

Universal Multiple-Octet Coded Character Set  
International Organization for Standardization  
Organisation Internationale de Normalisation  
Международная организация по стандартизации

**Doc Type: Working Group Document****Title: Preliminary proposal for encoding the Sorang Sompeng script in the UCS****Source: Michael Everson****Status: Individual Contribution****Action: For consideration by JTC1/SC2/WG2 and UTC****Replaces: N1957 (1999-01-29)****Date: 2008-04-08**

This is a preliminary proposal to encode the Sorang Sompeng script in the BMP of the UCS.

**1. Introduction.** The Sorang Sompeng script is used to write the Sora language. Sora is a member of the Munda family of languages, which, together with the Mon-Khmer languages, makes up Austro-Asiatic. The Sora people live between the Oriya- and Telugu-speaking populations in what is now the Orissa-Andhra border area. Sorang Sompeng was devised on 1936-06-18 by Mangei Gomango, son-in-law of charismatic community leader Malia Gomango, on the inspiration of a vision of the 24 letters he had. The script was promulgated as part of a comprehensive cultural program, and was offered as an improvement over IPA-based scripts used by linguists and missionaries. Sorang Sompeng is used in primary and adult education, and is published in a variety of printed materials.

**2. Structure.** Sorang Sompeng is of the Brahmic type: the consonant letters contain an inherent vowel. Consonant clusters are written not with conjunct characters, and there is no visible vowel killer to show the deletion of the inherent vowel; the reader must determine the presence or absence of the inherent schwa word by word. The character repertoire does not match the phonemic repertoire of Sora very well; LETTER I is used for both [i] and [ī], and LETTER O is used for both [o] and [ō], for instance. The glottal stop is written with LETTER HAH, and the sequence of LETTER RAH and LETTER DAH is used to write retroflex [ɽ].

**3. Collating order.** The letters get their names from 24 gods in the Sora pantheon, e.g. *s* for Sundañ, *t* for Tənod; no rationale is apparent for the relative ordering, but collation order is as in the code chart.

**4. Character names.** Consonant letter names are derived by adding [aʔ] (written *ah*) to the consonant.

**5. Punctuation.** I have seen no text in Sorang Sompeng. It can be surmised that punctuation common to Latin, Telugu, and Oriya may be used.

**6. Linebreaking.** Letters and digits behave as in Latin and other alphabetic scripts.

**7. Unicode Character Properties.**

```
AB40;SORANG SOMPENG LETTER SAH;Lo;0;L;;;;N;;;;;
AB41;SORANG SOMPENG LETTER TAH;Lo;0;L;;;;N;;;;;
AB42;SORANG SOMPENG LETTER BAH;Lo;0;L;;;;N;;;;;
AB43;SORANG SOMPENG LETTER CAH;Lo;0;L;;;;N;;;;;
AB44;SORANG SOMPENG LETTER DAH;Lo;0;L;;;;N;;;;;
```

AB45;SORANG SOMPENG LETTER GAH;Lo;0;L;;;;N;;;;;  
 AB46;SORANG SOMPENG LETTER MAH;Lo;0;L;;;;N;;;;;  
 AB47;SORANG SOMPENG LETTER NGAH;Lo;0;L;;;;N;;;;;  
 AB48;SORANG SOMPENG LETTER LAH;Lo;0;L;;;;N;;;;;  
 AB49;SORANG SOMPENG LETTER NAH;Lo;0;L;;;;N;;;;;  
 AB4A;SORANG SOMPENG LETTER VAH;Lo;0;L;;;;N;;;;;  
 AB4B;SORANG SOMPENG LETTER PAH;Lo;0;L;;;;N;;;;;  
 AB4C;SORANG SOMPENG LETTER YAH;Lo;0;L;;;;N;;;;;  
 AB4D;SORANG SOMPENG LETTER RAH;Lo;0;L;;;;N;;;;;  
 AB4E;SORANG SOMPENG LETTER HAH;Lo;0;L;;;;N;;;;;  
 AB4F;SORANG SOMPENG LETTER KAH;Lo;0;L;;;;N;;;;;  
 AB50;SORANG SOMPENG LETTER JAH;Lo;0;L;;;;N;;;;;  
 AB51;SORANG SOMPENG LETTER NYAH;Lo;0;L;;;;N;;;;;  
 AB52;SORANG SOMPENG LETTER A;Lo;0;L;;;;N;;;;;  
 AB53;SORANG SOMPENG LETTER E;Lo;0;L;;;;N;;;;;  
 AB54;SORANG SOMPENG LETTER I;Lo;0;L;;;;N;;;;;  
 AB55;SORANG SOMPENG LETTER U;Lo;0;L;;;;N;;;;;  
 AB56;SORANG SOMPENG LETTER O;Lo;0;L;;;;N;;;;;  
 AB57;SORANG SOMPENG LETTER AE;Lo;0;L;;;;N;;;;;  
 AB60;SORANG SOMPENG DIGIT ZERO;Nd;0;L;0;0;0;N;;;;;  
 AB61;SORANG SOMPENG DIGIT ONE;Nd;0;L;1;1;1;N;;;;;  
 AB62;SORANG SOMPENG DIGIT TWO;Nd;0;L;2;2;2;N;;;;;  
 AB63;SORANG SOMPENG DIGIT THREE;Nd;0;L;3;3;3;N;;;;;  
 AB64;SORANG SOMPENG DIGIT FOUR;Nd;0;L;4;4;4;N;;;;;  
 AB65;SORANG SOMPENG DIGIT FIVE;Nd;0;L;5;5;5;N;;;;;  
 AB66;SORANG SOMPENG DIGIT SIX;Nd;0;L;6;6;6;N;;;;;  
 AB67;SORANG SOMPENG DIGIT SEVEN;Nd;0;L;7;7;7;N;;;;;  
 AB68;SORANG SOMPENG DIGIT EIGHT;Nd;0;L;8;8;8;N;;;;;  
 AB69;SORANG SOMPENG DIGIT NINE;Nd;0;L;9;9;9;N;;;;;

## 8. Bibliography.

Zide, Norman. 1996. "Scripts for Munda languages", in Peter T. Daniels and William Bright, eds. *The world's writing systems*. New York; Oxford: Oxford University Press. ISBN 0-19-507993-0

**9. Acknowledgements.** This project was made possible in part by a grant from the U.S. National Endowment for the Humanities, which funded the which funded the Universal Scripts Project (part of the Script Encoding Initiative at UC Berkeley) in respect of the Sorang Sompeng encoding. Any views, findings, conclusions or recommendations expressed in this publication do not necessarily reflect those of the National Endowment of the Humanities.

## A. Administrative

1. Title

**Preliminary proposal for encoding the Sorang Sompeng script in the UCS**

2. Requester's name

**Michael Everson**

3. Requester type (Member body/Liaison/Individual contribution)

**Individual contribution.**

4. Submission date

**2008-04-08**

5. Requester's reference (if applicable)

6. Choose one of the following:

6a. This is a complete proposal

**No.**

6b. More information will be provided later

**Yes.**

## B. Technical – General

1. Choose one of the following:

1a. This proposal is for a new script (set of characters)

**Yes.**

1b. Proposed name of script

**Sorang Sompeng.**

1c. The proposal is for addition of character(s) to an existing block

**No.**

1d. Name of the existing block

2. Number of characters in proposal

**34.**

3. Proposed category (A-Contemporary; B.1-Specialized (small collection); B.2-Specialized (large collection); C-Major extinct; D-Attested extinct; E-Minor extinct; F-Archaic Hieroglyphic or Ideographic; G-Obscure or questionable usage symbols)

**Category A.**

4a. Is a repertoire including character names provided?

**Yes.**

4b. If YES, are the names in accordance with the "character naming guidelines" in Annex L of P&P document?

**Yes.**

4c. Are the character shapes attached in a legible form suitable for review?

**Yes.**

5a. Who will provide the appropriate computerized font (ordered preference: True Type, or PostScript format) for publishing the standard?

**Michael Everson.**

5b. If available now, identify source(s) for the font (include address, e-mail, ftp-site, etc.) and indicate the tools used:

**Michael Everson, Fontographer.**

6a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided?

**Yes.**

6b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached?

**Yes.**

7. 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.**

8. 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 behaviour information such as line breaks, widths etc., Combining behaviour, Spacing behaviour, Directional behaviour, Default Collation behaviour, 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 Unicode Character Database <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.

**See above.**

## C. Technical – Justification

1. Has this proposal for addition of character(s) been submitted before? If YES, explain.

**Yes, in N1957.**

2a. Has contact been made to members of the user community (for example: National Body, user groups of the script or characters, other experts, etc.)?

**No.**

2b. If YES, with whom?

2c. 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?

**People living at the Orissa-Andhra border in India.**

4a. The context of use for the proposed characters (type of use; common or rare)

**Evidently common.**

4b. Reference

5a. Are the proposed characters in current use by the user community?

**Yes.**

5b. If YES, where?

**In India.**

6a. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP?

**Yes.**

6b. If YES, is a rationale provided?

**Yes.**

6c. If YES, reference

**Contemporary use and accordance with the Roadmap.**

7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)?

**Yes.**

8a. Can any of the proposed characters be considered a presentation form of an existing character or character sequence?

**No.**

8b. If YES, is a rationale for its inclusion provided?

8c. If YES, reference

9a. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters?

**No.**

9b. If YES, is a rationale for its inclusion provided?

9c. If YES, reference

10a. Can any of the proposed character(s) be considered to be similar (in appearance or function) to an existing character?

**No.**

10b. If YES, is a rationale for its inclusion provided?

10c. If YES, reference

11a. 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)?

**No.**

11b. If YES, is a rationale for such use provided?

11c. If YES, reference

11d. Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided?

**No.**

11e. If YES, reference

12a. Does the proposal contain characters with any special properties such as control function or similar semantics?



































**No.**

12b. If YES, describe in detail (include attachment if necessary)



















13a. Does the proposal contain any Ideographic compatibility character(s)?

**No.**

13b. If YES, is the equivalent corresponding unified ideographic character(s) identified?

	AB4	AB5	AB6
0	 AB40	 AB50	 AB60
1	 AB41	 AB51	 AB61
2	 AB42	 AB52	 AB62
3	 AB43	 AB53	 AB63
4	 AB44	 AB54	 AB64
5	 AB45	 AB55	 AB65
6	 AB46	 AB56	 AB66
7	 AB47	 AB57	 AB67
8	 AB48		 AB68
9	 AB49		 AB69
A	 AB4A		
B	 AB4B		
C	 AB4C		
D	 AB4D		
E	 AB4E		
F	 AB4F		

**Consonants**

- AB40  SORANG SOMPENG LETTER SAH
- AB41  SORANG SOMPENG LETTER TAH
- AB42  SORANG SOMPENG LETTER BAH
- AB43  SORANG SOMPENG LETTER CAH
- AB44  SORANG SOMPENG LETTER DAH
- AB45  SORANG SOMPENG LETTER GAH
- AB46  SORANG SOMPENG LETTER MAH
- AB47  SORANG SOMPENG LETTER NGAH
- AB48  SORANG SOMPENG LETTER LAH
- AB49  SORANG SOMPENG LETTER NAH
- AB4A  SORANG SOMPENG LETTER VAH
- AB4B  SORANG SOMPENG LETTER PAH
- AB4C  SORANG SOMPENG LETTER YAH
- AB4D  SORANG SOMPENG LETTER RAH
- AB4E  SORANG SOMPENG LETTER HAH
- AB4F  SORANG SOMPENG LETTER JAH
- AB50  SORANG SOMPENG LETTER NYAH
- AB51  SORANG SOMPENG LETTER NYAH

**Vowels**

- AB52  SORANG SOMPENG LETTER A
- AB53  SORANG SOMPENG LETTER E
- AB54  SORANG SOMPENG LETTER I
- AB55  SORANG SOMPENG LETTER U
- AB56  SORANG SOMPENG LETTER O
- AB57  SORANG SOMPENG LETTER AE

**Digits**

- AB60  SORANG SOMPENG DIGIT ZERO
- AB61  SORANG SOMPENG DIGIT ONE
- AB62  SORANG SOMPENG DIGIT TWO
- AB63  SORANG SOMPENG DIGIT THREE
- AB64  SORANG SOMPENG DIGIT FOUR
- AB65  SORANG SOMPENG DIGIT FIVE
- AB66  SORANG SOMPENG DIGIT SIX
- AB67  SORANG SOMPENG DIGIT SEVEN
- AB68  SORANG SOMPENG DIGIT EIGHT
- AB69  SORANG SOMPENG DIGIT NINE