JTC1/SC2/WG2 N3293R
L2/07-233R
2007-07-26

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 Batak script in the UCS**
Source: **Michael Everson, SEI (Universal Scripts Project)**
Status: **Individual Contribution**
Action: **For consideration by JTC1/SC2/WG2 and UTC**
Date: **2007-07-26**

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

1. Introduction. The Batak script is (or was) used to write Toba, Mandailing, Dairi, and possibly other languages on the island of Sumatra. The alphabet is called *si-sija-sija* in Toba (van der Tuuk). Batak is read from left to right, but is often written similarly to Tagalog and Buhid, by writing vertically along the length of a piece of bamboo.

2. Unification. The different language groups share most of their letters in common, though sometimes a letter with a value in one language has a different value in another. This proposal encodes the superset of forms, regardless of pronunciation.

2.1. Mandailing. The Mandailing alphabetical order differs somewhat from Toba, and North Mandailing again differs slightly from South Mandailing. Some of the letter shapes are likewise slightly different; these are *ha* and *sa*. The rendering forms for the consonant vowel-sign combinations *pa+u*, *sa+u*, and *la+u* may differ from the forms used for Toba Batak. Mandailing uses two other letters for *ka* and *ca* sounds. These two letters are produced by putting a mark called *tompi* onto the normal letters for *ha* and *sa*. It is not known whether the *tompi* is otherwise productive.

2.2. Dairi. Dairi alphabetical order also differs from Toba and Mandailing. Dairi does not include the letter *nya*. The forms for *ta* and *wa* differ significantly from those used for Toba. The vowel sign listed in the chart as *u* is pronounced more like a closed *e* and written after the associated consonant rather than under (or attached to) the consonant. The sign *sikordjan*, which is pronounced as a soft *h* following the associated vowel, is placed over the consonant. When final *ng* is used in Dairi, it goes over the previous consonant rather than over the vowel sign. In Toba, it may optionally go over the vowel if the vowel is not a non-spacing mark.

3. Structure. The Batak script is of the Brahmic type. Like Tagalog and other scripts of the archipelagos between Southeast Asia and Australia, Batak ultimately derives from scripts of India. Batak has a vowel killer called a *pangolat* used to express final consonants. There is no consonant conjunct formation. Batak has three independent vowels (A, I, U) and makes use of a number of vowel signs and two consonant signs. Van der Tuuk gives an order which may be an alphabetical order; it differs from the Brahmic order. The accompanying chart is in the order given for Toba, but it may be useful to implementors to use Brahmic order.

4. Dependent vowel signs. The dependent vowels are as follows (shown with KA):

ba	=	ba
bë	=	ba + -ë
bë	=	ba + -ë
be	=	ba + -e
bi	=	ba + -i
bi	=	ba + -i
bo	=	ba + -o
bo	=	ba + -o
bou	=	ba + -ou
bu	=	ba + -u
bang	=	ba + -ng
bah	=	ba + -h
b	=	ba + killer
b	=	ba + killer

5. Rendering. The vowel signs *i*, *i*, *o*, and the two killer *pangolat*s and are spacing marks although they are combining characters. The vowel signs *e* and final *ng* are non-spacing marks, the former to the left and the latter to the right. (When *e* and *ng* occur together on a consonant, there are two marks above: *beng*) The vowel sign *u* is placed under a consonant and somewhat to the right; it can ligate with its base consonant. [Table to be supplied.] The *hamisaran* is usually written above the vowels *i* and *o*. When *pangolat* is used to close a syllable, the vowel sign for the previous vowel is placed either under the final consonant or after the final consonant, and before the *pangolat* itself. This is a bit strange: looks like *borati*\ but it is *borit*; looks like *gako*\ but it is *gok*; looks like *sni**tak*\ but it is *sintak*.

6. Punctuation. Punctuation is not normally used, all letters simply running together, but a *bindu* does exist and is occasionally used to disambiguate similar words or phrases. (This *bindu* is unfortunately known by the same name as the killer, *pangolat*.) The *bindu* apparently appears in several forms. One is called *bindu pinardjolma* and is used to separate sections of text; another is *bindu pinarulok*, and a third is *bindu pinarboras*, again used to separate sections of text. These marks can be written as large signs that physically separate sections of text. A sign called *pustaha* (Sanskit *pustaka*) is also sometimes used to separate a title from the main text which normally begins on the same line. [There are other punctuation marks in some of the materials I have seen. Other names include *bindu godong* 'large bindu' and *bindu na metek*. 'small bindu']

7. Collating order. Alphabetical order differs somewhat amongst the different languages. All sorting elements are treated with primary weight. [Ordering may have to be syllabic, given the unusual way final consonants are handled.]

8. Character names. The character names used follow Kozok 1999. Language identifiers are used to distinguish. Usually this was SIMALUNGUN because Simalungun is the most variant – but the use of the modifier does not imply that a character is only used in Simalungun Batak; the designation is arbitrary.

9. Linebreaking. Opportunities for line-break occur after any full orthographic syllable. Batak punctuation marks can be expected to have behaviour similar to that of Devanagari DANDA.

10. Unicode Character Properties.

[To be supplied]

11. Bibliography.

Daniels, Peter T., and William Bright, eds. 1996. *The world's writing systems*. New York; Oxford: Oxford University Press. ISBN 0-19-507993-0

Faulmann, Carl. 1990 (1880). *Das Buch der Schrift*. Frankfurt am Main: Eichborn. ISBN 3-8218-1720-8

Haarmann, Harald. 1990. *Die Universalgeschichte der Schrift*. Frankfurt: Campus. ISBN 3-593-34346-0

Kozok, Uli. 1999. *Warisan leluhur: sastra lama dan aksara Batak*. Jakarta: École française d'Extrême-Orient. ISBN 979-9023-33-5

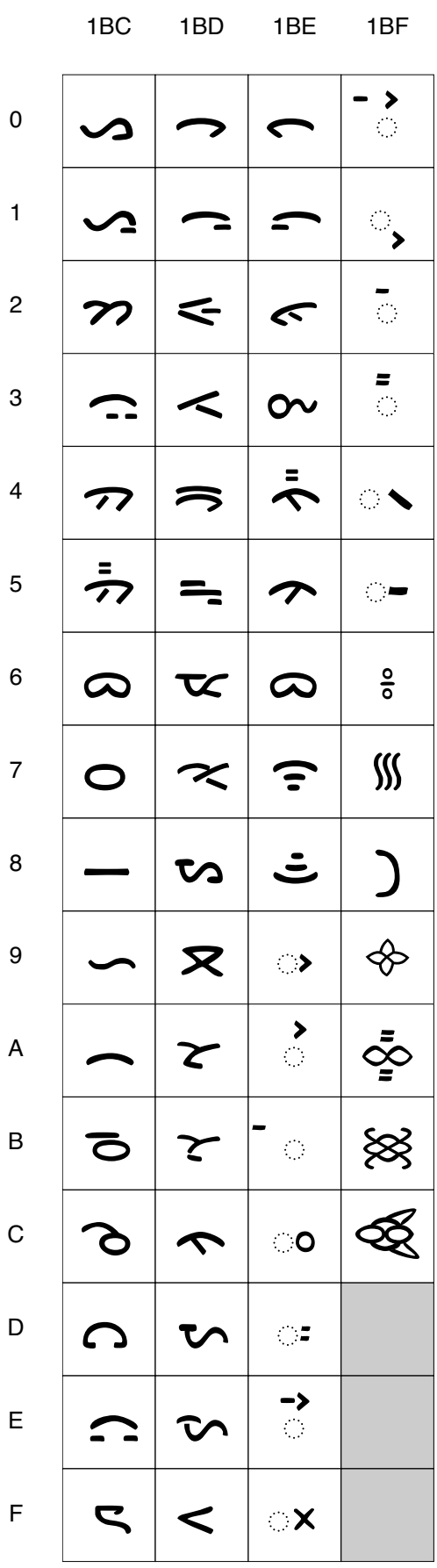
Nakanishi, Akira. 1990. *Writing systems of the world: alphabets, syllabaries, pictograms*. Rutland, VT: Charles E. Tuttle. ISBN 0-8048-1654-9

Unicode Consortium. 1992. *Unicode Technical Report #3: exploratory proposals*.

van der Tuuk, H. N. A Grammar of Toba Batak.

12. Acknowledgements. This project was made possible in part by a grant from the U.S. National Endowment for the Humanities, which funded the Script Encoding Initiative in respect of the Batak encoding.

Row 1B: BATAK DRAFT



hex	Name
C0	BATAK LETTER A
C1	BATAK LETTER SIMALUNGUN A
C2	BATAK LETTER KA
C3	BATAK LETTER SIMALUNGUN KA
C4	BATAK LETTER MANDAILING HA
C5	BATAK LETTER MANDAILING KA
C6	BATAK LETTER BA
C7	BATAK LETTER KARO BA
C8	BATAK LETTER PA
C9	BATAK LETTER SIMALUNGUN PA
CA	BATAK LETTER MANDAILING PA
CB	BATAK LETTER NA
CC	BATAK LETTER MANDAILING NA
CD	BATAK LETTER WA
CE	BATAK LETTER SIMALUNGUN WA
CF	BATAK LETTER PAKPAK WA
D0	BATAK LETTER GA
D1	BATAK LETTER SIMALUNGUN GA
D2	BATAK LETTER JA
D3	BATAK LETTER DA
D4	BATAK LETTER RA
D5	BATAK LETTER SIMALUNGUN RA
D6	BATAK LETTER MA
D7	BATAK LETTER SIMALUNGUN MA
D8	BATAK LETTER TA
D9	BATAK LETTER SIMALUNGUN TA
DA	BATAK LETTER SA
DB	BATAK LETTER SIMALUNGUN SA
DC	BATAK LETTER YA
DD	BATAK LETTER SIMALUNGUN YA
DE	BATAK LETTER CA TOBA NYA
DF	BATAK LETTER NGA
E0	BATAK LETTER LA
E1	BATAK LETTER SIMALUNGUN LA
E2	BATAK LETTER NYA
E3	BATAK LETTER CA
E4	BATAK LETTER MANDAILING CA
E5	BATAK LETTER NDA
E6	BATAK LETTER MBA
E7	BATAK LETTER I
E8	BATAK LETTER U
E9	BATAK VOWEL SIGN E
EA	BATAK VOWEL SIGN PAKPAK E
EB	BATAK VOWEL SIGN EE
EC	BATAK VOWEL SIGN I
ED	BATAK VOWEL SIGN SIMALUNGUN I
EE	BATAK VOWEL SIGN KARO O
EF	BATAK VOWEL SIGN O
F0	BATAK VOWEL SIGN OU
F1	BATAK VOWEL SIGN U
F2	BATAK VOWEL SIGN NG
F3	BATAK VOWEL SIGN H
F4	BATAK VIRAMA
F5	BATAK SIMALUNGUN VIRAMA
F6	BATAK SYMBOL
F7	BATAK SYMBOL
F8	BATAK SYMBOL
F9	BATAK SYMBOL
FA	BATAK SYMBOL
FB	BATAK SYMBOL
FC	BATAK SYMBOL
FD	(This position shall not be used)
FE	(This position shall not be used)
FF	(This position shall not be used)

A. Administrative

1. Title

Proposal for encoding the Batak script in the BMP of the UCS

2. Requester's name

Michael Everson

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

Individual contribution.

4. Submission date

2007-07-26

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

Batak.

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

59.

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.

No.

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.)?

Yes.

2b. If YES, with whom?

Ulrich Kozok

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 in Sumatra.

4a. The context of use for the proposed characters (type of use; common or rare)
Traditional use.
4b. Reference
5a. Are the proposed characters in current use by the user community?
Yes.
5b. If YES, where?
In Sumatra.
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?