

# Towards a Computer Encoding for Brāhmī\*

STEFAN BAUMS

## 1. INTRODUCTION

This paper describes a project to develop a computer coding system for the Brāhmī script and to submit the result to the international standardisation body responsible for specifying the computer representation of the world's writing systems.<sup>1</sup> Official standardisation of the coding system will encourage its support across a wide range of computer systems, and is the best way to ensure its long-term survival.

---

\* I would like to thank Lore Sander, who furnished advice on Khotanese and Tumshuqese Brāhmī; Gudrun Melzer, who provided an image of the rare *ṛ mātrā* from the Gupta period manuscripts that she studies; and Jost Gippert, Christiane Schaefer and Georges-Jean Pinault, who suggested improvements to the description of Tocharian Brāhmī. Andrew Glass carefully went through a draft of this paper and has been involved in the encoding project from the beginning. Any remaining inadequacies should be attributed to the present author. Contributions by specialists in particular forms of Brāhmī to the encoding project described here are not only highly welcome but strongly encouraged. It is important to ensure that the proposed encoding in fact meets the needs of the Indological community before it is submitted, accepted and implemented.

<sup>1</sup> The present paper is complementary in purpose to the formal proposal document (available at the URL <http://depts.washington.edu/ebmp/brahmi-encoding/>): the former aims to present the technical issues to the scholarly world, who are the future users of the technology, while the purpose of the latter is to represent our scholarly needs to a technical audience, who will implement the technology. The proposal document has now also been presented by my colleague Andrew Glass at the "International Symposium on Indic Scripts: Past and Future", 17 to 19 December 2003, convened by the research group "Grammatological Informatics Based on Corpora of Asian Scripts (GICAS)" at the Tokyo University of Foreign Studies, and it is being discussed on the Unicode Consortium's "indic" email list.

This will enable Indologists working with the Brāhmī script to fully use the generic text-processing functions of their software, such as searching and indexing. It will also ensure that Brāhmī data can be exchanged between different computer systems without the need for conversion, and that it will still be legible many years from now, even as general computer technology changes.

At the root of a computer coding system lies what is called a "coded character set". A coded character set is nothing more than a table that assigns to each character of a script a unique numerical value called its "code point". In the coded character set that is now commonly used throughout the world, the Latin character A, for instance, is assigned the code point 41, the Devanagari character अ the code point 915, and the Chinese character 佛 the code point 4F5B<sup>2</sup>; such a mapping between characters and numerical code points is necessary because numbers are the only entities that computers can handle internally. In addition to defining a coded character set, a coding system will also specify other properties of a script, such as how characters combine with each other in a line of text, which of course is particularly important for writing systems of the Indian type.

Historically, there have been dozens of mutually incompatible coding systems. One used the code point C7 for the French character Ç, another used the very same code point for the Russian character Ч and a third one for the Greek character Η. This confusing situation, which seriously impeded the exchange of textual data across linguistic boundaries, was partly due to technical reasons that have since become irrelevant<sup>3</sup> and partly to lack of international coordination. In the late 1980s, however, an initiative was started to remedy the situation by

---

<sup>2</sup> It is customary to write code points not as decimal, but as hexadecimal numbers, that is, numbers using the base 16, with digits from 0 to 9 and then A to F. This is a matter of technical convenience (two hexadecimal digits will always represent one byte), but has no deeper significance. In the examples used above, hexadecimal 41, 915 and 4F5B would correspond to decimal 65, 2325 and 20,315.

<sup>3</sup> Especially the use of only one byte for the representation of characters, imposing an upper limit of 256 on the number of characters that could be contained in the same coded character set.

developing one single coding system that would be capable of representing all the world's writing systems in a uniform and systematic fashion. This work was carried out jointly by the Unicode Consortium (representing the computing industry, academic institutions and government bodies<sup>4</sup>) and by the International Organization for Standardization<sup>5</sup>. The first version of this new coding system standard, known as the Unicode Standard or IS 10646, saw the light of the day in 1991, covering all the major modern scripts of the world with a total of 28,302 assigned characters. Since then it has been continually improved, and in its present version (2003) coverage has expanded significantly to include modern minority scripts and historical characters and scripts, with a total of 96,382 assigned characters (TUS, 1355f.).

## 2. THE COMPUTER ENCODING OF INDIAN SCRIPTS IN GENERAL

One of the defining principles of the Unicode Standard is that it encodes characters, the abstract units of a writing system, not glyphs, the graphical shapes that realise those units in writing. To take an example from the Indian sphere, क and ण are glyphs of the Devanagari writing system that represent one character each, namely U+0915 DEVANAGARI LETTER KA and U+0937 DEVANAGARI LETTER SSA<sup>6</sup> respectively, or *ka* and *ṣa* in Indological transliteration, whereas क्ṣ is a single glyph that represents a sequence of characters, namely <U+0915 DEVANAGARI LETTER KA, U+094D DEVANAGARI SIGN VIRAMA<sup>7</sup>, U+0937 DEVANAGARI LETTER SSA>, or *kṣa* in Indological transliteration. It is clear that in the case of Indian writing systems, there is no one-to-one correspondence of characters and glyphs. The mapping

---

<sup>4</sup> Including the Governments of India, Pakistan and Tamil Nadu.

<sup>5</sup> More precisely, the working group ISO/IEC/JTC1/SC2/WG2.

<sup>6</sup> Here and in the following, code points are prefixed with "U+" to indicate that they refer to the Unicode coding system. Every character is also assigned an official name in the Unicode Standard, and such names are as here and in the following written in small caps. Sequences of Unicode characters are presented between angle brackets.

<sup>7</sup> See below for further discussion of the VIRAMA control character.

from a sequence of Unicode characters in a text file to a sequence of glyphs, for example on the display of a computer, is not a subject matter of the Unicode Standard, but rather is handled by complementary font technology such as the widespread OpenType format.<sup>8</sup> Nevertheless, it is the responsibility of Unicode to provide a coded character set that both represents the units of a real-world writing system as faithfully as possible and lays the foundations for a natural glyph rendering of that writing system at the font layer.

Akṣara writing systems such as Brāhmī and Kharoṣṭhī have two characteristic properties that a computer coding system needs to take account of: postconsonantal vowels are written in the form of diacritic elements added to the core of the akṣaras; and sequences of more than one consonant are written as consonant clusters, which are graphic combinations that often involve simplifications of their constituent elements. Unicode handles the first property by providing, in addition to the initial vowel characters (e.g. उ U+0909 DEVANAGARI LETTER U), a separate series of vowel diacritics (e.g. ु U+0941 DEVANAGARI VOWEL SIGN U) that have the property of being “combining characters”, meaning that they cannot occur independently but are added to and to be interpreted in combination with the preceding character. Thus, the Devanagari akṣara कु will be encoded as a sequence of two Unicode characters, the independent U+0915 DEVANAGARI LETTER KA followed by the combining character U+0941 DEVANAGARI VOWEL SIGN U. It is an important principle of the Unicode Standard that characters always follow each other in the phonetic or logical order, disregarding the physical order of their glyphs when rendered. Therefore, the akṣara कि *ki* will be encoded exactly parallel to कु *ku*, even though the *i* diacritic precedes the *ka* core of the akṣara when rendered (<U+0915 DEVANAGARI LETTER KA, U+093F DEVANAGARI VOWEL SIGN I>); and the parallelism even holds in the case of those Indian scripts that have vowel diacritics consisting of more than one

---

<sup>8</sup> An open standard initially developed by Adobe Systems and Microsoft Corporation.



physical part, such as Bengali, কো ko being encoded as <U+0995 BENGALI LETTER KA, U+09CB BENGALI VOWEL SIGN O>.

Consonant clusters, the other characteristic feature of akṣara writing systems, are handled with the help of a special control character called VIRAMA which is placed between any two consonants that are to be joined. The consonant cluster न्ध ndha, for instance, will be encoded as <U+0928 DEVANAGARI LETTER NA, U+094D DEVANAGARI SIGN VIRAMA, U+0927 DEVANAGARI LETTER DHA>, and the same applies to longer consonant sequences such as न्ध्र ndhra <U+0928, U+094D, U+0927, U+094D, U+0930 DEVANAGARI LETTER RA>. The abstract virama control character is thus by no means necessarily displayed as a virāma glyph (्). In practice, a sequence of consonant characters joined by VIRAMA will be rendered as a proper conjunct wherever made possible by the font, and with a visible virāma mark only in those cases where the font lacks a conjunct glyph for the consonant sequence in question (e.g. ण्म ñma <U+0919 DEVANAGARI LETTER NGA, U+094D, U+092E DEVANAGARI LETTER MA>). It may be noted that this behaviour closely mirrors traditional typographic practice which resorted to aesthetically less pleasing (but semantically equivalent) virāma combinations where the metal type lacked a particular conjunct. Final consonants are similarly encoded with the VIRAMA control character (उपनिषत् upaniṣat <..., U+0924 DEVANAGARI LETTER TA, U+094D>).<sup>9</sup>

The Unicode Standard currently covers the following South Asian Brāhmī-derived scripts (2003, version 4.0): Devanagari, Bengali, Gurmukhi, Gujarati, Oriya, Tamil, Telugu, Kannada, Malayalam, Sinhala, Tibetan and Limbu (as well as the major Brāhmī-derived scripts of Southeast Asia). As the case of Limbu shows, efforts are now underway to add the modern minority scripts to the Unicode standard, and the same is true of historic writing systems: Ogham, Old Italic, Runic, Gothic, Ugaritic, Linear B, the Cypriot Syllabary and a very large number of premodern Han characters have recently been added to its

---

<sup>9</sup> For a more detailed and technical discussion of the modern Indian scripts in Unicode see TUS, 217-262.

repertoire. Thus, with the completion of the modern national scripts, the focus of the Unicode development effort is in effect shifting from the commercial and political to the scholarly.<sup>10</sup>

### 3. THE COMPUTER ENCODING OF KHAROṢṬHĪ AND BRĀHMĪ

On this background, Andrew Glass and the author of the present paper have recently co-authored a proposal for the addition of Kharoṣṭhī to the Unicode Standard,<sup>11</sup> in close and productive cooperation with members of the Unicode Technical Committee and other experts.<sup>12</sup> The Kharoṣṭhī proposal has now been accepted by the Unicode Consortium and is currently awaiting ratification by the International Association for Standardization; it is expected to be published as part of version 4.1 of the Unicode Standard.

Based on our positive experience with the Kharoṣṭhī proposal, it was felt that the more formidable task of developing an encoding for the variants of Brāhmī should now also be attempted. There have been several efforts in the past to formulate encodings for particular sub-varieties of Brāhmī, but only one of these has advanced to the proposal stage.<sup>13</sup> In contrast to previous approaches, we are convinced of the importance of providing a unitary encoding for all subvarieties of Brāhmī: a single encoding that provides a character repertoire sufficient for the representation of all the forms of the script that occur in manuscripts and inscriptions from its beginnings in the 3rd century BCE until the emergence of the modern Indian scripts by around 1000

---

<sup>10</sup> Everson (2002) gives an overview of the historical and minority scripts that remain to be encoded, and refers to the Script Encoding Initiative that has been set up at the University of California at Berkeley to support and coordinate work on these scripts.

<sup>11</sup> Available at <http://depts.washington.edu/ebmp/downloads/Kharoshthi.pdf>.

<sup>12</sup> Especially Rick McGowan (Unicode Consortium), Kenneth Whistler (Sybase Inc.), Deborah Anderson (University of California at Berkeley) and Michael Everson (Everson Typography).

<sup>13</sup> Everson 1998, for Aśokan Brāhmī.

CE. Considering the strict separation in Unicode between an abstract character layer defined by the encoding and a concrete glyph layer produced by a computer's rendering system on the basis of specific fonts, it would be a duplication of effort and waste of architectural resources to reflect those palaeographic differences at the encoding layer that will already be reflected at the font layer:

Aśokan Brāhmī font	Kuṣāṇa Brāhmī font	Gupta Brāhmī font
Aśokan Brāhmī encoding	Kuṣāṇa Brāhmī encoding	Gupta Brāhmī encoding

The varieties of premodern Brāhmī should instead be given a unitary encoding, their differences being expressed at the font level only:

Aśokan Brāhmī font	Kuṣāṇa Brāhmī font	Gupta Brāhmī font
Brāhmī encoding		

This architecture has the great advantage of reflecting not only the differences of the variants of Brāhmī at one level, but also their unity at another level. The fact that in this model there is one sense in which, for example, a *ka* is always a *ka*, whether it appears in Eastern Gupta Brāhmī or in Gilgit-Bamiyan type II Brāhmī, makes it possible to have the computer search for all *kas* across script varieties, to index them and to group them together – an immense practical benefit for any sort of palaeographic work.<sup>14</sup>

---

<sup>14</sup> In the palaeographic treatment of the Buddhist manuscripts in the Schøyen Collection, for example, we have to do with a large number of varieties of the Brāhmī script (see Sander 2000 for a description of some of these) that nonetheless all were used in related textual traditions, had a well-defined user group, and evidently all found their way into the same monastic library. From a practical perspective, the advantage of a unitary encoding for those who constantly have to handle these script variants in parallel is readily apparent.

Moreover, the historical development and differentiation of Brāhmī was of such a gradual nature that any of the various suggested sub-classifications inevitably involves a strong element of subjectivity (cf. Salomon 2003: 73f.), just as the ongoing discovery of significant new manuscript<sup>15</sup> and inscriptional<sup>16</sup> source material keeps changing our understanding of the historical processes involved. It has to be kept in mind in this connection that in order to ensure stability of the standard the Unicode Consortium has a strict policy never to remove a character or script once it has been encoded. Any particular subclassification of Brāhmī encoded now would therefore be set in stone for all time, and would inevitably lose congruity with our emerging understanding, even assuming that a suitable consensus on classification could be reached among Brāhmī scholars at the present time. For technical, scholarly and practical reasons, it is therefore strongly preferable to express the variety of Brāhmī at the font level, and to express its unity at the encoding level.

#### 4. BRIEF OVERVIEW OVER THE HISTORY OF BRĀHMĪ

The earliest undisputed examples of writing from historical India are the edicts of Aśoka from the middle of the 3rd century BCE. While most of these inscriptions are in Brāhmī script, in the Indian Northwest Kharoṣṭhī, Aramaic and Greek were used as well. It would appear that the earliest known form of Brāhmī presupposes the existence of Kharoṣṭhī: Brāhmī follows the same system of vowel marking as Kharoṣṭhī, but has a greater number of distinct vowel signs that allow for a better representation of Indian languages, and Kharoṣṭhī has clear historical associations (with the Aramaic script) that Brāhmī lacks. It has been suggested that the Brāhmī script was specially invented for use in the royal inscriptions of Aśoka or similar documents, on the

---

<sup>15</sup> Such as the several thousand fragments of Buddhist manuscripts now in the Schøyen Collection, Oslo (Braarvig 2000).

<sup>16</sup> For example the Tamil Brāhmī inscriptions that have come to light in the course of the last century (Maḥādevan 2003).

basis of an acquaintance with the Kharoṣṭhī and possibly also Aramaic and Greek scripts.<sup>17</sup>

The further development of Brāhmī script is characterised by very gradual changes in the forms of letters conditioned by cursivisation and modification of stroke order, and by changes in the writing utensils used. The characteristic headmarks of the modern Devanagari and Bengali scripts, for instance, have their origin in the mark left where a reed pen first touches the writing surface,<sup>18</sup> while the trend towards round letter forms in the southern varieties of Brāhmī is attributed to the South Indian technique of incising the letters into palm leaves, where straight lines would have tended to split the leaf.<sup>19</sup>

As a chronological framework, the system employed by Salomon 2003: 94-99 (based on Sircar 1970-1971: 113) is adopted here, distinguishing Old, Middle and Late Brāhmī, Transitional Scripts, and the modern Indian scripts. While spreading towards southern India in the Old Brāhmī period (3rd to 1st centuries BCE), the script underwent experimental and rather shortlived systemic innovations and saw the introduction of new characters in the Old Tamil and Bhattiprolu inscriptions (see below). In the Middle Brāhmī period (1st to 3rd centuries CE), regional variation increased; Dani distinguishes between East Indian, Northwest Indian, Northwest Deccan and South Indian styles.<sup>20</sup> Brāhmī was now for the first time being used to represent Sanskrit, and for this purpose five or six new characters were added to the script (*r*, *ṛ*, *ḷa*, *au*, *ḥ* and maybe *ṇa*<sup>21</sup>). A special device was introduced for the marking of vowelless consonants, used both in Sanskrit, where it is called *virāma* and first occurs in manuscripts of the 1st century CE, and in Tamil, where it is called *pulli* and attested in inscriptions from the 2nd century CE (Mahadevan 2003: 198). In the course of trade relations and cultural exchange, the Brāhmī script was

---

<sup>17</sup> Falk 1993: 109-112.

<sup>18</sup> Dani 1986: 79-81; Salomon 1998: 32, 34.

<sup>19</sup> Salomon 2003: 83.

<sup>20</sup> Dani 1986: 53.

<sup>21</sup> Unless the character *ṇa* was first introduced in Tamil Brāhmī; see section 6.

being exported to Central Asia and Southeast Asia. For several centuries, Indian forms of the script continued to be used in both these regions, primarily for the writing of Sanskrit texts. It was first during the Late Brāhmī period (4th to 7th centuries CE) that distinct Central Asian and Southeast Asian forms of Brāhmī developed, which then also began being used to write local languages. While the Central Asian tradition of Brāhmī came to an end with the Muslim invasions of the region at the end of the first millenium, the Southeast Asian forms of Brāhmī developed further into the modern Southeast Asian scripts. In the period of the Transitional Scripts (7th to 10th centuries CE), the Indian Northwest saw the emergence of the proto-Śāradā form of Brāhmī that became the precursor of Śāradā and other regional scripts such as Takri and Landa, which in turn inspired the development of the modern Gurmukhi script. In the rest of northern India, a style called Siddhamātrkā predominated that gave rise to the modern Devanagari and Bengali scripts. In the Deccan, a proto-Kannada-Telugu script began to take form, while further south the Grantha script was developed for the writing of Sanskrit, and the Vaṭṭeḷuttu and Tamil scripts for the writing of Tamil.








## 5. THE ENCODING OF COMMON BRĀHMĪ SCRIPT FEATURES

The Brāhmī script shares many properties with Devanagari and its other descendants, and it should therefore as far as possible follow the encoding principles developed for the modern Indian scripts and described above. In almost all varieties of Brāhmī (except for Tamil and Bhattiprolu Brāhmī; see section 6), the basic consonantal graphemes denote the consonant in combination with an inherent *a* vowel. The presence of other vowels is indicated by adding vowel diacritics to the base consonant, as illustrated below from the Gilgit-Bamiyan type I variety of Brāhmī (6th/7th century CE, Northwestern India).<sup>22</sup> As with





---

<sup>22</sup> The font is drawn from a manuscript in the Schøyen Collection (edited in Baums 2002). Here and in the following, illustrations of Brāhmī letters are only used to identify the character under discussion, not to indicate any normative rendering of







the modern scripts, vowel diacritics are encoded as combining characters following the core of the akṣara. For the code points and character names used here and in the following, refer to the code charts and names list at the end of this paper.

						
<i>ta</i>	<i>tā</i>	<i>ti</i>	<i>tī</i>	<i>tu</i>	<i>tū</i>	<i>tr</i>
U+1101F	U+1101F, U+11035	U+1101F, U+11036	U+1101F, U+11037	U+1101F, U+11038	U+1101F, U+11039	U+1101F, U+1103A







			
<i>te</i>	<i>tai</i>	<i>to</i>	<i>tau</i>
U+1101F, U+1103E	U+1101F, U+1103F	U+1101F, U+11040	U+1101F, U+11041

A sequence of consonants without intervening vowels is written as a consonant cluster. In parallel with the other Indian scripts, these consonant clusters are to be encoded with the help of U+1104D BRAHMI SIGN VIRAMA. Consonant ligatures are written from top left to bottom right:



					
<i>tma</i>	<i>tsa</i>	<i>tkṣa</i>	<i>dgr</i>	<i>śma</i>	<i>sthā</i>
U+1101F, U+1104D, U+11028	U+1101F, U+1104D, U+1102F	U+1101F, U+1104D, U+11010, U+1104D, U+1102E	U+11021, U+1104D, U+11012, U+1103A	U+1102D, U+1104D, U+11028	U+1102F, U+1104D, U+11020, U+11035

them. As stressed throughout this paper, the visual display of Brāhmī characters will vary considerably according to the font or fonts used in actual applications. That being said, care has been taken to base the code tables at the end of this paper on a geographically and chronologically “central” font (from Sander 1968: Tafel 9 to 20, “Gupta-Alphabete der Gruppe B, h–k (Schrifttypus II)”).

Pre- and postconsonantal *r* and postconsonantal *y* assume special reduced shapes in all but the earliest varieties of Brāhmī; the *kṣa* and *jña* ligatures, however, are in contrast to most modern Brāhmī-derived scripts usually transparent. Just like other consonant clusters, these are both to be encoded with the VIRAMA control character:

					
<i>rtu</i>	<i>tra</i>	<i>tya</i>	<i>rya</i>	<i>kṣa</i>	<i>jña</i>
U+1102A, U+1104D, U+1101F, U+11038	U+1101F, U+1104D, U+1102A	U+1101F, U+1104D, U+11029	U+1102A, U+1104D, U+11029	U+11010, U+1104D, U+1102E	U+11017, U+1104D, U+11019





When a consonant without inherent vowel cannot be written as non-final part of a ligature, such as when that consonant occurs at the end of a verse or paragraph, a visible virāma device is used. This device consists primarily of writing the vowelless consonant smaller and lower than other akṣaras, and often also of drawing a connecting line from the vowelless consonant to the preceding akṣara. Secondarily, a short horizontal line is frequently added above the vowelless consonant; it is this horizontal line that developed into the primary virāma marks of the modern Brāhmī-derived scripts.

	
<i>cat</i>	<i>jet</i>
U+11015, U+1101F, U+1104D	U+11017, U+1103E, U+1101F, U+1104D

The anusvāra sign (U+11045) is used to indicate that a vowel is nasalised (when the next syllable starts with a fricative) or that it is followed by a nasal segment (when the next syllable starts with a stop). The need for a separate encoding of candrabindu (indicating only nasalisation of a vowel) has not yet been demonstrated, but the code point following anusvāra has been left unassigned in case the



need should arise.<sup>23</sup> The visarga sign (U+11047) is used to write syllable-final voiceless [h]. The velar and labial allophones of [h], preceding voiceless velar and labial stops respectively, are sometimes written with the separate signs jihvāmūlīya and upadhmāñīya (U+11048 and U+11049); in contrast to visarga, these two signs are not combining diacritics, but behave like ordinary consonant signs, entering into ligatures with the following stop. (The third and fourth illustrations in the following table are from Gupta period manuscripts of the 4th/5th century CE.)

			
<i>tam</i>	<i>tah</i>	<i>hka</i>	<i>hpha</i>
U+1101F, U+11045	U+1101F, U+11047	U+11048, U+1104D, U+11010	U+11049, U+1104D, U+11025

The independent vowel signs *ṛ*, *ḷ* and *ī* and the dependent vowel signs *ḥ* and *ḥ* hardly ever occur in ordinary written texts, and therefore could not be illustrated in the code charts. The sounds they represent are, however, recognised by Indian systems of phonetics and therefore could in theory be written (cf. Dani 1986: 24f.).<sup>24</sup> This situation is exactly paralleled by that of the corresponding characters in Devanagari, with the exception that the encoding of the Devanagari characters had already been sanctioned by the pre-Unicode ISCII<sup>25</sup> standard and that they therefore had to be included in Unicode for compatibility reasons; in the case of Devanagari, illustrative glyphs were more readily available.

<sup>23</sup> Readers of this paper are encouraged to provide evidence for such a need, if indeed it exists, from manuscripts or inscriptions that employ a variant of Brāhmī that cannot be considered an early form of any particular one of the modern scripts.

<sup>24</sup> The contribution of illustrations of these characters from any relevant written source, including manuscript abecedaries, would be most welcome.

<sup>25</sup> Indian Script Code for Information Interchange, first published in 1988, and then in a revised form as IS (Indian Standard) 13194:1991 by the Bureau of Indian Standards in November 1991; see TUS, 2.9.

The Brāhmī script has separate number signs not only for the digits from 1 to 9, but also for the tens from 10 to 90, and for 100, 200, 300, 1000, 2000 and presumably 3000 (though the last has not yet been found attested). Numbers are written additively, with higher number signs preceding lower ones. Multiples of 100 above 300, and of 1000 above 3000, are expressed multiplicatively, with the multiplier following and forming a ligature with *100* or *1000*; I suggest that these ligatures be encoded with the control character ZERO WIDTH JOINER (U+200D).<sup>26</sup> (The first five illustrations in the following table are based on a Gupta period manuscript from the 4th/5th century CE;<sup>27</sup> the last two are from Western Kṣatrapa coin legends of the 1st to 4th centuries CE.<sup>28</sup>)

𑀓𑀭	𑀓𑀭𑀮	𑀓	𑀓𑀭	𑀓𑀭𑀮𑀮	𑀓	𑀓𑀭𑀮𑀮𑀮
<i>10 6</i> (= 16)	<i>50 1</i> (= 51)	100	200	<i>100 4</i> (= 104)	1000	<i>1000-4</i> (= 4000)
U+1105A, U+11056	U+1105E, U+11051	U+11063	U+11064	U+11063, U+11054	U+11064	U+11064, U+200D, U+22054

Later in the history of Brāhmī, a special sign for zero (U+11050) was invented, and the positional system came gradually into use.<sup>29</sup>

Seven punctuation marks should be encoded, namely single (𑀮, U+11070) and double (𑀮𑀮, U+11071) *daṇḍa*, delimiting clauses and verses; dot (𑀮̣, U+11072), double dot (𑀮̣𑀮̣, U+11073) and horizontal line (𑀮̣̣, U+11074), delimiting smaller textual units; and the crescent (𑀮̣̣̣, U+11075) and lotus (𑀮̣̣̣̣, U+11076) marks, delimiting larger textual units.<sup>30</sup> The scribes of Brāhmī manuscripts use additional devices, such

<sup>26</sup> See TUS, 389-391 for a detailed description of the function of this control character.

<sup>27</sup> Sander 1968: Tafel 20, alphabet h.

<sup>28</sup> Salomon 1998: 58 (table 2.6, "Numerical Notation in Brāhmī and Kharoṣṭhī").

<sup>29</sup> Bühler 1904: 82f.; Salomon 1998: 61-63.

<sup>30</sup> The glyphs for the Brāhmī punctuation marks are taken from Kuṣāṇa- and Gupta-period manuscripts in the Schøyen Collection.

as horizontal wavy lines and larger floral designs, to structure their texts, but these are of very disparate appearance and often their shape and presence is determined by physical features of the manuscript. Therefore they should be considered graphic elements rather than punctuation proper, comparable to vignettes in European manuscripts and prints.

## 6. OLD TAMIL BRĀHMĪ

In the 2nd century BCE, as the Brāhmī script spread southwards, speakers of Old Tamil became acquainted with it and adapted it to write their own language. The Tamil form of Brāhmī is known to us from numerous stone inscriptions recording the donation of caves to Jaina monastic communities, mostly in southern Tamil Nadu; from pottery graffiti found at Arikamedu, Kodumanal and other ancient trading sites; and from coin legends and inscriptions on objects such as seals and rings. In contrast to the Middle Indo-Aryan dialects for which Brāhmī had been originally invented and used so far, the Tamil language has word-final consonants and non-homorganic consonant clusters that needed to be represented in the writing system. In its first phase of development, Early Tamil Brāhmī (2nd century BCE to 1st century CE), two competing modifications of Brāhmī orthography were used to achieve this aim. One system, Mahadevan's (2003) TB-I, does away with the inherent vowel of Brāhmī consonant signs, using the vowel mātrā *ā* to represent both short and long [a] / [a:]; in this orthography consonant signs without the *ā* mātrā always represent the bare consonant. In the second orthographic system, Mahadevan's TB-II, the *ā* mātrā always represents long [a:], whereas vowelless consonant signs can be read either with inherent short [a] or as bare consonants, depending on the context. The element of ambiguity in both these systems (of *ā* in TB-I and of bare consonant signs in TB-II), as well as pressure to conform with standard forms of Brāhmī that had been adopted in neighbouring regions, led to a further orthographic modification in Late Tamil Brāhmī (2nd to 4th centuries CE; Mahadevan's TB-III) with the adoption of the *pulli* diacritic to unambigu-

ously mark vowelless consonants. *Puḷḷi* takes the form of a dot above or within the upper part of the akṣara. In addition to this normal virāma function, *puḷḷi* is also used with the vowels *e* and *o* in order to mark them as short, since in contrast to Sanskrit and most Middle-Indo-Aryan dialects, the Dravidian languages have short as well as long *e* and *o* phonemes. Just as in other forms of Brāhmī, short [a] is always inherent in TB-III consonant signs, and *ā* always indicates long [a:].

The orthographic peculiarities of Old Tamil Brāhmī do not concern the elements of the writing system itself, but are a matter of the conventional phonetic interpretation of those elements. The encoding of Old Tamil Brāhmī should not reflect this phonetic interpretation, but should be based on what is actually written; bare akṣaras and akṣaras with *ā* mātrā should be encoded as such, just as with the other varieties of Brāhmī and the Brāhmī-derived scripts. This is in accordance with Mahadevan 2003, who in his edition of the Old Tamil inscriptions provides first a close transliteration (corresponding to the proposed computer encoding of Old Tamil Brāhmī) and then a phonetic transcription. The following example is the second line of inscription no. 1 in Mahadevan 2003: 315, illustrating the TB-I system:






† ॐ † ॐ ॐ ॐ

ku va a ṇa ke dha ma mā ma

kuv aṇkē dhammam

The same encoding principle obtains already in the case of Devanagari as used for Hindi and of Gurmukhi script used to write Panjabi, where by conventional phonetic interpretation morpheme-final unmarked consonant signs are pronounced without the *a* vowel, although this is not reflected at the encoding level. The two functions of Late Tamil Brāhmī *puḷḷi* can be subsumed under the heading of “vowel reduction” (short to zero, and long to short), and *puḷḷi* should be encoded as U+1104D BRAHMI SIGN VIRAMA; the Brāhmī VIRAMA character can thus follow both consonant characters and the vowel characters

E and O, in contrast to the modern scripts' VIRAMA characters, used only after consonant characters to mark the absence of a vowel.

For the representation of sounds peculiar to Dravidian, the makers of Old Tamil Brāhmī added four or five new consonant signs to the original repertoire of Brāhmī:<sup>31</sup>  *l*,  *ɭ*,  *ɻ*,  *ɳ* and possibly  *ɺ*.<sup>32</sup> The second of these, *ɭ*, is a retroflex lateral, phonetically identical to the *la* that somewhat later appears in North Indian Brāhmī used for Sanskrit, and that also occurs in the Bhattiprolu inscriptions. Moreover, both the Tamil Brāhmī and the Bhattiprolu *l* appear to be graphically derived from the regular letter *la*, the former by adding a hook to the lower right of *la*, the latter by modifying a horizontally mirrored *la*; this in contrast to the north-Indian *la*, which is derived from the letter *ḍa*.<sup>33</sup> Old Tamil, Bhattiprolu and north-Indian *la* should therefore all be encoded as U+11031. Additional code points are provided for *ɭ*, *ɻ* and *ɳ* in the positions U+11080 to U+11082.



## 7. BHATTIPROLU BRĀHMĪ

Nine short inscriptions in a Middle Indo-Aryan dialect, found in a stūpa at Bhattiprolu in Andhra Pradesh and dating probably from the 2nd century BCE, show an orthographic system that seems to be based on the Tamil Brāhmī system TB-I. Apparently in order to avoid the phonetic ambiguity of the latter's *ā* mātrā (standing for either [a] or [a:]), the Bhattiprolu inscriptions introduce a separate dedicated mātrā

---

<sup>31</sup> The Tamil Brāhmī glyphs are based on Mahadevan 2003: 217 (palaeographic chart 2, "The Tamil-Brāhmī Script").

<sup>32</sup> Falk (1993: 203) assumes that the first occurrence of *ṇa* in Tamil Brāhmī pre-dates that in Kuṣāṇa Brāhmī, and takes this as an indication that the Sanskrit character *ṇa* was borrowed from South India. There are, however, only two possible occurrences of *ṇa* in the pre-Kuṣāṇa part of the Old Tamil corpus (Mahadevan 2003: 327, 348), both of them in what have been interpreted as proper names, and both of them originally read differently (as *ṇā* and *ṇa*, respectively).

<sup>33</sup> See Lüders 1912: 809f. In this case too, Falk (1993: 214) argues for a borrowing of the Kuṣāṇa sign from Tamil Brāhmī, not, however, from the letter  *l*, but rather from  *ɻ*; I consider neither the phonetic nor the graphic similarity strong enough to adopt this suggestion.

for long [a:] by adding a vertical stroke to the end of the *ā* mātrā: 𑀅.<sup>34</sup> Thus in these inscriptions, *ā* unambiguously writes [a], and 𑀅 (here transliterated as *Ā*) writes [a:]. The following illustration is line 2 of inscription V in Bühler 1894, while the reading follows Lüders 1912:



hi rā nā kĀ rā gĀ mā nī pu to bū bo  
 hiraṇakāra gāmaṇīputo būbo

Puzzlingly, the main reason for abandoning inherent [a] in Tamil Brāhmī, namely the ability to write word-final consonants and non-homorganic consonant clusters conveniently, does not apply in the case of the Bhattiprolu inscriptions, since Middle Indo-Aryan has neither of these phonetic features. This would seem to imply that the dedicated long *Ā* mātrā, too, was first introduced in a Tamil context, and that the resulting system was only later imitated in Bhattiprolu. But no such Tamil inscription has yet been discovered.

The shapes of five Bhattiprolu letters (*gha*, *ja*, *ma*, *la* and *sa*) differ to a certain degree from those seen in other varieties of Old Brāhmī; the *ma*, for instance, is upside-down. But only in the case of *gha*, which is graphically derived from the unaspirated *ga*, is there real innovation. Even *gha*, however, should be encoded as in the other varieties of Brāhmī, since its graphemic identity is not in doubt. The experimentation with letter shapes seen in Bhattiprolu and other forms of Old Brāhmī is entirely typical of early writing systems, such as the various Greek alphabets before the Athenian orthographic reform. The [ks] sound, for instance, was written X in the Western part of the Greek world (hence the Latin letter X) but Ξ in Greece itself, a situation not unlike that of the contrast between Bhattiprolu and standard Brāhmī *gha*.

<sup>34</sup> The Bhattiprolu 𑀅 glyph is based on the script table in Bühler 1894.

## 8. CENTRAL ASIAN BRĀHMĪ

The first Central Asian ethnic groups that modified Brāhmī for the writing of their own languages were the Khotanese on the Southern Silk Road and the Tocharians on the Northern (Hitch 1981; Sander 1986; Maue 1997).

The Khotanese writing system adds a diacritic double dot ̣̣ to the common Brāhmī repertoire, and shares with Uigur (on which below) the non-Indian orthographic practice of adding two vowel mātrās to a single akṣara to represent its set of falling diphthongs. Khotanese also developed an alternative analytic way of writing word-initial vowels, using initial *a* as a vowel carrier in combination with the various vowel mātrās, instead of separate initial signs for each vowel (Hitch 1981: 42-44).<sup>35</sup> In addition, Khotanese makes use of a hook-shaped diacritic sign placed below the akṣara, whose phonetic value is uncertain; this sign has not yet been included in the proposed encoding.






Tocharian (Pinault 1989: 33-36) added a set of ten new characters (the so-called *Fremdzeichen*, i.e., foreign or special signs) which differ from the corresponding regular Brāhmī characters in having as their inherent vowel not [a], but a different vowel transliterated as *ä* or by placing a line under the whole akṣara: 𑖅 *kä*, 𑖆 *tä*, 𑖇 *nä*, 𑖈 *pä*, 𑖉 *mä*, 𑖊 *rä*, 𑖋 *lā*, 𑖌 *śā*, 𑖍 *ṣā* and 𑖎 *ṣā*.<sup>36</sup> An alternative notation for the different vowel is the diacritic double dot seen already in Khotanese (̣̣, U+11095). In addition, the Tocharian script has an eleventh special sign 𑖏 *wä*.

The Uigur variant of Brāhmī (von Gabain 1950) adopted the Tocharian special signs (mostly using them word-finally with virāma), and also employs the double-dot diacritic ̣̣ to indicate high unrounded vowels. It added six further signs to write special consonants of the

<sup>35</sup> The same system had been employed by Kharoṣṭhī and is later also found (for some of the initial vowels) in the Devanagari, Gujarati and Tibetan scripts (Salomon 2003: 79f.).

<sup>36</sup> The glyphs for the Tocharian and Uigur special signs follow Sander 1968: Tafel 41, "Fremdzeichen".

Uigur language:  $\text{𐰚}$  *qa*,  $\text{𐰚}$  *ya*,  $\text{𐰚}$  *ḍa*,  $\text{𐰚}$  *dza*,  $\text{𐰚}$  *za* and  $\text{𐰚}$  *ža*.<sup>37</sup> The Uigur short vowels *ä*, *ü* and *ö* are spelled *-ya-*, *-yu-* and *-yo-* postconsonantly. The long vowels *ā*, *ū* and *ō* are written like the short ones, but with the addition of an *ā* mātrā (U+11035), so that in the case of *ū* and *ō* the akṣara carries not one but two vowel mātrās (<U+11038, U+11035> and <U+11040, U+11035> respectively). The initial vowels *ā*, *ū*, *ō* and *ṛ* are written by adding *-ya-*, *-yu-*, *-ya-* and *-yo-* respectively directly to the initial vowel signs *a* or *e*, *u*, *o* and *o*; therefore the resulting complexes *aya-* or *eya-*, *uyu-*, *oya-* and *oyo-* are single akṣaras that should be encoded with the control character U+200D ZERO WIDTH JOINER (cf. the discussion of Brāhmī numerals in section 5) between the initial vowel character and the *-y-*.<sup>38</sup>

				
aya (= <i>ā</i> )	eya (= <i>ā</i> )	uyu (= <i>ū</i> )	oya (= <i>ō</i> )	oyo (= <i>ṛ</i> )
U+11000, U+200D, U+11029	U+1100A, U+200D, U+11029	U+11004, U+200D, U+11029, U+11038	U+1100C, U+200D, U+11029	U+1100C, U+200D, U+11029, U+11040

Tumshuqese, closely related to Khotanese, employs a large number of special signs; scholarly discussion of the precise inventory has focused on the following manuscript sign list, written on the verso of a Tocharian alphabet table (*dvādaśākṣari*) and containing twelve entries (Konow 1935, 1947; Hitch 1981: 60-76; Maue 2004):

<sup>37</sup> Maue 1997: 3 argues that  $\text{𐰚}$  *dza* was actually pronounced [β], and the discussion in Maue 2004: 209 (on the Tumshuqese sign no. 4) seems to imply a retroflex articulation [ʒ] also for Uigur  $\text{𐰚}$  *ža*.

<sup>38</sup> Alternatively and on analogy with the postconsonantal vowels, these initial-vowel complexes could be encoded with U+1104D BRAHMĠ SIGN VIRAMA; this would however be contrary to the usual “vowel reduction” function of the VIRAMA control character (cf. the discussion of Tamil Brāhmī *puḷḷi* above), as pointed out by Gautam Sengupta of Jadavpur University, Calcutta, at the GICAS symposium in Tokyo in December 2003.





At least five of these signs (𑖀 *za*, 𑖁 *ya*, 𑖂 *za*, 𑖃 *da* and 𑖄 *dza*) are shared with Uigur and therefore do not need to be encoded separately (their codepoints are U+1109D, U+1109A, U+1109E, U+1109B and U+1109C).<sup>39</sup> Three other signs (nos. 3, 8 and 9 from the left) would appear to be mere copies of signs no. 2, 4 and 7 (𑖁 = 𑖁 *ya*, 𑖂 = 𑖂 *za* and 𑖃 = 𑖃 *da*), and are according to Konow not independently attested in Tumshuqese manuscripts.<sup>40</sup> The status of signs no. 5, 6 and 11 (𑖅, 𑖆 and 𑖇) is disputed. Hitch 1981: 67-77 interpreted them as *la*, *khu* and *śu* instead of Konow's (1947) *zya*, *za* and *gwa*; Hitch 1989 and Maue 2004 argue that no. 10 represents a voiced palatal fricative [j]. Because of the remaining uncertainty, these signs are not yet included in the set of proposed Brāhmī characters. Sign no. 12 (𑖈) is however generally agreed to be a genuine special character with the value *χša*, and accordingly is included at codepoint U+110A2.

The Central Asian varieties of Brāhmī share a ligature *rra* that does not occur in Indian Brāhmī. Although *rra* tends to be treated as a unit in Khotanese, probably representing a phoneme of that language distinct from the one written *ra*, it should be encoded as the ligature that orthographically it is.

<sup>39</sup> The Tumshuqese glyphs used in the text of this paragraph are from the manuscript sign list published in Konow 1935 and reproduced above.

<sup>40</sup> See, however, Skjærvø 1987 and Maue 2004: 210.

## 9. NOTE ON VEDIC CHARACTERS

Accent marks and other characters peculiar to Vedic texts have been excluded from the proposal described here because they are no more closely associated with pre-modern Brāhmī than with the modern Brāhmī-derived scripts, and in fact the oldest extant Vedic manuscripts lack any accent marking. A code block for Vedic characters, combinable both with Brāhmī and with the modern scripts, should become the subject of a separate proposal by a specialist in Vedic codicology.

## 10. EXPECTED USE OF THE PROPOSED CHARACTER CODING

It is anticipated that the main initial use of the proposed encoding of Brāhmī as part of the Unicode standard will be in the area of palaeographic work. Most of the fonts produced for the purpose of palaeographic discussion will aim to reproduce a particular manuscript hand or epigraphic ductus as closely as possible. Every akṣara occurring in the source is expected to be assigned a single comprehensive glyph in such fonts, and the use of glyphs for subparts of akṣaras will thus be minimal. The main operation to be performed at the rendering level will therefore be the mapping from a sequence of character code points to one particular akṣara glyph, not the relative positioning of subparts of akṣaras as is the case with most fonts for the modern Indian scripts. This greatly reduces the complexity of the rendering software that will need to be written for the display of Brāhmī.

Most fonts produced for specialised palaeographic purposes will not contain glyphs for every single Brāhmī codepoint, but only for those characters that are to be discussed or that occur in the manuscript hand to be described. Beyond such specialised fonts, however, the production and distribution of comprehensive fall-back fonts for the main subvarieties of Brāhmī is desirable. These fonts will contain normalised glyph shapes, and in their case the use of combining glyphs for subparts of akṣaras is feasible. The unitary nature of the

proposed encoding, covering a range of visually distinctive forms of the script, has the consequence that a single fall-back font cannot be appropriate for every Brāhmī text.<sup>41</sup> Nonetheless, it is felt that the advantages of a unitary encoding, as outlined in section 3, far outweigh this one inconvenience.<sup>42</sup>

The strongest case for a separate encoding of a subvariety of Brāhmī would have been presented by Tamil Brāhmī, due to the systemic characteristics that distinguish it from other forms of Brāhmī (apart from the Bhattiprolu variant). In fact, however, it adds a strong theoretical argument in favour of a unitary encoding; as has been shown above, the only possible way to encode the three subvarieties of Tamil Brāhmī (TB-I, TB-II and TB-III) uniformly and naturally is to regard the Tamil Brāhmī orthographic system as a matter of phonetic interpretation, not of character coding. Any special encoding for this orthography would have separated TB-I and TB-II from TB-III, obscuring the historical development which after a period of experimentation reintegrated the Tamil variety into the mainstream of Brāhmī script history.

As a result of accepting a unitary encoding for premodern Brāhmī, there will inevitably arise borderline cases involving late documents from the transitional period between the varieties of Brāhmī that are covered by the present encoding proposal and the modern Indian scripts that are already encoded separately in the Unicode standard. In such cases it will be left to the users' discretion in each case whether their documents are most naturally encoded as Brāhmī or as a suitable modern Brāhmī-derived script; the question is analogous to the linguistic problem of when exactly to start regarding texts to be written in New Indo-Aryan instead of Middle Indo-Aryan. In practice, the set of characters provided by the Brāhmī range and by the modern-script ranges respectively will guide the user's decision. A medieval

---

<sup>41</sup> The display of, for instance, a 3rd-century Mahāyāna sūtra with an Aśokan Brāhmī font or of an Aśokan edict with a Gilgit-Bamiyan type II font would be visually quite jarring.

<sup>42</sup> The participants in the GICAS symposium in Tokyo shared this assessment and strongly approved of a unitary encoding.

Sri Lankan text, for instance, which uses the special Sinhalese vowel *ā* could not be encoded as Brāhmī, since the proposed encoding does not provide a code point for this vowel, and would have to be encoded as Sinhalese script using the Unicode Sinhala code range, which does provide a code point for *ā*.<sup>43</sup> It is worth pointing out yet again that this kind of decision-making problem would be far worse if the historical varieties of Brāhmī were encoded in a non-unitary manner.

## REFERENCES

- BAUMS, Stefan 2002. Jyotiṣkāvadāna. In: Jens Braarvig (ed.), *Buddhist Manuscripts, Vol. II* (Manuscripts in the Schøyen Collection): 287-302. Oslo: Hermes Publishing.
- BRAARVIG, Jens (ed.) 2000. *Buddhist Manuscripts*. (Manuscripts in the Schøyen Collection.) Oslo: Hermes Publishing.
- BÜHLER, G. 1894. The Bhattiprolu inscriptions. *Epigraphia Indica* 2: 323-329.
- 1896. *Indische Palaeographie von circa 350 a. Chr. – circa 1300 p. Chr.* (Grundriss der indo-arischen Philologie und Altertums:unde, I. Band, 11. Heft.) Strassburg: Verlag von Karl J. Trübner.
- 1904. *Indian Paleography*. Bombay: Bombay Education Society's Press, Byculla.<sup>44</sup>
- DANI, Ahmad Hasan 1986. *Indian Palaeography*. Second edition. New Delhi: Munshiram Manoharlal Publishers.
- EVERSON, Michael 1998. Proposal to encode Brahmi in Plane 1 of ISO/IEC 10646-2. <http://std.dkuug.dk/JTC1/SC2/WG2/docs/n1685/n1685.htm>.
- 2002. Leaks in the Unicode pipeline: script script script... <http://www.unicode.org/notes/tn4/>.<sup>45</sup>
- FALK, Harry 1993. *Schrift im alten Indien: ein Forschungsbericht mit Anmerkungen*. (ScriptOralia, 56.) Tübingen: Gunter Narr Verlag.
- HITCH, Douglas A. 1981. *Central Asian Brahmi Palaeography: The Relationships Among the Tocharian, Khotanese, and Old Turkic Gupta scripts*. MA thesis, Department of Linguistics, University of Calgary.
- 1989. Brāhmī. In: Ehsan Yarshater (ed.), *Encyclopædia Iranica, Vol. IV*: 432-433. London: Routledge & Kegan Paul.

<sup>43</sup> U+0D87 for the initial, U+0DD0 for the vowel diacritic.

<sup>44</sup> Translation of Bühler 1896, edited as an appendix to *Indian Antiquary* 33 by John Faithfull Fleet.
























<sup>45</sup> This is Unicode Technical Note #4, presented at the 21st International Unicode Conference, May 2002, Dublin.

- KONOW, Sten 1935. Ein neuer Saka-Dialekt. *Sitzungsberichte der Preußischen Akademie der Wissenschaften, philosophisch-historische Klasse*: 772-823.
- 1947. The oldest dialect of Khotanese Saka. *Norsk Tidsskrift for Sprogvidenskap* 14: 156-190.
- LÜDERS, Heinrich 1912. Epigraphische Beiträge. *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften*: 806-831.
- MAHADEVAN, Irvatham 2003. *Early Tamil Epigraphy: From the Earliest Times to the Sixth Century A.D.* (Harvard Oriental Series, 62.) Chennai: Cre-A:.
- MAUE, Dieter 1997. A tentative stemma of the varieties of Brāhmī script along the Northern Silk Road. In: Shirin Akiner and Nicholas Sims-Williams (eds.), *Languages and Scripts of Central Asia*: 1-15. London: School of Oriental and African Studies.
- 2004. Konows Zeichen Nr. 10. In: Desmond Durkin-Meisterernst et al. (eds.), *Turfan Revisited – The First Century of Research into the Arts and Cultures of the Silk Road* (Monographien zur indischen Archäologie, Kunst und Philologie, 17): 208-212. Berlin: Dietrich Reimer Verlag.
- PINAULT, Georges-Jean 1989. Introduction au tokharien. *LALIES : actes des sessions de linguistique et de littérature* 7: 5-224.
- SALOMON, Richard 1998. *Indian Epigraphy: A Guide to the Study of Inscriptions in Sanskrit, Prakrit, and the Other Indo-Aryan Languages*. (South Asia Research.) New York: Oxford University Press.
- 2003. Writing systems of the Indo-Aryan languages. In: George Cardona and Dhanesh Jain (eds.), *The Indo-Aryan Languages* (Routledge Language Family Series): 67-103. London: Routledge.
- SANDER, Lore 1968. *Paläographisches zu den Sanskrithandschriften der Berliner Turfansammlung*. (Verzeichnis der orientalischen Handschriften in Deutschland, Supplementband 8.) Wiesbaden: Franz Steiner Verlag.
- 1986. Brāhmī scripts on the Eastern Silk Roads. *Studien zur Indologie und Iranistik* 11/12: 159-192.
- 2000. A brief palaeographical analysis of the Brāhmī manuscripts in volume I. In: Jens Braarvig (ed.), *Buddhist Manuscripts*. (Manuscripts in the Schøyen Collection): 285-300. Oslo: Hermes Publishing.
- SIRCAR, D.C. 1970-1971. Introduction to Indian epigraphy and palaeography. *Journal of Ancient Indian History* 4: 72-136.
- SKJÆRVØ, P.O. 1987. On the Tumshuqese karmavācānā text. *Journal of the Royal Asiatic Society*: 77-90.
- TUS = *The Unicode Standard: version 4.0*. Edited by Joan Aliprand et al. Boston: Addison-Wesley, 2003.
- VON GABAIN, A. 1950. *Altürkische Grammatik: mit Bibliographie, Lesestücken und Wörterverzeichnis, auch Neutürkisch*. 2. verbesserte Auflage. (Porta linguarum Orientalium: Sammlung von Lehrbüchern für das Studium der orientalischen Sprachen, 23.) Leipzig: Otto Harrassowitz.

## APPENDIX 1:

## Proposed Unicode code charts for Brāhmī

	1100	1101	1102	1103	1104	1105	1106	1107
0	𑀀	𑀁	𑀂	𑀃	𑀄	𑀅	𑀆	𑀇
	11000	11010	11020	11030	11040	11050	11060	11070
1	𑀈	𑀉	𑀊	𑀋	𑀌	𑀍	𑀎	𑀏
	11001	11011	11021	11031	11041	11051	11061	11071
2	𑀐	𑀑	𑀒			𑀓	𑀔	𑀕
	11002	11012	11022			11052	11062	11072
3	𑀖	𑀗	𑀘			𑀙	𑀚	𑀛
	11003	11013	11023			11053	11063	11073
4	𑀜	𑀝	𑀞			𑀟	𑀠	𑀡
	11004	11014	11024			11054	11064	11074
5	𑀢	𑀣	𑀤	𑀥	𑀦	𑀧	𑀨	𑀩
	11005	11015	11025	11035	11045	11055	11065	11075
6	𑀪	𑀫	𑀬	𑀭		𑀮		𑀯
	11006	11016	11026	11036		11056		11076
7	(𑀰)	𑀱	𑀲	𑀳	𑀴	𑀵	𑀶	
	11007	11017	11027	11037	11047	11057	11067	
8	(𑀷)	𑀸	𑀹	𑀺	𑀻	𑀼	𑀽	
	11008	11018	11028	11038	11048	11058	11068	
9	(𑀾)	𑀿	𑁀	𑁁	𑁂	𑁃		
	11009	11019	11029	11039	11049	11059		
A	𑁄	𑁅	𑁆	𑁇		𑁈		
	1100A	1101A	1102A	1103A		1105A		
B	𑁉	𑁊	𑁋	𑁌		𑁍		
	1100B	1101B	1102B	1103B		1105B		
C	𑁎	𑁏	𑁐	(-𑁑)		𑁒		
	1100C	1101C	1102C	1103C		1105C		
D	𑁓	𑁔	𑁕	(-𑁖)	𑁗	𑁘		
	1100D	1101D	1102D	1103D	1104D	1105D		
E		𑁙	𑁚	𑁛		𑁜		
		1101E	1102E	1103E		1105E		
F		𑁝	𑁞	𑁟		𑁠		
		1101F	1102F	1103F		1105F		

	1108	1109	110A	110B	110C	110D	110E	110F
0	 11080	 11090						
1	 11081	 11091						
2	 11082	 11092	 110A2					
3		 11093						
4		 11094						
5		 11095						
6	 11086							
7								
8								
9		 11099						
A	 1108A	 1109A						
B	 1108B	 1109B						
C	 1108C	 1109C						
D	 1108D	 1109D						
E	 1108E	 1109E						
F	 1108F							

**APPENDIX 2:****Proposed Unicode character names list for Brāhmī**

The format of the names list follows that used in TUS. The first column gives the proposed code point of the character, the second an illustrative glyph, the third the proposed Unicode name, and the fourth the usual Indological transliteration.

**Independent vowel signs**

11000	𑀅	BRAHMI LETTER A	<i>a</i>
11001	𑀆	BRAHMI LETTER AA	<i>ā</i>
11002	𑀇	BRAHMI LETTER I	<i>i</i>
11003	𑀈	BRAHMI LETTER II	<i>ī</i>
11004	𑀉	BRAHMI LETTER U	<i>u</i>
11005	𑀊	BRAHMI LETTER UU	<i>ū</i>
11006	𑀋	BRAHMI LETTER VOCALIC R	<i>ṛ</i>
11007	(𑀌)	BRAHMI LETTER VOCALIC RR	<i>ṝ</i>
11008	(𑀍)	BRAHMI LETTER VOCALIC L	<i>ḷ</i>
11009	(𑀎)	BRAHMI LETTER VOCALIC LL	<i>ḹ</i>
1100A	𑀏	BRAHMI LETTER E	<i>e</i>
1100B	𑀐	BRAHMI LETTER AI	<i>ai</i>
1100C	𑀑	BRAHMI LETTER O	<i>o</i>
1100D	𑀒	BRAHMI LETTER AU	<i>au</i>
1100E		<reserved>	
1100F		<reserved>	

**Consonants**

11010	𑀓	BRAHMI LETTER KA	<i>ka</i>
11011	𑀔	BRAHMI LETTER KHA	<i>kha</i>
11012	𑀕	BRAHMI LETTER GA	<i>ga</i>
11013	𑀖	BRAHMI LETTER GHA	<i>gha</i>
11014	𑀗	BRAHMI LETTER NGA	<i>ṇa</i>
11015	𑀘	BRAHMI LETTER CA	<i>ca</i>



11016	𑀲	BRAHMI LETTER CHA	<i>cha</i>
11017	𑀳	BRAHMI LETTER JA	<i>ja</i>
11018	𑀴	BRAHMI LETTER JHA	<i>jha</i>
11019	𑀵	BRAHMI LETTER NYA	<i>ña</i>
1101A	𑀶	BRAHMI LETTER TTA	<i>ṭa</i>
1101B	𑀷	BRAHMI LETTER TTIIA	<i>ṭha</i>
1101C	𑀸	BRAHMI LETTER DDA	<i>ḍa</i>
1101D	𑀹	BRAHMI LETTER DDHA	<i>ḍha</i>
1101E	𑀺	BRAHMI LETTER NNA	<i>ṇa</i>
1101F	𑀻	BRAHMI LETTER TA	<i>ta</i>
11020	𑀼	BRAHMI LETTER THA	<i>tha</i>
11021	𑀽	BRAHMI LETTER DA	<i>da</i>
11022	𑀾	BRAHMI LETTER DHA	<i>dha</i>
11023	𑀿	BRAHMI LETTER NA	<i>na</i>
11024	𑁀	BRAHMI LETTER PA	<i>pa</i>
11025	𑁁	BRAHMI LETTER PHA	<i>pha</i>
11026	𑁂	BRAHMI LETTER BA	<i>ba</i>
11027	𑁃	BRAHMI LETTER BHA	<i>bha</i>
11028	𑁄	BRAHMI LETTER MA	<i>ma</i>
11029	𑁅	BRAHMI LETTER YA	<i>ya</i>
1102A	𑁆	BRAHMI LETTER RA	<i>ra</i>
1102B	𑁇	BRAHMI LETTER LA	<i>la</i>
1102C	𑁈	BRAHMI LETTER VA	<i>va</i>
1102D	𑁉	BRAHMI LETTER SHA	<i>śa</i>
1102E	𑁊	BRAHMI LETTER SSA	<i>ṣa</i>
1102F	𑁋	BRAHMI LETTER SA	<i>sa</i>
11030	𑁌	BRAHMI LETTER HA	<i>ha</i>
11031	𑁍	BRAHMI LETTER LLA	<i>ḷa</i>
11032		<reserved>	
11033		<reserved>	
11034		<reserved>	



11054	𑀤	BRAHMI DIGIT FOUR	4
11055	𑀥	BRAHMI DIGIT FIVE	5
11056	𑀦	BRAHMI DIGIT SIX	6
11057	𑀧	BRAHMI DIGIT SEVEN	7
11058	𑀨	BRAHMI DIGIT EIGHT	8
11059	𑀩	BRAHMI DIGIT NINE	9
1105A	𑀪	BRAHMI NUMBER TEN	10
1105B	𑀫	BRAHMI NUMBER TWENTY	20
1105C	𑀬	BRAHMI NUMBER THIRTY	30
1105D	𑀭	BRAHMI NUMBER FOURTY	40
1105E	𑀮	BRAHMI NUMBER FIFTY	50
1105F	𑀯	BRAHMI NUMBER SIXTY	60
11060	𑀰	BRAHMI NUMBER SEVENTY	70
11061	𑀱	BRAHMI NUMBER EIGHTY	80
11062	𑀲	BRAHMI NUMBER NINETY	90
11063	𑀳	BRAHMI NUMBER ONE HUNDRED	100
11064	𑀴	BRAHMI NUMBER TWO HUNDRED	200
11065	𑀵	BRAHMI NUMBER THREE HUNDRED	300
11066		<reserved>	
11067	𑀶	BRAHMI NUMBER ONE THOUSAND	1000
11068	𑀷	BRAHMI NUMBER TWO THOUSAND	2000
11069		<reserved>	
1106A		<reserved>	
1106B		<reserved>	
1106C		<reserved>	
1106D		<reserved>	
1106E		<reserved>	
1106F		<reserved>	

**Punctuation**

11070	𑀸	BRAHMI DANDA	
11071	𑀹	BRAHMI DOUBLE DANDA	

11072	•	BRAHMI PUNCTUATION DOT	•
11073	∴	BRAHMI PUNCTUATION DOUBLE DOT	:
11074	~	BRAHMI PUNCTUATION LINE	—
11075	☾	BRAHMI PUNCTUATION CRESCENT BAR	☾
11076	☸	BRAHMI PUNCTUATION LOTUS	☸
11077		<reserved>	
11078		<reserved>	
11079		<reserved>	
1107A		<reserved>	
1107B		<reserved>	
1107C		<reserved>	
1107D		<reserved>	
1107E		<reserved>	
1107F		<reserved>	

### Tamil Brāhmī signs

11080	𑌀	BRAHMI LETTER TAMIL LLLA	<i>la</i>
11081	𑌁	BRAHMI LETTER TAMIL RRA	<i>ra</i>
11082	𑌂	BRAHMI LETTER TAMIL NNA	<i>na</i>
11083		<reserved>	
11084		<reserved>	
11085		<reserved>	

### Bhattiprolu Brāhmī sign

11086	𑀧	BRAHMI VOWEL SIGN BHATTIPROLU AAA	<i>ā</i>
11087		<reserved>	
11088		<reserved>	
11089		<reserved>	

**Central Asian Brāhmī signs**

1108A	𑀓	BRAHMI LETTER CENTRAL ASIAN KA	<i>kā</i>
1108B	𑀔	BRAHMI LETTER CENTRAL ASIAN TA	<i>tā</i>
1108C	𑀕	BRAHMI LETTER CENTRAL ASIAN NA	<i>nā</i>
1108D	𑀖	BRAHMI LETTER CENTRAL ASIAN PA	<i>pā</i>
1108E	𑀗	BRAHMI LETTER CENTRAL ASIAN MA	<i>mā</i>
1108F	𑀘	BRAHMI LETTER CENTRAL ASIAN RA	<i>rā</i>
11090	𑀙	BRAHMI LETTER CENTRAL ASIAN LA	<i>lā</i>
11091	𑀚	BRAHMI LETTER CENTRAL ASIAN SHA	<i>śā</i>
11092	𑀛	BRAHMI LETTER CENTRAL ASIAN SSA	<i>ṣā</i>
11093	𑀜	BRAHMI LETTER CENTRAL ASIAN SA	<i>śā</i>
11094	𑀝	BRAHMI LETTER CENTRAL ASIAN WA	<i>wā</i>
11095	𑀞	BRAHMI SIGN CENTRAL ASIAN DOUBLE DOT	<i>ä</i>
11096		<reserved>	
11097		<reserved>	
11098		<reserved>	
11099	𑀟	BRAHMI LETTER CENTRAL ASIAN QA	<i>qā</i>
1109A	𑀠	BRAHMI LETTER CENTRAL ASIAN GA	<i>gā</i>
1109B	𑀡	BRAHMI LETTER CENTRAL ASIAN DA	<i>ḍā/dā</i>
1109C	𑀢	BRAHMI LETTER CENTRAL ASIAN DZA	<i>dza</i>
1109D	𑀣	BRAHMI LETTER CENTRAL ASIAN ZA	<i>zā</i>
1109E	𑀤	BRAHMI LETTER CENTRAL ASIAN ZHA	<i>žā</i>
1109F		<reserved>	
110A0		<reserved>	
110A1		<reserved>	
110A2	𑀥	BRAHMI LETTER CENTRAL ASIAN KSHA	<i>χśā</i>
110A3		<reserved>	
...			
110FF		<reserved>	

# Script and Image

## Papers on Art and Epigraphy

*Edited by*

ADALBERT J. GAIL  
GERD J.R. MEVISSSEN  
RICHARD SALOMON

PAPERS OF THE 12TH WORLD SANSKRIT CONFERENCE  
HELD IN HELSINKI, FINLAND, 13-18 JULY 2003  
VOL. 11.1

General editors:  
PETTERI KOSKIKALLIO & ASKO PARPOLA

*First Editon : Delhi, 2006*

© THE AUTHORS  
All Rights Reserved

ISBN: 81-208-2944-1

**MOTILAL BANARSIDASS**

41 U.A. Bungalow Road, Jawahar Nagar, Delhi 110 007  
8 Mahalaxmi Chamber, 22 Bhulabhai Desai Road, Mumbai 400 026  
203 Royapettah High Road, Mylapore, Chennai 600 004  
236, 9th Main III Block, Jayanagar, Bangalore 560 011  
Sanas Plaza, 1302 Baji Rao Road, Pune 411 002  
8 Camac Street, Kolkata 700 017  
Ashok Rajpath, Patna 800 004  
Chowk, Varanasi 221 001

PRINTED IN INDIA

BY JAINENDRA PRAKASH JAIN AT SHRI JAINENDRA PRESS,  
A-45 NARAINA, PHASE-I, NEW DELHI 110 028  
AND PUBLISHED BY NARENDRA PRAKASH JAIN FOR  
MOTILAL BANARSIDASS PUBLISHERS PRIVATE LIMITED,  
BUNGALOW ROAD, DELHI 110 007