ISO/IEC JTC 1/SC 2/WG 2

PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS

FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646[1]

Please fill all the sections A, B and C below.

(Please read Principles and Procedures Document for guidelines and details before filling this form.)

See http://www.dkuug.dk/JTC1/SC2/WG2/docs/summaryform.html for latest *Form*.

See http://www.dkuug.dk/JTC1/SC2/WG2/docs/principles.html for latest *Principles and Procedures* document.

See http://www.dkuug.dk/JTC1/SC2/WG2/docs/roadmaps.html for latest roadmaps.

A. Administrative

1. **Title: Proposal for the Encoding of Brāhmī in Plane 1 of ISO/IEC 10646.**
2. Requesters' names: **Stefan Baums, Andrew Glass.**
3. Requester type (Member body/Liaison/Individual contribution): **Liaison contribution.**
4. Submission date: **9 October 2007.**
5. Requester's reference (if applicable): **N/A.**
6. This is a complete proposal. **Yes.**

B. Technical - General

1. This proposal is for a new script (set of characters).
   Proposed name of script: **Brāhmī.**
2. Number of characters in proposal: **121.**
3. Proposed category (see section II, Character Categories): **C.**
4. Proposed Level of Implementation (1, 2 or 3) (see clause 14, ISO/IEC 10646-1: 2000): **3.**
   Is a rationale provided for the choice? **Yes.**
   If Yes, reference: **Combining marks are used.**
5. Is a repertoire including character names provided? **Yes.**
   a. If Yes, are the names in accordance with the 'character naming guidelines in Annex L of ISO/IEC 10646-1: 2000? **Yes.**
   b. Are the character shapes attached in a legible form suitable for review? **Yes.**
6. Who will provide the appropriate computerized font (ordered preference: True Type, or PostScript format) for publishing the standard? **Andrew Glass.**
   If available now, identify source(s) for the font (include address, e-mail, ftp-site, etc.) and indicate the tools used: **Department of Asian Languages and Literature, University of Washington, Box 353521, Seattle, WA 98195-3521, USA, asg@u.washington.edu. Photoshop, Fontlab.**
7. References:
   a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided? **Yes.**
   b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached? **Yes.**
8. Special encoding issues: Does the proposal address other aspects of character data processing (if applicable) such as input, presentation, sorting, searching, indexing, transliteration etc. (if yes please enclose information)? **Yes.**
9. Additional Information:

Submitters are invited to provide any additional information about Properties of the proposed Character(s) or Script that will assist in correct understanding of and correct linguistic processing of the proposed character(s) or script. Examples of such properties are: Casing information, Numeric information, Currency information, Display behavior information such as line breaks, widths etc., Combining behavior, Spacing behavior, Directional behavior, Default Collation behavior, relevance in Mark Up contexts, Compatibility equivalence and other Unicode normalization related information. See the Unicode standard at http://www.unicode.org for such information on other scripts. Also see http://www.unicode.org/Public/UNIDATA/UnicodeCharacterDatabase.html and associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for inclusion in the Unicode Standard.

[1] Form number: N2352-F (Original 1994-10-14; Revised 1995-01, 1995-04, 1996-04, 1996-08, 1999-03, 2001-05, 2001-09).

C. Technical - Justification

1. Has this proposal for addition of character(s) been submitted before? **No.**
2. Has contact been made to members of the user community (for example: National Body, user groups of the script or characters, other experts, etc.)? **Yes.**
   If YES, with whom? Richard Salomon, Lore Sander, Jost Gippert, Gudrun Melzer.
   If YES, available relevant documents:
3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included? **Indologists.**
4. The context of use for the proposed characters (type of use; common or rare): **Scholarly; Rare.**
5. Are the proposed characters in current use by the user community? **Yes.**
   If Yes, where? **Reference: Paleographic, epigraphic and philological studies; text editions.**
6. After giving due considerations to the principles in *Principles and Procedures document* (a WG 2 standing document) must the proposed characters be entirely in the BMP? **No.**
7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)? **Yes.**
8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence? **No.**
9. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters? **No.**
10. Can any of the proposed character(s) be considered to be similar (in appearance or function) to an existing character? **Yes.**
    If Yes, is a rationale for its inclusion provided? **Yes.**
    If Yes, reference: **See below.**
11. Does the proposal include use of combining characters and/or use of composite sequences (see clauses 4.12 and 4.14 in ISO/IEC 10646-1: 2000)? **Yes.**
    If Yes, is a rationale for such use provided? **Yes.**
    If Yes, reference: **See below.**
    Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided? **Yes.**
    If Yes, reference: **See below.**
12. Does the proposal contain characters with any special properties such as control function or similar semantics? **Yes.**
    If Yes, describe in detail (include attachment if necessary). **11046 BRAHMI SIGN VIRAMA, see below.**
13. Does the proposal contain any Ideographic compatibility character(s)? **No.**

Proposal for the Encoding of Brāhmī in Plane 1 of ISO/IEC 10646

1. A common encoding for Brāhmī

In spite of superficial historical and regional variation in the form of letters and their combinations, the members of the pre-modern Brāhmī script family agree very closely in character repertoire and systemic principles. The variation that does exist is of a gradual nature that would make it a very difficult and rather arbitrary task to break the Brāhmī script continuum into subvarieties. While in the study of Brāhmī palaeography, questions of subclassification and variation do need to be discussed, we are convinced that in digital form this variation is most suitably represented at the font level, not at the encoding level.

The history of Brāhmī may be compared to the history of the Latin script. Clear subvarieties of the Latin script are well established in palaeographical studies (see Bischoff 1990), such as the Anglo-Saxon, Visigothic, Beneventan, and Caroline minuscule scripts. Further, like Brāhmī, these subvarieties were used to write a variety of regional languages as well as the lingua franca, in this case Latin as opposed to Sanskrit. These subvarieties of the Latin script are not encoded separately—to do so would present unnecessary barriers to humanistic scholars. In the case of Brāhmī, the subvarieties are less well established and continue to be revised. Setting in stone a particular set of subvarieties by encoding them separately would greatly hinder rather than help palaeographical study of the Brāhmī script.

It must also be kept in mind that in premodern India there was to a very large extent no natural connection between script varieties on the one hand and languages and their texts on the other: any given script variety would typically be used for the writing down of texts in multiple languages (such as Sanskrit and one or more regional languages), and any given text would be written in different parts of India in the respective regional scripts. Therefore the scholarly community of Indologists—the main potential users of a Brāhmī character coding—typically have to handle manuscript and epigraphical material in a multitude of script varieties in their investigation of a single text or group of texts. An artificially non-unified encoding of the written source material for this sort of study would greatly complicate searching and general data-processing.

2. Overview of the History of Brāhmī

The earliest examples of writing from historical India are the edicts of Emperor Aśoka from the third century BCE. Most of his inscriptions are in the Brāhmī writing system, but in the Indian northwest Kharoṣṭhī, Aramaic and Greek are 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 much better representation of Indian speech; 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 documents of their kind, on the basis of an acquaintance with Kharoṣṭhī and maybe also with the Aramaic or Greek scripts. The name ‘Brāhmī’ has been applied to this script family by modern scholars and is taken

from the list of scripts that the young Buddha is claimed to have mastered in the *Lalitavistara*; the first script on this long list is called *brāhmī* and said to be written from left to right, while the second is called *kharoṣṭhī* and said to be written from right to left.

The further historical development of the Brāhmī script is characterized by gradual changes in the forms of letters conditioned by cursivization and modification of stroke order, and by changes in the writing utensils used. The characteristic headmarks of the modern Devanāgarī and Bengali scripts, for instance, have their origin in the onset mark left where a reed pen first touches the writing surface, and the trend towards round letter forms in the southern varieties of Brāhmī is attributed to the southern technique of incising letters into palm leaves, where straight lines would have tended to split the leaf.

One widely used system of paleographical subclassification is that developed by A. H. Dani, distinguishing Old, Middle, and Late Brāhmī periods, Transitional Scripts, and the modern Indian scripts. In southern India in the Old Brāhmī period (third to first centuries BCE), the script was subject to experimental and rather short-lived systemic innovations attested in the Old Tamil and Bhattiprolu inscriptions (see below). In the Middle Brāhmī period (first to third centuries CE), regional variation increased; Dani distinguishes between Mathurā, Kauśāmbī, Western Deccan and Eastern Deccan styles. During this period Brāhmī was used for the first time to represent Sanskrit, and for this purpose four new letters were added to the script (𑀋 / 𑀾 *ṛ*, 𑀏 / 𑁆𑀏 *au*, 𑁂𑀃 *ḥ*, and 𑀦 *ṅa*). A special device was introduced for the marking of vowelless consonants, used both for Sanskrit and Tamil. In Sanskrit, this sign is called *virāma,* and is first attested in manuscripts of the first century CE. In Tamil, it is called *puḷḷi* and is attested in inscriptions from the second century CE (Mahadevan 2003, p. 198). In the course of trade relations and cultural exchange, the Brāhmī script was 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 during the Late Brāhmī period (fourth to seventh centuries CE) that the distinct Central Asian and Southeast Asian forms of Brāhmī developed, which then also began to be used for the writing of 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 millennium, the Southeast Asian forms of Brāhmī developed further into the modern Southeast Asian scripts. In the period of the Transitional Scripts (seventh to tenth 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 Siddhamātṛkā predominated that gave rise to the modern Devanāgarī and Bengali scripts. In the Deccan, a proto-Kannada-Telugu script began to take shape, while further south the Grantha script developed for the writing of Sanskrit, and the Vaṭṭeḻuttu and Tamil scripts for the writing of Tamil.

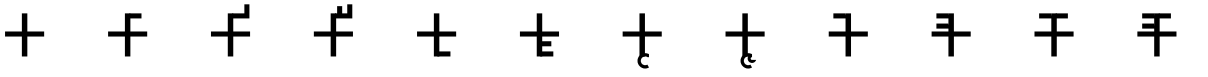
This proposal provides an encoding for the Old, Middle, and Late Brāhmī periods as defined above. It is intended, and suitable for encoding documents and citations from documents written in Brāhmī from the time of Aśoka until the seventh century of the Common Era, including the Old Tamil and Bhattiprolu inscriptions, and documents from Central Asia written in Sanskrit, Khotanese, Tocharian, Uigur, and Tumshuqese. Unless otherwise specified, illustrations in this proposal are given using glyph shapes based on a

variety of Late Brāhmī called Gilgit-Bamiyan type I, as this type covers many of the code points identified in this proposal.

3. General properties of the Brāhmī script

3.1. Dependent Vowel Signs

The Brāhmī script shares many properties with Devanāgarī and its other descendants. Lines are usually written from left to right and pages filled from top to bottom. In almost all varieties of Brāhmī (but see below on Tamil and Bhattiprolu Brāhmī), the basic consonant 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:



ka	*kā*	*ki*	*kī*	*ku*	*kū*	*kṛ*	*kṝ*	*ke*	*kai*	*ko*	*kau*
11010	11010, 11033	11010, 11034	11010, 11035	11010, 11036	11010, 11037	11010, 11038	11010, 11039	11010, 1103C	11010, 1103D	11010 1103E	11010, 1103F

The independent vowel signs r̥̄, l̥, and l̥̄ and the dependent vowel signs l̥ and l̥̄ hardly ever occur in ordinary written texts, no examples could be found on which to base examples for the code charts. The sounds they represent are, however, recognized by the indigenous Indian systems of grammar, and therefore could in theory be written (cf. Dani 1986: 24f.). Therefore this proposal reserves space for these signs in case they need to be added at some point in the future.

3.2. Consonant Ligatures

A sequence of consonants without intervening vowels is written as a consonant ligature. As with the other Indic scripts, these consonant ligatures are to be encoded with the help of 11046 BRAMI SIGN VIRAMA. It is to be noted that up to a very late date, Brāhmī used vertical conjuncts exclusively; there is thus no parallel series of 'half-consonants' as in Devanāgarī and other modern scripts. Consonant ligatures are written from top-left to bottom-right:

tma	*tsa*	*tkṣa*	*dgṛ*	*śma*	*sthā*
1101F, 11046, 11028	1101F, 11046, 1102F	1101F, 11046, 11010, 11046, 1102E	11021, 11046, 11012, 11038	1102D, 11046, 11028	1102F, 11046, 11020, 11033

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 often transparent:

rtu	*tra*	*tya*	*rya*	*kṣa*	*jña*
1102A, 11046, 1101F, 11036	1101F, 11046, 1102A	1101F, 11046, 11029	1102A, 11046, 11029	11010, 11046, 1102E	11017, 11046, 11019

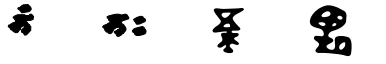
3.3. Vowel Cancellation

When a consonant without an inherent vowel cannot be written as a 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 consonants, 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 visible *virāma* marks of the modern Brāhmī-derived scripts. This device is not used in the Old Brāhmī period as consonants do not occur in final position in the Prakrits documented in this period. The encoding should follow the akṣara model common to the other Indian scripts, where the *virāma* is also used to join consonants into conjunct signs. As such, this device will is required for all periods of the Brāhmī script.

jet
11017, 1103C, 1101F, 11046

3.4. Vowel Modifiers

The *anusvāra* sign (11040) 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) could not yet be demonstrated, but the codepoint following *anusvāra* has been left unassigned in case the need should arise. The *visarga* sign (11042) is used to write syllable-final voiceless [h]. The velar and labial allophones of [h], followed by voiceless velar and labial stops respectively, are sometimes written with the separate signs *jihvāmūlīya* and *upadhmānīya* (11043 and 11044); 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 a Gupta dynasty manuscript of the fourth/fifth c. CE.)



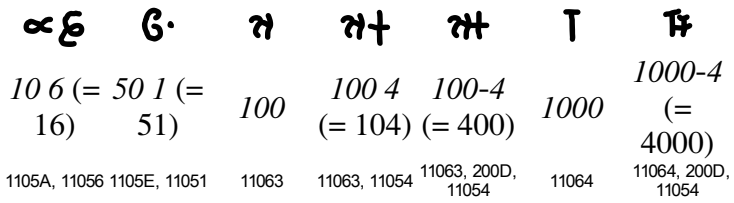
taṃ 1101F, 11040 | *taḥ* 1101F, 11042 | *ḫka* 11043, 11046, 11010 | *ḫpha* 11044, 11046, 11025

3.5. Number Signs

Two sets of numbers, used for different numbering systems are attested in Brāhmī documents. The first set is the old additive/multiplicative system that goes back to the very beginning of the Brāhmī script. The second is a set of decimal numbers that occurs side by side with the earlier numbering system in manuscripts and inscriptions during the late Brāhmī period.

The set of additive/multiplicative numbers of the Brāhmī script contains separate numbers signs for the digits from 1 to 9, the decades from 10 to 90, as well as signs for 100 and 1000. Numbers are written additively, with higher number signs preceding lower ones. Multiples of 100 and of 1000 are expressed multiplicatively, with the multiplier following and forming a ligature with *100* or *1000*; we suggest that these ligatures be encoded with ZERO WIDTH JOINER (200D). There are examples from the

Middle and Later Brāhmī periods in which the signs for 200, 300, and 2000 appear in special forms and are not obviously connected with a ligature of the component parts. Separate code points have been assigned for these numbers.



10 6 (= 16)	*50 1* (= 51)	*100*	*100 4* (= 104)	*100-4* (= 400)	*1000*	*1000-4* (= 4000)
1105A, 11056	1105E, 11051	11063	11063, 11054	11063, 200D, 11054	11064	11064, 200D, 11054

Later in the history of Brāhmī, a special sign for zero was invented, and the positional system came into use. This system is believed to be the ancestor of the modern decimal number system. Due to the different systemic features and different shapes for the signs in this set (see Melzer 2006: 64–68), we feel that a separate encoding is necessary. The features of this system should be the same as for the modern Indian number signs.

0	*1*	*2*	*3*	*5*	*10*	*248*
11060	11061	11062	11063	11064	11061, 11060	11062, 11064, 11068

3.6. Punctuation Signs

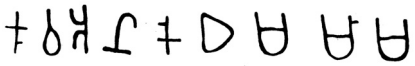
Seven punctuation marks should be encoded, namely single (𑁍, 1106A) and double (𑁎, 1106B) *daṇḍa*, delimiting clauses and verses; dot (𑁏, 1106C), double dot (𑁐, 1106D) and horizontal line (𑁑, 1106E), delimiting smaller textual units; and the crescent with a bar through it (𑁒, 1106F) and lotus (𑁓, 11070) marks, delimiting larger textual units. The shape of the single daṇḍa varies among the inscriptions and manuscripts; sometimes appearing as little more than a dot, sometimes it is a horizontal line, and sometimes it is a vertical line. Due to this variation in shape we feel it is appropriate to assign a dedicated code point (11070) rather than using the Devanagari daṇḍa (0964). The scribes of Brāhmī manuscripts use additional devices, such 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 graphical elements rather than punctuation proper, comparable to vignettes in European manuscripts and prints.

4. Tamil Brāhmī

In the second century BCE, as Brāhmī spread southwards, speakers of Old Tamil became acquainted with it and adapted it to the writing of their own language. The Tamil form of Brāhmī is known to us from a number of inscriptions donating 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 that needed to be represented in the writing system. In its first phase of development (Early Tamil Brāhmī, second century BCE – first century CE), two competing modifications of Brāhmī orthography were used to achieve this aim. The one

system (Mahadevan 2003's '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:]; consonant signs without *mātrā* always represent the bare consonant in this orthography. 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 regular forms of Brāhmī that had been adopted in neighboring regions, led to a further orthographic modification (Late Tamil Brāhmī, second – fourth centuries CE, Mahadevan's 'TB-III') with the adoption of the *puḷḷi* diacritic to unambiguously mark vowelless consonants. *Puḷḷi* takes the form of a dot above or in 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: 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 means 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 these elements. The encoding of Old Tamil Brāhmī should not reflect this phonetic interpretation, but should be based on what is actually written; bare *akṣara*s and *akṣara*s with *ā mātrā* should be encoded as such, just as in other varieties of Brāhmī. 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, on p. 315, illustrating the TB-I system):



ku va a n̲a ke dha ma mā ma

kuv aṉ̲kē dhammam

A similar encoding principle obtains already in the case of Devanāgarī as used for Hindi and of the Gurmukhi script, where by conventional phonetic interpretation morpheme-final bare *akṣara*s are pronounced vowellessly without this being 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 11046 BRAHMI SIGN VIRAMA; the Brāhmī *virāma* character can thus follow both consonant characters and the vowel characters *e* and *o*, in contrast to the modern scripts' *virāma* characters.

For the representation of sounds particular to Dravidian, the makers of Old Tamil Brāhmī added four new consonant signs to the repertoire of Brāhmī: 𑀴 *l̲*, 𑀮 *ḷ*, 𑀶 *r̲* and 𑀷 *n̲*. The second of these, *ḷ*, is phonetically identical (a retroflex lateral) to the *ḷ* that somewhat later appears in north-Indian Brāhmī for the writing of Sanskrit, and that also occurs in the Bhattiprolu inscriptions. Moreover, both the Tamil Brāhmī and the Bhattiprolu *ḷ* are graphically derived from the regular letter *l*, the former by adding a hook to the lower right of *l*, the latter by mirroring *l* horizontally (while the north-Indian *ḷ* is derived from the letter *ḍ*). Old Tamil, Bhattiprolu and north-Indian *ḷ* should therefore all be encoded as

11031. Additional code points are provided for l̲, r̲ and n̲ in the positions 11072 to 11074.

5. Bhattiprolu Brāhmī

Ten short Middle Indo-Aryan inscriptions from the second century BCE, found in a stūpa at Bhattiprolu in Andhra Pradesh, show an orthography that seems to be derived from the Tamil Brāhmī system TB-I. To avoid the phonetic ambiguity of the latter's *ā mātrā* (standing for either [a] or [a:]), the Bhattiprolu inscriptions introduce a separate *mātrā* for long [a:] by adding a vertical stroke to the end of the *ā mātrā*: ◌𑀸. Thus in these inscriptions, *ā* unambiguously means [a], and ◌𑀸 (here transliterated as *Ā*) means [a:]. (The following illustration is line 2 of inscription V in Bühler 1894; the reading follows Lüders 1912.)

hi rā ṇā kĀ rā	gĀ mā ṇī pu to	bū bo
hiraṇakāra	*gāmaṇīputo*	*būbo*

Puzzlingly, the main reason for abandoning inherent [a], namely the ability to write word-final consonants does not apply in the case of the Bhattiprolu inscriptions since Middle Indo-Aryan has neither of these phonetic features. This makes it likely 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. However, no such Tamil inscription has been discovered yet.

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 other varieties of Brāhmī as its graphemic identity is not in doubt. The experimentation with letter shapes that we see in Bhattiprolu and other 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 and Ξ in Greece itself, a situation not unlike that of Bhattiprolu and regular *gha*.

6. Central Asian Brāhmī

The first Central Asian people to have modified Brāhmī for the writing of their own language were the Khotanese on the Southern Silk Road and the Tocharians on the Northern (Hitch 1981, Sander 1986, Maue 1997).

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.

6.1. Khotanese Brāhmī

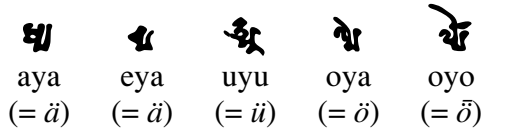
The Khotanese writing system adds the diacritic double dot ◌̈ (11083) to the common Brāhmī repertoire, and shares with Uighur (see below) the un-Indian orthographic practice of adding two vowel *mātrā*s to a single *akṣara* for the writing of its set of falling diphthongs. Khotanese also developed an alternative analytic way of writing word-initial vowels, using not the dedicated initial signs for all vowels, but just initial *a* as vowel bearer in combination with the various vowel *mātrā*s (see Hitch 1981: 42—44). The same system was used for Kharoṣṭhī, and later (for some of their initial vowels) for Gujarati, Devanāgarī, and Tibetan. In addition, Khotanese makes use of a diacritic sign with the shape of a hook below the *akṣara* and of uncertain phonetic value; this sign has not yet been included in the proposed encoding pending further research.

6.2. Tocharian Brāhmī

The Tocharians (Pinault 1989: 33–36) added a set of 10 new characters (the so-called *Fremdzeichen*, i.e., foreign or special signs) that differ from the corresponding regular Brāhmī characters by having inherent not an [a] sound, but a modified vowel [ə] or similar, transliterated as *ä* or by a line under the whole akṣara: 𑀓 *ka̲*, 𑀝 *ta̲*, 𑀦 *na̲*, 𑀧 *pa̲*, 𑀫 *ma̲*, 𑀭 *ra̲*, 𑀮 *la̲*, 𑀰 *śa̲*, 𑀲 *ṣa̲*, 𑀲 *sa̲*. An alternative notation for [ə] after these and other consonants is a diacritic double dot (◌̈). In addition, the Tocharian script has an eleventh special sign 𑀯 *wa*.

6.3. Uighur Brāhmī

Uighur Brāhmī (von Gabain 1950) adopted the Tocharian special signs in their word-final use 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 Uighur language: 𑀑 *qa*, 𑀒 *γa*, 𑀣 *δa*, 𑀜 *dza*, 𑀵 *za*, and 𑀱 *źa*. (Maue 1997: 3 argues that 𑀜 *dza* was actually pronounced [β], and Maue 2004: 209, on the Tumshuqese sign no. 4, implies a retroflex articulation [ʐ] also for Uighur 𑀱 *źa*.) The Uighur short vowels *ä*, *ü* and *ö* are spelled *-ya-*, *-yu-* and *-yo-* postconsonantally. The long vowels *ā̈*, *ṻ* and *ō̈* are written like the corresponding short vowels but with the addition of an *ā mātrā* (11033) to the same *akṣara*, which means that in the case of *ṻ* and *ō̈*, the *akṣara* carries not one but two vowel *mātrā*s (11036, 11033 and 1103E, 11033, respectively). The initial vowels *ä*, *ü*, *ö* and *ō̈* are written by adding *-ya-*, *-yu-*, *-ya-* and *-yo-* directly to the initial vowel signs *a* or *e*, *u*, *o* and *o*; this means that the resulting complexes *aya-*, *eya-*, *uyu-*, *oya-* and *oyo-* are single *akṣara*s that should, on analogy with the postconsonantal vowels, be encoded with the control character 11046 BRAHMI SIGN VIRAMA between the initial vowel character and the *-y-*:



aya (= *ä*)	eya (= *ä*)	uyu (= *ü*)	oya (= *ö*)	oyo (= *ō̈*)
11000, 11046, 11029	1100A, 11046, 11029	11004, 11046, 11029, 11036	1100C, 11046, 11029	1100C, 11046, 11029, 1103E

6.4. Tumshuqese Brāhmī

Tumshuqese is closely related to Khotanese and employs a large number of special signs. Scholarly discussion of the precise inventory of these signs has focused on the following manuscript sign list, The following sign list is written on the recto of a Tocharian alphabet table (Staatsbibliothek zu Berlin, T III M 58, published in Konow 1935). It contains twelve items (Konow 1935, 1947; Hitch 1981: 60–76; Maue 2004).

At least five of these signs (𑀵 *za*, 𑀬 *γa*, 𑀰 *źa*, 𑀝 *ḏa* and 𑀡 *dza*) are shared with Uighur and do not need to be encoded separately (their codepoints are 1108A, 11087, 1108B, 11088 and 11089). Three other signs (nos. 3, 8 and 9 from the left) appear to be mere copies of signs no. 2, 4 and 7 (𑀭 = 𑀬 *γa*, 𑀵 = 𑀰 *źa* and 𑀞 = 𑀝 *ḏa*), and are, according to Konow 1947, not independently attested in Tumshuqese manuscripts. The status of signs no. 5, 6 and 10 (𑀲, 𑀶 and 𑀴) is disputed. Hitch (1981: 67–77) interpreted them as *la*, *khu* and *śu* instead of Konow's *źya*, *ẓa* and *gwa* (1947); Hitch (1989) and Maue (2004) argue that no. 10 represents a voiced palatal fricative [ʝ]. Because of the remaining uncertainty, they are not yet included in the present proposal. Sign no. 12 (𑀴), however, is generally agreed to be a genuine special character with the value *χša*; it is included at codepoint 1108E.

7. Implementation and Usage

It is anticipated that initially the main use of the Unicode Brāhmī encoding will be in the area of scholarly paleographical work. Most of the fonts produced in this area of study will aim to reproduce a particular epigraphic ductus as closely as possible. Every occurring *akṣara* instance (consonant-vowel-diacritic combination) will be assigned a single glyph in the font, and the use of combining vowel-sign glyphs and the like will be minimal. The main operation to be performed at the rendering level will therefore be the substitution of a sequence of character code points by one particular *akṣara* glyph, not the relative positioning of subparts of *akṣara*s as with modern Indic scripts. Most fonts produced for paleographic purposes will not contain glyphs for every Brāhmī codepoint, and will usually not be applied to texts much different from the inscriptions they are based on.

Ultimately, however, the production and distribution of comprehensive fall-back fonts for the main varieties of Brāhmī is desirable. These fonts will contain normalised glyph shapes, and in their case the use of combining glyphs for subparts of *akṣara*s is feasible. As with the other scripts included in the Unicode Standard, the memory representation of strings will follow their phonetic order. For most *akṣara*s in most varieties of Brāhmī, no display reordering such as for Devanāgarī *i* will be required, because the dependent vowel sign for *i* had not yet descended from its original position on top of the base

consonant. Exceptions do, however, occur even in one and the same script, cf. Gilgit-Bamiyan type I  *dhi* with  *ti*; and in the medieval South Indian forms of Brāhmī, the *e* and *ai mātrā*s are regularly written on the left side of the *akṣara*. These exceptional cases might still be handled with substitution routines such as the GSUB OpenType feature rather than by reordering glyphs.

It has been our aim to present a unified proposal for all pre-modern forms of Brāhmī, for the reasons set out at the beginning of this document. Looking back, possibly the strongest case for a separate encoding of a Brāhmī variety would have been Tamil Brāhmī due to the systemic characteristics that distinguish it from other forms of Brāhmī. As has been shown, however, the only 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 that after a period of experimentation reintegrates the Tamil variety into the mainstream of Brāhmī script history. The other varieties of Brāhmī diverge far less from the original model, and to unify their encoding should be even less controversial.

We strongly suggest that all historical documents written in a variety of Brāhmī be encoded following the codepoints and principles set out in this document. Additional characters that may become necessary for the encoding of future discoveries of Brāhmī texts can easily be added to the code range; no major additions are, however, expected.

It remains up to the user's discretion whether in individual cases his or her documents are most naturally encoded using the Brāhmī code range or the code range of one of the modern Brāhmī-derived scripts, an issue similar to the linguistic dilemma of when exactly to start regarding texts as written in a New Indo-Aryan language rather than a Middle Indo-Aryan one. (It is worth pointing out again that this problem of decision would be exacerbated manifold if the historical varieties of Brāhmī were encoded in a non-unified manner.) In practice, the set of characters provided respectively by the Brāhmī range and by the modern-script ranges will have an influence on the user's decision. For example, an early Sri Lankan text containing the special Sinhalese vowel *ä* could not be encoded as Brāhmī, since the present proposal does not contain a codepoint for this vowel, but only as Sinhalese using the Unicode Sinhala code range (0D80 to 0DFF). It is part of our responsibility to make this sort of delimitation imposed by the contents of the Brāhmī code range coincide as closely as possible with the boundaries suggested by linguistic and other scholarly criteria.

8. Sorting

Alphabetically ordered word lists (such as dictionaries) in the Brāhmī script are not preserved and maybe never existed. We do, however, know the traditional way of arranging the letters of the Brāhmī script from ancient abecedaries (*varṇamālā*s or *dvādaśākṣarī*s) which are based on phonetic principles. The sort order of the modern Indian scripts, as well as of Indological transliteration, is based on the *varṇamālā* order, but varies in some details. The conjuncts *kṣ* and *jñ*, for instance, are considered so basic that they are included in their own right at the end of the ancient *varṇamālā*s; this is not

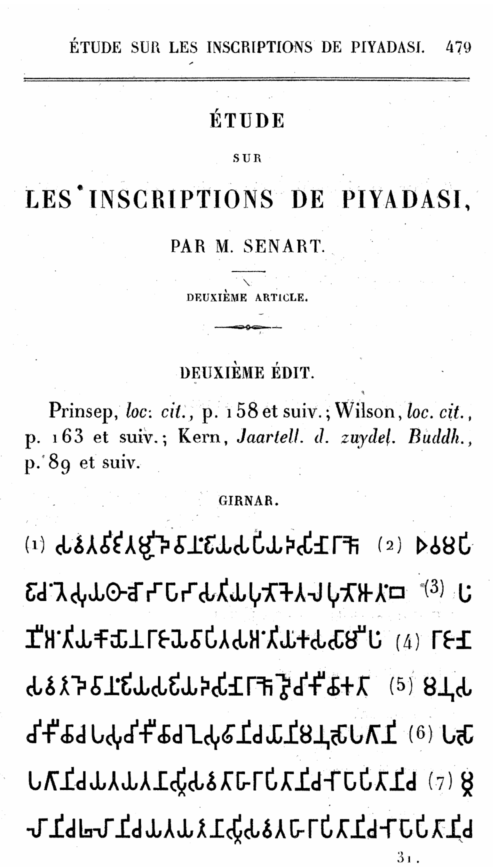
imitated in modern usage.

It is most practical to specify the Brāhmī sort order in terms of an ordered list of Indological transliteration units, where some transliteration units correspond to a single Brāhmī Unicode character (e.g., *ḥ* = 11042 BRAHMI SIGN VISARGA); some to a particular sequence of Brāhmī Unicode characters (e.g., *k* = 11010 BRAHMI LETTER KA + 11046 BRAHMI SIGN VIRAMA); and some to either one of two alternate Brāhmī Unicode characters (e.g., *o* = either 1100C BRAHMI LETTER O or 1103E BRAHMI VOWEL SIGN O). Please compare the descriptions of the individual writing systems above, and the transliterations given in the right-hand column of the character name list below. Note that when *ṃ* is immediately followed by a stop, it is pronounced and sorted like the nasal consonant homorganic with that stop: like *ṅ* when followed by *k*, *kh*, *g*, *gh*, or *ṅ*; like *ñ* when followed by *c, ch, j, jh,* or *ñ*; like *ṇ* when followed by *ṭ, ṭh, ḍ, ḍh,* or *ṇ*; like *n* when followed by *t, th, d, dh,* or *n*; like *m* when followed by *p, ph, b, bh,* or *m*.

Brāhmī sort order: *a*, *ā*, *i*, *ī*, *u*, *ū*, *e*, *ai*, *o*, *au, ṃ*, *k*, *kh*, *g*, *gh*, *ṅ*, *c*, *ch*, *j*, *jh*, *ñ*, *ṭ*, *ṭh*, *ḍ*, *ḍh*, *ṇ*, *t*, *th*, *d*, *dh*, *n*, *p*, *ph*, *b*, *bh*, *m*, *y*, *r*, *l*, *v*, *w*, *ś*, *ṣ*, *s*, *h*, *l̲*, *l̲̄*, *r̲*, *n̲*.

Appendix 1: Samples of Brāhmī texts

Illustration 1: Minor Rock Edict of Aśoka at Brahmagiri (Mahadevan 2003, p. 174).



ÉTUDE SUR LES INSCRIPTIONS DE PIYADASI. 479

ÉTUDE

SUR

LES INSCRIPTIONS DE PIYADASI,

PAR M. SENART.

DEUXIÈME ARTICLE.

DEUXIÈME ÉDIT.

Prinsep, *loc. cit.*, p. 158 et suiv.; Wilson, *loc. cit.*, p. 163 et suiv.; Kern, *Jaartell. d. zuydel. Buddh.*, p. 89 et suiv.

GIRNAR.

31.

Illustration 2: Sample of typeset Brāhmī, the Second Rock Edict of Aśoka at Girnar (Senart 1880, p. 479).

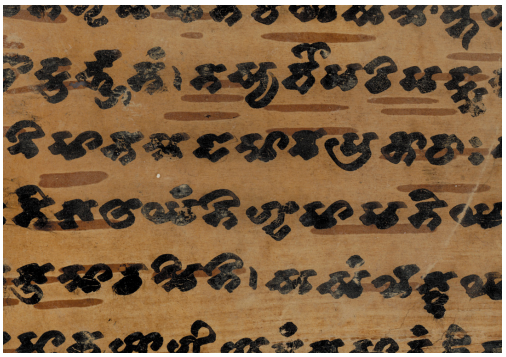


Illustration 3: Manuscript of the Jyotiṣkāvadāna in Gilgit-Bamiyan type I Brāhmī (Baums 2003, pl. XVI.1).

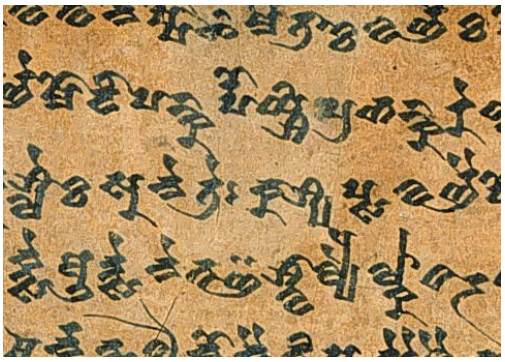


Illustration 4: Tocharian manuscript from Shorchuq (Staatsbibliothek zu Berlin).

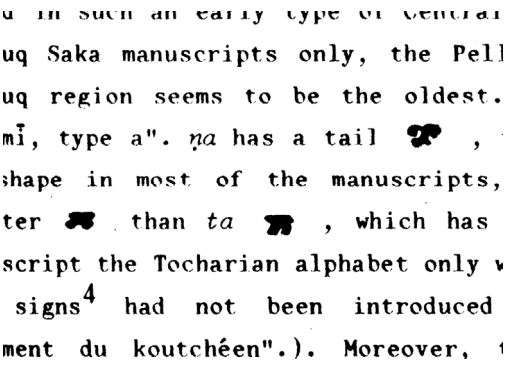


Illustration 5: Example of Brāhmī characters in modern scholarly use (Sander 1986, p. 165).

Appendix 2: Code Chart

	1100	1101	1102	1103	1104	1105	1106	1107	1108
0	11000	11010	11020	11030	11040	11050	11060	11070	11080
1	11001	11011	11021	11031		11051	11061		11081
2	11002	11012	11022		11042	11052	11062	11072	11082
3	11003	11013	11023	11033	11043	11053	11063	11073	11083
4	11004	11014	11024	11034	11044	11054	11064	11074	
5	11005	11015	11025	11035		11055	11065		
6	11006	11016	11026	11036	11046	11056	11066	11076	11086
7		11017	11027	11037		11057	11067		11087
8		11018	11028	11038	11048	11058	11068	11078	11088
9		11019	11029	11039	11049	11059	11069	11079	11089
A	1100A	1101A	1102A		1104A	1105A	1106A	1107A	1108A
B	1100B	1101B	1102B		1104B	1105B	1106B	1107B	1108B
C	1100C	1101C	1102C	1103C	1104C		1106C	1107C	
D	1100D	1101D	1102D	1103D	1104D		1106D	1107D	
E		1101E	1102E	1103E	1104E		1106E	1107E	1108E
F		1101F	1102F	1103F	1104F		1106F	1107F	

Independent vowel signs

11000	𑀀	BRAHMI LETTER A	*a*
11001	𑀁	BRAHMI LETTER AA	*ā*
11002	𑀂	BRAHMI LETTER I	*i*
11003	𑀃	BRAHMI LETTER II	*ī*
11004	𑀄	BRAHMI LETTER U	*u*
11005	𑀅	BRAHMI LETTER UU	*ū*
11006	𑀆	BRAHMI LETTER VOCALIC R	*r̥*
11007		<reserved>	
11008		<reserved>	
11009		<reserved>	
1100A	𑀇	BRAHMI LETTER E	*e*
1100B	𑀈	BRAHMI LETTER AI	*ai*
1100C	𑀉	BRAHMI LETTER O	*o*
1100D	𑀊	BRAHMI LETTER AU	*au*
1100E		<reserved>	
1100F		<reserved>	

Consonants

11010	𑀋	BRAHMI LETTER KA	*ka*
11011	𑀌	BRAHMI LETTER KHA	*kha*
11012	𑀍	BRAHMI LETTER GA	*ga*
11013	𑀎	BRAHMI LETTER GHA	*gha*
11014	𑀏	BRAHMI LETTER NGA	*ṅa*
11015	𑀐	BRAHMI LETTER CA	*ca*
11016	𑀑	BRAHMI LETTER CHA	*cha*
11017	𑀒	BRAHMI LETTER JA	*ja*
11018	𑀓	BRAHMI LETTER JHA	*jha*
11019	𑀔	BRAHMI LETTER NYA	*ña*
1101A	𑀕	BRAHMI LETTER TTA	*ṭa*
1101B	𑀖	BRAHMI LETTER TTHA	*ṭha*
1101C	𑀗	BRAHMI LETTER DDA	*ḍa*
1101D	𑀘	BRAHMI LETTER DDHA	*ḍha*
1101E	𑀙	BRAHMI LETTER NNA	*ṇa*
1101F	𑀚	BRAHMI LETTER TA	*ta*
11020	𑀛	BRAHMI LETTER THA	*tha*
11021	𑀜	BRAHMI LETTER DA	*da*
11022	𑀝	BRAHMI LETTER DHA	*dha*
11023	𑀞	BRAHMI LETTER NA	*na*
11024	𑀟	BRAHMI LETTER PA	*pa*
11025	𑀠	BRAHMI LETTER PHA	*pha*
11026	𑀡	BRAHMI LETTER BA	*ba*
11027	𑀢	BRAHMI LETTER BHA	*bha*
11028	𑀣	BRAHMI LETTER MA	*ma*
11029	𑀤	BRAHMI LETTER YA	*ya*

1102A	𑀭	BRAHMI LETTER RA	*ra*
1102B	𑀮	BRAHMI LETTER LA	*la*
1102C	𑀯	BRAHMI LETTER VA	*va*
1102D	𑀰	BRAHMI LETTER SHA	*śa*
1102E	𑀱	BRAHMI LETTER SSA	*ṣa*
1102F	𑀲	BRAHMI LETTER SA	*sa*
11030	𑀳	BRAHMI LETTER HA	*ha*
11031	𑀴	BRAHMI LETTER LLA	*ḷa*
11032		<reserved>	

Dependent vowel signs

11033	𑀵	BRAHMI VOWEL SIGN AA	*ā*
11034	𑀶	BRAHMI VOWEL SIGN I	*i*
11035	𑀷	BRAHMI VOWEL SIGN II	*ī*
11036	𑀸	BRAHMI VOWEL SIGN U	*u*
11037	𑀹	BRAHMI VOWEL SIGN UU	*ū*
11038	𑀺	BRAHMI VOWEL SIGN VOCALIC R	*ṛ*
11039	𑀻	BRAHMI VOWEL SIGN VOCALIC RR	*ṝ*
1103A		<reserved>	
1103B		<reserved>	
1103C	𑀼	BRAHMI VOWEL SIGN E	*e*
1103D	𑀽	BRAHMI VOWEL SIGN AI	*ai*
1103E	𑀾	BRAHMI VOWEL SIGN O	*o*
1103F	𑀿	BRAHMI VOWEL SIGN AU	*au*

Various signs

11040	𑁀	BRAHMI SIGN ANUSVARA	*ṃ*
11041		<reserved>	
11042	𑁁	BRAHMI SIGN VISARGA	*ḥ*
11043	𑁂	BRAHMI LETTER JIHVAMULIYA	*ḫ*
11044	𑁃	BRAHMI LETTER UPADHMANIYA	*ḯ*
11045			
11046	𑁆	BRAHMI SIGN VIRAMA	
11047		<reserved>	

Number Signs

11048	𑁒	BRAHMI NUMERAL ONE	*1*
11049	𑁓	BRAHMI NUMERAL TWO	*2*
1104A	𑁔	BRAHMI NUMERAL THREE	*3*
1104B	𑁕	BRAHMI NUMERAL FOUR	*4*
1104C	𑁖	BRAHMI NUMERAL FIVE	*5*
1104D	𑁗	BRAHMI NUMERAL SIX	*6*
1104E	𑁘	BRAHMI NUMERAL SEVEN	*7*
1104F	𑁙	BRAHMI NUMERAL EIGHT	*8*

11050	𑁐	BRAHMI NUMBER NINE	*9*
11051	𑁑	BRAHMI NUMBER TEN	*10*
11052	𑁒	BRAHMI NUMBER TWENTY	*20*
11053	𑁓	BRAHMI NUMBER THIRTY	*30*
11054	𑁔	BRAHMI NUMBER FOURTY	*40*
11055	𑁕	BRAHMI NUMBER FIFTY	*50*
11056	𑁖	BRAHMI NUMBER SIXTY	*60*
11057	𑁗	BRAHMI NUMBER SEVENTY	*70*
11058	𑁘	BRAHMI NUMBER EIGHTY	*80*
11059	𑁙	BRAHMI NUMBER NINETY	*90*
1105A	𑁚	BRAHMI NUMBER ONE HUNDRED	*100*
1105B	𑁛	BRAHMI NUMBER ONE THOUSAND	*1000*
1105C		<reserved>	
1105D		<reserved>	
1105E		<reserved>	
1105F		<reserved>	

Decimal Numbers

11060	𑁠	BRAHMI DIGIT ZERO	*0*
11061	𑁡	BRAHMI DIGIT ONE	*1*
11062	𑁢	BRAHMI DIGIT TWO	*2*
11063	𑁣	BRAHMI DIGIT THREE	*3*
11064	𑁤	BRAHMI DIGIT FOUR	*4*
11065	𑁥	BRAHMI DIGIT FIVE	*5*
11066	𑁦	BRAHMI DIGIT SIX	*6*
11067	𑁧	BRAHMI DIGIT SEVEN	*7*
11068	𑁨	BRAHMI DIGIT EIGHT	*8*
11069	𑁩	BRAHMI DIGIT NINE	*9*

Punctuation

1106A	𑁪	BRAHMI DANDA	।
1106B	𑁫	BRAHMI DOUBLE DANDA	॥
1106C	𑁬	BRAHMI PUNCTUATION DOT	·
1106D	𑁭	BRAHMI PUNCTUATION DOUBLE DOT	:
1106E	𑁮	BRAHMI PUNCTUATION LINE	–
1106F	𑁯	BRAHMI PUNCTUATION CRESCENT BAR	☪
11070	𑁰	BRAHMI PUNCTUATION LOTUS	❀
11071		<reserved>	

Tamil Brāhmī signs

11072	𑁱	BRAHMI LETTER TAMIL LLLA	*l̲a*
11073	𑁲	BRAHMI LETTER TAMIL RRA	*r̲a*
11074	𑁳	BRAHMI LETTER TAMIL NNA	*n̲a*

11075		<reserved>	

Bhattiprolu Brāhmī sign

11076	◌𑀾	BRAHMI VOWEL SIGN BHATTIPROLU AAA	*Ā*
11077		<reserved>	

Central Asian Brāhmī signs

11078	𑁸	BRAHMI LETTER CENTRAL ASIAN KA	*k̲a̲*
11079	𑁹	BRAHMI LETTER CENTRAL ASIAN TA	*t̲a̲*
1107A	𑁺	BRAHMI LETTER CENTRAL ASIAN NA	*n̲a̲*
1107B	𑁻	BRAHMI LETTER CENTRAL ASIAN PA	*p̲a̲*
1107C	𑁼	BRAHMI LETTER CENTRAL ASIAN MA	*m̲a̲*
1107D	𑁽	BRAHMI LETTER CENTRAL ASIAN RA	*r̲a̲*
1107E	𑁾	BRAHMI LETTER CENTRAL ASIAN LA	*l̲a̲*
1107F	𑁿	BRAHMI LETTER CENTRAL ASIAN SHA	*ś̲a̲*
11080	𑂀	BRAHMI LETTER CENTRAL ASIAN SSA	*s̲a̲*
11081	𑂁	BRAHMI LETTER CENTRAL ASIAN SA	*s̲a̲*
11082	𑂂	BRAHMI LETTER CENTRAL ASIAN WA	*wa*
11083	◌𑂃	BRAHMI SIGN CENTRAL ASIAN DOUBLE DOT	*ä*
11084		<reserved>	
11085		<reserved>	
11086	𑂆	BRAHMI LETTER CENTRAL ASIAN QA	*qa*
11087	𑂇	BRAHMI LETTER CENTRAL ASIAN GA	*γa*
11088	𑂈	BRAHMI LETTER CENTRAL ASIAN DA	*δa / d̲a*
11089	𑂉	BRAHMI LETTER CENTRAL ASIAN DZA	*dza*
1108A	𑂊	BRAHMI LETTER CENTRAL ASIAN ZA	*za*
1108B	𑂋	BRAHMI LETTER CENTRAL ASIAN ZHA	*ź̲a*
1108C		<reserved>	
1108D		<reserved>	
1108E	𑂎	BRAHMI LETTER CENTRAL ASIAN KHSHA	*χša*
1108F		<reserved>	

Appendix 3: Usage of Characters

11000–1100D These are independent vowel signs. They do not combine with dependent vowel signs (11033–1103F) or the Bhattiprolu Brāhmī sign (11076), but may combine with BRAHMI SIGN ANUSVARA (11040), BRAHMI SIGN VISARGA (11042), and BRAHMI SIGN CENTRAL ASIAN DOUBLE DOT (11083).

11010–11031 These are the consonant signs. All unmarked consonants include the inherent vowel *a*. Other vowels are indicated by one of the dependent vowel signs (11033–1103F). Consequently these signs may combine with the dependent vowel signs, BRAHMI VOWEL SIGN BHATTIPROLU AAA (11076), BRAHMI SIGN ANUSVARA (11040), BRAHMI SIGN VISARGA (11042), and BRAHMI SIGN CENTRAL ASIAN DOUBLE DOT (11083). These signs may be followed by BRAHMI SIGN VIRAMA (11046), see below.

11033–1103F These are the dependent vowel signs. In principle, only one may be applied to each syllable, however, multiple vowels are used in some varieties of Brāhmī, see §§ 6.1, 6.3. These signs should only combine with the Brāhmī consonants (11010–11031), the Tamil Brāhmī signs (11072–11074), and the Central Asian Brāhmī signs (11078–8E). These signs may be followed by BRAHMI SIGN ANUSVARA (11040) and the BRAHMI SIGN VISARGA (11042).

11040 This is the Brāhmī *anusvāra*, indicating either a vowel nasalization or a nasal consonant segment. The order of this glyph is thus context dependent, see § 7 Sorting. It may combine with any Brāhmī sign except BRAHMI SIGN VISARGA (11042), the Brāhmī numerals (11048–1105B) and digits (11060–11069), and the Brāhmī punctuation signs (1106A–11070).

11042 This is the Brāhmī *visarga.* It has the same combining properties as the BRAHMI SIGN ANUSVARA (11040).

11043–11044 These are the Brāhmī *jihvāmūlīya* and *upadhmānīya.* They may enter into conjuncts with other consonant characters (11010–11031, 11072–11078–11082, 11086–1108E).

11046 This is the Brāhmī *virāma*. It is used to indicate the suppression of the inherent vowel, and as a device to join consonants into conjunct signs, see § 2. During the Old Brāhmī period, it does not appear as a mark or sign in itself. In later periods is appears as a horizontal mark over a reduced sized consonant. For all periods, it should function as a control character that causes the consonant which it follows to appear as a subscript to the preceding akṣara. When followed immediately by another consonant it triggers a conjunct form representing both consonants, see § 2. It can only follow a consonant (11010–11031, 11072–11078–11082, 11086–1108E), or the BRAHMI LETTER JIHVAMULIYA (11043) and BRAHMI LETTER UPADHMANIYA (11044). The Brāhmī virāma may follow an independent vowel sign (11000–1100D) in Uighur Brāhmī, see § 6.3.

11048–1105B These are the Brāhmī numbers for the older additive/multiplicative number system, see § 2.

1105A–1105B BRAHMI NUMBER ONE HUNDRED and BRAHMI NUMBER ONE THOUSAND may be followed by the ZERO WIDTH JOINER (200D) and another

number (a multiplier). Such cases should trigger a conjunct form showing a multiple of a hundred or a thousand.

11060–11069 These are the Brāhmī digits for the decimal system. These digits function exactly like modern Arabic numerals (0030–0039).

1106A–11070 These are the Brāhmī punctuation signs. Automatic line breaks should come after these signs, not before.

11072–11074 These are the Tamil Brāhmī signs. They function just like the Brāhmī consonant signs (11010–11031).

11076 This is a Bhattiprolu Brāhmī sign. It functions just like the Brāhmī dependent vowel signs (11035–11040).

11078–11082 These are Central Asian Brāhmī signs, they function just like the Brāhmī consonant signs (11010–11031).

11083 This Central Asian Brāhmī sign may combine with full Brāhmī signs, and functions just like the BRAHMI SIGN ANUSVARA (11040) and the BRAHMI SIGN VISARGA (11042).

11086–1108E These are the additional Central Asian Brāhmī signs, they function just like the Brāhmī consonant signs (11010–11031).

Appendix 4: Word Breaks, Line Breaks and Hyphenation

Most Brāhmī inscriptions are written as continuous text with no indication of word boundaries. Line breaks may occur at word boundaries, but not always. There are no examples of anything akin to hyphenation in Brāhmī documents. In cases where a word would not completely fit into a line, its continuation simply appears at the beginning of the next line. Modern scholarly practice will in most cases make use of spaces and hyphenation. When necessary, hyphenation should be applied on the model of Sanskrit.

Acknowledgements

Work on this proposal was made possible in part by a grant from the U. S. National Endowment for the Humanities (PA-511171-05) to the Universal Scripts Project. The authors are grateful to Deborah Anderson and Rick McGowan for their comments on a draft of this proposal.

The main font used in the code tables is based on Aśokan Brāhmī. Signs that do not occur in Aśokan Brāhmī, but which are needed for this proposal have been designed with a view to blending Aśokan style with the attested forms. The Tamil Brāhmī glyphs are based on Mahadevan 2003 (palaeographic chart 2, 'The Tamil-Brāhmī script').

On 18 January 1998, Michael Everson submitted a proposal for the separate encoding of 58 Brāhmī characters occurring in the edicts of Aśoka (available at http://www.dkuug.dk/JTC1/SC2/WG2/docs/n1685/n1685.htm). We hope that our proposal, with its more comprehensive coverage of the variants of Brāhmī will be found to be a worthy successor to and replacement of his pioneering effort.

Bibliography

Baums, Stefan. 2006. "Towards a computer encoding for Brāhmī." In Adalbert J. Gail, Gerd J. R. Mevissen and Richard Salomon, eds., *Script and Image: Papers on Art and Epigraphy*. Papers of the 12th World Sanskrit Conference, vol. 11.1, pp. 111–143. Delhi: Motilal Banarsidass Publishers. [This article is based on an earlier version of this proposal which provided a comprehensive encoding for all periods of the Brāhmī script.]

Bischoff, Bernard. 1990. *Latin Palaeography: Antiquity and the Middle Ages.* Cambridge: Cambridge University Press.

Bühler, G., 1894. The Bhattiprolu inscriptions. *Epigraphia Indica: a collection of inscriptions supplementary to the Corpus Inscriptionum Indicarum of the Archæological Survey*, II, pp. 323—329.

Bühler, G., 1896. *Indische Palaeographie von circa 350 a. Chr. – circa 1300 p. Chr.* Strassburg: Verlag von Karl J. Trübner. (Grundriss der indo-arischen Philologie und Altertumskunde, I. Band, 11. Heft

Dani, Ahmad Hasan, 1986. *Indian palaeography*. Second edition. New Delhi: Munshiram Manoharlal Publishers.

Hitch, Doug, 1981. *Central Asian Brahmi palaeography: the relationships among the Tocharian, Khotanese, and Old Turkic Gupta scripts*. MA thesis, Department of Linguistics, University of Calgary.

Konow, Sten, 1935. Ein neuer Saka-Dialekt. *Sitzungsberichte der Preußischen Akademie der Wissenschaften, philosophisch-historische Klasse*, pp. 772–823.

Konow, Sten, 1947. The oldest dialect of Khotanese Saka. *Norsk tidsskrift for sprogvidenskap*, XIV, pp. 156–190.

Lüders, Heinrich, 1912. Epigraphische Beiträge. *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften*, pp. 806ff.

Mahadevan, Iravatham, 2003. *Early Tamil epigraphy: from the earliest times to the sixth century A.D.* Chennai, India: Cre-A:. (Harvard Oriental Series, volume sixty-two.)

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, London: School of Oriental and African Studies, pp. 1–15.

Maue, Dieter, 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*, Berlin: Dietrich Reimer Verlag (Monographien zur indischen Archäologie, Kunst und Philologie, Band 17), pp. 208–212.

Melzer, Gudrun, 2006. *Ein Abschnitt aus dem Dīrghāgama.* PhD dissertation, Ludwig-Maximilians-Universität, München.

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*. New York: Oxford University Press. (South Asia Research.)

Sander, Lore, 1968. *Paläographisches zu den Sanskrithandschriften der Berliner Turfansammlung*. Wiesbaden: Franz Steiner Verlag. (Verzeichnis der orientalischen Handschriften in Deutschland, Supplementband 8.)

Sander, Lore, 1986. Brāhmī scripts on the Eastern Silk Roads. *Studien zur Indologie und Iranistik*, 11/12, pp. 159–192.

Senart, E., 1880. Étude sur les inscriptions de Piyadasi. *Journal asiatique* 15: 287–357, 479–509.

Skjærvø, P. O., 1987. On the Tumshuqese karmavācanā text. *Journal of the Royal Asiatic Society*, 77–90.

von Gabain, A., 1950. *Alttürkische Grammatik: mit Bibliographie, Lesestücken und Wörterverzeichnis, auch Neutürkisch*. 2. verbesserte Auflage. Leipzig: Otto Harrassowitz. (Porta linguarum Orientalium: Sammlung von Lehrbüchern für das Studium der orientalischen Sprachen, XXIII.)