Appendix E: Proposed Scripts for Future Versions of the Unicode Standard

This informative appendix presents draft layouts for several scripts that are currently under study for future inclusion in the Unicode standard. These scripts are not included in the version 1.0 of the Unicode standard. Comments and suggestions regarding these scripts are invited by the Unicode consortium.

Ethiopian

The Ethiopian script is used for writing several languages of the sub-Saharan area, including Amharic, Tigre, and Oromo. The script, which is based on the writing of a dead language, Ge'ez, is graphically consistent. However, it is a syllabary rather than an alphabet, which has several encoding consequences discussed in the following section.

Array Structure. The basic Ge'ez syllabary is traditionally arranged as an array of thirty-three consonant initials crossed with seven vowel finals. Since most of the consonants also take a labialized final, this can be expanded to a 33 x 8 array, which is ideal for encoding. This orderly array forms the basis for the Unicode Ethiopian block; other characters are added afterward in a less systematic fashion.

Encoding Structure. The Unicode character block for the Ethiopian script is divided into two adjacent blocks Ethiopian and Extended Ethiopian.

Diacritical marks. The Ethiopian syllabic letterforms in most cases reveal their origin as composites of a consonant base character plus a vowel diacritical mark, with labialization represented by a further diacritical mark. In the Unicode encoding, the syllabic letters are represented as whole codes, rather than by composition, because the composites have truly become the units of the script (and besides, the compositional rules are very irregular). However, a syllabary is more difficult to extend than an alphabet, and there may be merit in accomplishing some extensions via the application of diacritical marks. The few marks in this range appear to be the most effective in producing extensions, and are provided in case there is a desire to use them this way.

Extended Ethiopian Letters. This group includes some extensions of the basic syllabary, plus a set of labialized series that is now part of the standard script (and which in some cases replicates syllables in the main array). The characters are arranged according to the same N x 8 scheme as the main array. The names given to the extended Ethiopian characters are somewhat artificial, intended mainly to create a unique identifier. The Ethiopian script has been extended for some relatively obscure languages which may have little tradition of printed typography, and obsolete alternative

forms of some letters also exist. The available information on variant letter forms is often sporadic and inconsistent, so some of the codes may be regarded as unneeded (and/or invalid) for some applications. It is assumed that the encoding of various languages will make use of various different subsets of these extensions. Given the imperfection of information and the bulkiness of extensions to a syllabary, the currently unassigned range has been made larger for Ethiopian than for other scripts. (Enough singly-attested forms have already been collected to fill it).

Sinhala

The Sinhala script (also known as Sinhalese, the majority language of Sri Lanka) was removed from version 1.0 of the Unicode standard because new information about Sinhala encoding was received prior to publication of the Unicode standard. The Sri Lankan government is currently producing a standard encoding for Sinhala, and the proposed (draft) standard differs significantly from both the draft layout shown here and the Indian standard (ISCII) upon which the Unicode draft was based.

Mongolian

The proposed Unicode structure for encoding Mongolian includes no ideal forms for characters. The only forms for Mongolian are the contextual shapes. Nevertheless, the Unicode standard proposes an encoding which identifies a single form for each Mongolian letter for encoding; Mongolian thus requires contextual shaping rules for rendering, much as for Arabic.

Mongolian is traditionally rendered vertically.

Because the traditional Mongolian script is undergoing a revival in the Mongolian People's Republic, there may be standards-related activity in the near future which could significantly affect the encoding of the Mongolian script.

Burmese and Khmer are being investigated.