, Manchu, and any other rare necessities.

Alphabetical Order. The order of letters as assigned in the code chart is used in many modern sources from Inner Mongolia and China, although others use a slight variant in which L comes before M, and H comes before C. Traditional sources often arrange the main groups of consonants a bit differently:
modern: N B P Q/K GAMMA/G M L S SH T D CH ...
traditional: N Q/K GAMMA/G B P S SH T D L M CH ...

As generally, Mongolian implementations are expected to handle sorting via explicit processing, rather than relying on the code order to implicitly provide a desired alphabetical collation.

Glyphs Representing Individual Letters. Like many connected scripts, Mongolian has basic initial, medial, and final contextual forms of letters. However, the individual Mongolian letters are not normally presented alone in isolated forms (rather, the script is generally presented as a syllabary). To indicate the alphabetic letters in isolation or in abstraction, as in our

code charts, by convention the initial-form glyphs are used.

Latin Letter Names. There is a fairly well established tradition of Latin transliteration for Mongolian, which is here adapted for the character names. The alternative Latin transcriptions given in the names list are generally self-explanatory, except that the equivalence Q = X should be noted. The foreign K' or K* is denoted here by KK; this doubling is merely an artifice to avoid using ' or * in the name. The RH is a Chinese retroflex R.

Cyrillic Transcription. The generic mappings to the Cyrillic orthography are also given in the names list, even though the correspondence is not always unique in either direction. Although correlation of the traditional-script encoding system with the modern Cyrillic-based alphabet is desirable, it does not appear possible to design the encoding so as to permit algorithmic (or at least, simple) transcription between the two scripts. Thus, script convertibility is not taken as a goal of the current encoding system.

Character Shaping Behavior. An absolute requirement on any script encoding is that it be possible for a computer to take any valid sequence of underlying character codes and algorithmically render the appropriate visual form, given a repertoire of surface glyphs. In the case of the Mongolian traditional script, the required character shaping rules are particularly complex. Mongolian rendering systems have been built, but not yet in conjunction with the particular encoding approach proposed here. Until a reference implementation is thoroughly tested, there is no way to be certain that this proposed encoding is actually workable in all cases.

Issues of Alphabet Content and Spelling. Because the relationship between alphabetic letters and their rendition into glyphs is complex, often many-to-many, there are several cases where the choice of underlying spelling needs to be made clear. All of the following items are implicit questions for the reader, as to whether the model proposed here is accurate and workable.

* Content-based spelling. It is intended that the encoded spelling of Mongolian words be based on their underlying alphabetic content, not on their visual appearance. For example, the words ADA "devil" and ENDE "here" happen to appear identical, but they nevertheless should be spelled differently as indicated. The motivations for specifying content-based spelling are to avoid unnecessary arbitrariness, to preserve content for text processes such as search, and to align the spelling system insofar as possible with the more phonetic Cyrillic orthography.

* Spelling of vowels in non-initial syllables. This is a particular case of the preceding. The vowels A and E are written differently only in the first syllable, not in later syllables, likewise the vowel pair O and OE and the pair U and UE. Since the encoding is based on content and not appearance, the correct letter should be spelled regardless of how it might be rendered. Thus NARAN "sun" is a correct spelling, and *NAREN should be treated as a spelling error, regardless of the fact that *NAREN might be rendered with the same shape as NARAN.

* Spelling of vowel pairs O and U, OE and UE. This is another case of content-based spelling. The vowels O and U are always written identically, likewise the pair OE and UE. However, they are distinguished in the underlying alphabet, so the correct letters should be used in spelling.

* Spelling of consonant pairs Q and K, GAMMA and G. The sounds Q and K are in complementary distribution (in modern native Mongolian words), as are those represented by the pair GAMMA and G. Mongolian sources divide about evenly on whether to regard each pair as a single entity or two entities. In this encoding, each pair is resolved into two separate entities, primarily so that they can be distinguished if need be in spelling foreign or classical Mongolian words. In this system, all four characters are to be considered distinct in content-based spelling; this stricture also takes care of the fact that the initial glyphs for K and G appear identical.

* Spelling of digraphs NG and LH. There two digraphs NG (= N + G) and LH (= L + H). Mongolian sources divide about evenly on whether to regard each digraph as a single entity or two entities. In this encoding, each digraph is treated as a separate single letter, primarily because this makes it easier to accommodate to sources (e.g. dictionary listings) that so treat them.

* Spelling of consonants T and D. Although most listings of the Mongolian alphabet simply show a T letter and a D letter (as reflected in the current proposal), it appears that these are firmly distinguished only in foreign words, and that a content-based spelling for native Mongolian words should use instead a combined "Mongolian T/D". This distinction may surface in the assertion that initial "Mongolian T/D" has the appearance of U+1074 in our chart, while initial "foreign D" has the appearance of U+1075 in our chart. If this is true, then there would be even stronger reason for adding a separate "Mongolian T/D" letter, with the separate T and D then being reserved for transliteration of foreign words.

Possible Additional Letters. Various sources occasionally show other letterforms, which may perhaps be needed for classical Mongolian or for representing other languages. Any information on potential additional letters would be appreciated, below are a few of those under consideration.

* Aleph. Is it a rule of Mongolian spelling that every word must begin with a consonant? If so, as in semitic alphabets having this rule, there should be a "silent consonant" (aleph) to start words that phonetically begin with vowels. It could be said that the form of such an aleph is seen in our chart as the "cap" on the heads of the initial vowel forms U+1062 -> U+1068.

* "Left-tail" final forms of vowels A and E. Do there need to be separate letters for the foot-shaped leftward final forms of vowels A and/or E? This question means, does the occurrence of these forms depend on linguistic factors, such that if they were spelled via U+1062 MONGOLIAN LETTER A or U+1063 MONGOLIAN LETTER E, a computer could not algorithmically know how to render the appropriate visual form?

* Additional "z" consonant. Some sources show a form of U+107E MONGOLIAN LETTER Z that has its "toe" curling out-and-upward rather than down-and-inward. Is this a glyphic variant of U+107E MONGOLIAN LETTER Z, or is it a separate additional letter?

Separation and Concatenation of Word Components. The ASCII character U+0020 SPACE is intended to be used as the normal Mongolian wordspace. Also, a gap of white space often appears between a word and its endings, and sometimes even within a stem-morpheme. We need a fuller understanding of the semantics of Mongolian white spaces and the contextual-form behavior of the letters adjoining them. There is no doubt that some sort of special encodings will be needed at these junctures, the open *** question is whether one or more special "suffix-break" character(s) need be added, or whether the special mechanisms already available in the standard can suffice to represent these situations with reasonable semantics. Some of the relevant special characters already contained in the standard include:

* U+00A0 NON-BREAKING SPACE. Indicates a white space within a word, but does not break the word in two.

* U+200C ZERO WIDTH NON-JOINER. Invisible character that is regarded by adjacent letters as if it were a word boundary, causing them to assume the corresponding non-joining contextual form even in the middle of a word.

* U+200D ZERO WIDTH JOINER. Invisible character that is regarded by adjacent letters as if it were an ordinary letter, causing them to assume the corresponding joining contextual form even at the boundaries of a word. So, for example, if the spelling sequence is (reading down):

...
some letter
last letter before gap
U+200C ZERO WIDTH NON-JOINER
U+00A0 NON-BREAKING SPACE
U+200D ZERO WIDTH JOINER
first letter after gap
some letter
some letter
...

then the word would remain logically unbroken but it would contain a white gap, the last letter before the gap would assume final contextual form, and the first letter after the gap would assume medial contextual form. The question is whether such mechanisms are workable for Mongolian, and if so, appropriate.

DRAFT MONGOLIAN CHARACTER NAMES LIST

@ Punctuation
1060 MONGOLIAN COMMA
1061 MONGOLIAN PERIOD

@ Vowels
1062 MONGOLIAN LETTER A
= Cyrillic A
1063 MONGOLIAN LETTER E
= Cyrillic REVERSED E
1064 MONGOLIAN LETTER I
= Cyrillic II
1065 MONGOLIAN LETTER O
= Cyrillic O
1066 MONGOLIAN LETTER U
[= WG2 N628 # 5 "UO"]
= Cyrillic U
1067 MONGOLIAN LETTER OE
[= WG2 N628 # 6 "UE"]
= Cyrillic LATIN SMALL LETTER BARRED O (U+0275)
alternative Latin transcription: O UMLAUT
1068 MONGOLIAN LETTER UE
[= WG2 N628 # 7 "U"]
= Cyrillic STRAIGHT U
alternative Latin transcription: U UMLAUT

@ Basic consonants
1069 MONGOLIAN LETTER N
= Cyrillic EN
106A MONGOLIAN LETTER B
= Cyrillic BE
106B MONGOLIAN LETTER P
= Cyrillic PE
106C MONGOLIAN LETTER Q
[= WG2 N628 #11 "H"]
= Cyrillic KHA (which is also used for 107F)
alternative Latin transcription: X
back vowel harmony correspondent to the following
106D MONGOLIAN LETTER K
front vowel harmony correspondent to the preceding
106E MONGOLIAN LETTER GAMMA
[= WG2 N628 #12 "G"]
= Cyrillic GE
back vowel harmony correspondent to the following
106F MONGOLIAN LETTER G
front vowel harmony correspondent to the preceding
1070 MONGOLIAN LETTER M
= Cyrillic EM
1071 MONGOLIAN LETTER L
= Cyrillic EL
1072 MONGOLIAN LETTER S
= Cyrillic ES
1073 MONGOLIAN LETTER SH
= Cyrillic SHA
alternative Latin transcription: S CARON

1074 MONGOLIAN LETTER T
= Cyrillic TE
1075 MONGOLIAN LETTER D
= Cyrillic DE
1076 MONGOLIAN LETTER CH
= Cyrillic CHE
= Cyrillic TSE (which is also used for 107D)
alternative Latin transcription: C CARON
1077 MONGOLIAN LETTER JH
[= WG2 N628 #20 "ZH"]
= Cyrillic ZHE (which is also used for 1083)
= Cyrillic ZE (which is also used for 107E)
alternative Latin transcriptions: J CARON, ZH
1078 MONGOLIAN LETTER Y
= Cyrillic SHORT II
1079 MONGOLIAN LETTER R
= Cyrillic ER
107A MONGOLIAN LETTER V
[= WG2 N628 #23 "W"]
= Cyrillic VE
alternative Latin transcription: W

@ Consonants for foreign words
107B MONGOLIAN LETTER F
= Cyrillic EF
107C MONGOLIAN LETTER KK
[= WG2 N628 #25 "K"]
= Cyrillic KA
alternative Latin transcriptions: K', K*
107D MONGOLIAN LETTER C
[= WG2 N628 #26 "TS"]
= Cyrillic TSE (which is also used for 1076)
alternative Latin transcription: TS
107E MONGOLIAN LETTER Z
[= WG2 N628 #27 "DS"]
= Cyrillic ZE (which is also used for 1077)
alternative Latin transcription: DZ

107F MONGOLIAN LETTER H
= Cyrillic KHA (which is also used for 106C)

@ Additional letters
1080 MONGOLIAN LETTER NG
1081 MONGOLIAN LETTER LH
1082 MONGOLIAN LETTER EH
1083 MONGOLIAN LETTER RH
[= WG2 N628 #29 "R"]
= Cyrillic ZHE (which is also used for 1077)
alternative Latin transcriptions: Z CARON, ZH

Acknowledgments

The following individuals contributed greatly to the production of Technical Report #2: Lloyd Anderson, Ken Whistler, Peter Lofting, Rick McGowan

The content of the Mongolian proposal is s based on the document "General Information on Mongolian Characters" registered by China as ISO-IEC JTC1/SC2/WG2 N628, May 1990. A great deal of useful input and materials were supplied by Lloyd Anderson of Ecological Linguistics, Professor John Krueger of Indiana University, Professor John Street of University of Wisconsin, Mr. Ochir of Inner Mongolian University, and Wayne Richter of Western Washington University.

Copyright
