From: Stefan Baums
Date: 2002-11-02 20:19:35 -0800
Subject: Note for the UTC on the encoding of Brahmi in Unicode

Dear UTC,

Here are our remarks on the proposed encoding of Brāhmī in Unicode:

Overview of the history of the Brāhmī writing system:

We find the Brāhmī writing system first used in the inscriptions of the Indian emperor Aśoka (third century BCE), in the northwestern part of his empire also in parallel with Kharoṣṭhī, Greek and Aramaic. The Brāhmī script follows the same principles as the (possibly somewhat older) Kharoṣṭhī script, i.e., it is of the akṣara or abugida type, based on the akṣara as the graphical unit of written strings, with the vowel "a" inherent in consonant signs. Other vowels are indicated by the obligatory addition of combining vowel diacritics, and the absence of a vowel (in the case of consonant clusters – when one consonant immediately follows upon another) is indicated by combining consonant signs into ligatures. In contrast to Kharoṣṭhī, the Brāhmī script uses a separate set of letters for initial vowels (where Kharoṣṭhī combines a null-consonant with the vowel diacritics), and also in contrast to Kharoṣṭhī, Brāhmī uses separate vowel diacritics and letters for long and short vowels (the Kharoṣṭhī script does usually not differentiate between e.g. [ɪ] and [i□] of the spoken language). The origin of the actual graphical shapes that make up Brāhmī akṣaras remains unclear, but it seems not entirely unlikely that the design of the Brāhmī script was commissioned by the emperor Aśoka specifically for the production of his inscriptions, based on a background knowledge of Kharoṣṭhī (itself derived from Aramaic) and maybe other, non-Indian scripts.

The history of pre-modern Brāhmī (the subject of our upcoming Unicode proposal – the Brāhmī script as used before the evolution of its modern forms that began being associated with the emerging local literary languages of India around the year 1000 CE, and that are already encoded in Unicode) in India is usually subdivided into four phases: Early Brāhmī (3rd to 1st century BCE), Middle Brāhmī (1st to 3rd century CE), Late Brāhmī (4th to 6th century CE) and the Transitional Scripts (7th to 10th century CE). In the Old Brāhmī period the script is very uniform in appearance; in the Middle Brāhmī period local style differences begin to emerge; in the Late Brāhmī period this trend intensifies, and through the Transitional Scripts leads to the distinct modern scripts. Besides these developments in mainland India, the Brāhmī script was transplanted around the beginning of the Common Era to Sri Lanka (and the Maldives), Central Asia, and South East Asia. In these countries, Brāhmī was typically first used for the import and then local production of Indian-language

(Sanskrit and Pāli) texts, while palaeographically staying very closely linked to the Indian prototype. Only after a matter of one or several hundred years, these extra-Indian forms of Brāhmī began being applied to the local languages and to develop in directions of their own.

The coexistence of varieties of Brāhmī in ancient India, and in the life of the modern scholar

One very important fact to realise about the use of Brāhmī in ancient India and the Indian cultural world is that in any given place and time, only one variety or style of Brāhmī was used, regardless of the texts that were written in it. In North India, Sanskrit as well as Middle-Indian and later Modern Indian texts were all written in precisely the same North Indian script; in Central Asia, (sometimes the same) Sanskrit texts as well as texts in Central Asian languages were written in the very same Central Asian Brāhmī; and on Sri Lanka, Sanskrit texts (again in many cases the same as in other parts of the Indian world), Pāli texts and Sinhalese texts were written in the same Sri Lankan form of Brāhmī. For the modern scholar (the user of the upcoming Brāhmī encoding proposal), studying any one given text in e.g. Sanskrit, this means that he will have to deal with manuscript material written in e.g. the Central Asian variety of Brāhmī, a northwest Indian variety of Brāhmī, and a north Indian variety of Brāhmī, but all containing the same text in the same language. For the modern Indian manuscript scholar, the script varieties employed are thus a rather superficial aspect of their business (unless they are writing a palaeography), in a way quite similar to a Classicist editing a Classical Latin text from near-contemporary papyri or inscriptions, from minuscle and majuscle medieval manuscripts, and from Renaissance manuscripts. In the case of the Classical scholar, all the scripts of his manuscripts will be encoded using the same Unicode subset, and if and when he wishes to discuss scribal issues, he will just apply different fonts, but leave his text encoded in the same character sequences.

The proposed encoding strategy for pre-modern Brāhmī

To facilitate the work of the expected scholarly user community and on the parallel obtaining in e.g. the various script varieties that a Classicist deals with and their Unicode encoding, we suggest that in the encoding of the pre-modern varieties of Brāhmī, unification should be used to the greatest extent possible. A “ka” should be a “ka” and be represented by the same character code, regardless of which subvariety of pre-modern Brāhmī it occurs in. We also believe that this will do justice to the observed historical development of Brāhmī. While at first glance the diversity of pre-modern Brāhmī varieties seems greater than that of pre-modern Latin scripts, it is not so to a signifant degree. While it is true that in some outposts of Brāhmī use special diacritics were introduced for the representation of non-Indian languages (such as a double-dot-above for Tokharian), the

same is true for the Latin script in Europe (such as the umlaut diacritic in German). While it is true that in the history of Brāhmī some special characters were added for new uses (such as the "ṅa" character when Brāhmī began to be adapted to the writing of Sanskrit), the same is also true for the Latin script in Europe (such as the þ of Icelandic or the ø of Danish).

For the same historical and practial reasons that the Carolingian minuscule, or the German or Icelandic varieties of the Latin script, are not encoded separately in the Latin-script part of Unicode, we believe that pre-modern Brāhmī should receive a unified encoding.

The state of the SMP Roadmap

The current SMP Roadmap (http://www.unicode.org/roadmaps/smp-3-3.html) does not represent our ideas for the encoding of Brāhmī. It has one Brāhmī script outside the block labelled "Brahmic scripts":

> 10860–1087F Balti and others (an old discussion document is linked to: http://www.dkuug.dk/JTC1/SC2/WG2/docs/n2042.pdf)

and contains the following assortment of Brāhmī or similar scripts in the range 11000 to 117FF:

> 11000–1104F Brahmi

A short proposal by Michael Everson (http://www.dkuug.dk/JTC1/SC2/WG2/docs/n1685/n1685.htm) is linked to that suggests an encoding for what we called Early Brāhmī. The proposal, however, does not address the historical dimensions of Brāhmī, and does not take into account recent scholarly work on Brāhmī.

> 11080–110BF Pyu

Pyu is an early (ca. 5th century CE) South East Asian transplant of the Kadamba variety of Brāhmī, used for the Pyu language. The whole complex question of when historical South East Asian scripts can still be represented by our Common Brāhmī encoding, when they should be represented by the encodings for the modern South East Asian Scripts, and when an encoding separate from these two (and to be developed by someone else) should be employed, needs to be addressed in the scholarly community. Pyu is definitely not a priority for us.

> 110C0–110FF Balinese

Similar remarks apply to the historical, Brāhmī-based script of Bali.

> 11100–1113F Soyombo

Soyombo (on the Roadmap wrongly marked as having been proposed) was

invented in 1686 by a Mongolian monk for the writing of Mongolian, Tibetan, and Sanskrit. While supposedly modelled on a form of Brāhmī, it really remains to be investigated whether it actually shares the systemic features of the Brāhmī writing system. Again, Soyombo is not a priority for us.

11140–1117F Ahom

A local Brāhmī script used for Assamese before the adoption of the Bengali script for that purpose.

11180–111DF Turkestani

This presumably means the various Central Asian varieties of Brāhmī (North Turkestan Brāhmī and South Turkestan Brāhmī, adapted with various additional diacritics to the use of Central Asian languages such as Tokharian).

11200–1125F Kaithi

A local Brāhmī script.

11280–112DF Rejang

A South East Asian Brāhmī-based(?) script of Sumatra. Cf. the remarks above, peripheral to an encoding of Brāhmī.

11300–1135F Landa
11380–113DF Modi

Local Brāhmī scripts of the Punjab and Maharashtra, respectively.

11400–1145F Chalukya (Box-Headed)
11480–114DF Chola
11500–1155F Satavahana

Three south Indian varieties of Brāhmī, named after the dynasties that employed them in their inscriptions. These three names are in no way representative of the script culture of pre-modern South India; a classification into e.g. Proto-Kannada-Telugu, Grantha, Tamil, and Vaṭṭeẕuttu would be more appropriate.

11580–115DF Takri

A north Indian local script, used for the writing of Western Pahari dialects.

It will have become clear that the names in the SMP Roadmap in no way represent a satisfactory classification of pre-modern varieties of the Brāhmī script. More importantly, any such subclassification into small local varieties runs counter to our case for greatest possible unification.

Work underway

We are in the process of writing a formal proposal for the encoding of Pre-modern Brāhmī. To do justice to all the local varieties of the script, and to ensure that they all will ultimately be encodable in Unicode, will require communication with experts in other subfields of Brāhmī palaeography. We have established contact regarding the upcoming Unicode proposal with interested colleagues in the relevant subfields – representative of the targeted user community – and our next step will be a personal meeting in connection with the XIIIth Conference of the International Association of Buddhist Studies, 7th to 13th December in Bangkok.

This note is intended to make the UTC aware of the issues involved. We welcome any technical and – to the extent possible – factual remarks that you wish to make regarding the encoding of Brāhmī, and will be more than happy to answer any questions you may have.

Best regards,

Stefan Baums
Andrew Glass

Asian Languages and Literature
University of Washington