JTC1/SC2/WG2 N3320R
L2/08-011R
2008-10-07

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

Doc Type: **Working Group Document**
Title: **Proposal for encoding the Batak script in the UCS**
Source: **UC Berkeley Script Encoding Initiative (Universal Scripts Project)**
Author: **Michael Everson and Uli Kozok**
Status: **Liaison Contribution**
Action: **For consideration by JTC1/SC2/WG2 and UTC**
Replaces: **UTR#3, N3293R**
Date: **2008-10-07**

1. Introduction. The Batak script is used on the island of Sumatra to write the five Batak dialects Karo, Mandailing, Pakpak, Simalungun, and Toba. (These dialects can differ as much as the related languages English and Dutch do.) The script is called *surat na sampulu sia* 'the nineteen letters', or *si-sia-sia*. Batak is read from left to right. (Descriptions of Batak writing, like those of Tagalog and Buhid, which talk about writing vertically bottom-to-top along the length of a piece of bamboo, are based on an observation of practical writing behaviour. Anyone engraving Latin script with the point of a knife on bamboo in the same way would do likewise.) The Batak script is taught in schools more for cultural purposes than as a practical writing system for Batak, which, when written, uses Latin orthography (though the overwhelming majority of writing by Bataks is in Indonesian, as elsewhere in Indonesia). Batak script does enjoy public display for instance in the signage of shops and governmental institutions.

2. Structure. The Batak script is of the Brahmic type. It has a vowel killer which is called *pangolat* in Mandailing, Pakpak, and Toba (where it has the shape ◌᳭); the Karo call the killer *pěnĕngĕn*, and the Simalungen call it *panongonan* (it has the shape ◌᳟ for those groups). Consonant conjuncts are not formed. (It is worth noting that this simplification, found also in other insular Southeast Asian scripts outside of Java and Bali, is a sensible and appropriate response to the CV(C) structure of the languages in the region, and is by no means a "corruption" of the original Brahmic prototype.) Batak has three independent vowels (A, I, U) and makes use of a number of vowel signs and two consonant signs.

3. Dependent vowel signs. The dependent vowels are as follows (shown with 𑲂 RA and 𑲒 SIMALUNGUN RA and with 𑲓 SIMALUNGUN SA for VOWEL SIGN U FOR SIMALUNGUN SA):

𑲂◌𑲹	rĕ	=	𑲂	ra	+	◌𑲹	-ĕ	𑲂◌𑲺	rĕ	=	𑲂	ra	+ ◌𑲺 -ĕ (Pakpak)
𑲂◌᳟	re	=	𑲂	ra	+	◌᳟	-e						
𑲂◌𑲸	ri	=	𑲂	ra	+	◌𑲸	-i	𑲒◌𑲸	ri	=	𑲒	ra	+ ◌𑲸 -i (Simalungun)
𑲂◌𑲻	ro	=	𑲂	ra	+	◌𑲻	-o	𑲂◌𑲼	ro	=	𑲂	ra	+ ◌𑲼 -o (Karo)
𑲂◌𑲽	ru	=	𑲂	ra	+	◌𑲽	-u	𑲓◌𑲽	su	=	𑲓	sa	+ ◌𑲽 -u (Simalungun)
𑲂◌᳞	rang	=	𑲂	ra	+	◌᳞	-ng						
𑲂◌᳠	rah	=	𑲂	ra	+	◌᳠	-h						
𑲂◌᳭	r	=	𑲂	ra	+	◌᳭	killer	𑲒◌᳟	r	=	𑲒	ra	+ ◌᳟ (Simalungun)

It should be noted that some of the vowel signs are limited to use by certain groups. Only the Karo and Pakpak have the sound *ĕ*, and use ◌ᩮ VOWEL SIGN E for it, though the Pakpak sometimes use ◌ᩯ VOWEL SIGN PAKPAK E instead. Karo writers use either the ◌ᩯ VOWEL SIGN PAKPAK E or the ◌ᩰ VOWEL SIGN KARO O for *o*; VOWEL SIGN KARO O is used by the Simulungun for *ou*. Karo writers always use ◌ᩬ VOWEL SIGN O for *u* (though the other groups use it for *o*); Karo writers may use either ◌ᩥ VOWEL SIGN I or ◌ᩦ VOWEL SIGN KARO I for *i*.

4. Rendering. The vowel signs ◌ᩥ VOWEL SIGN I, ◌ᩦ VOWEL SIGN KARO I, ◌ᩬ VOWEL SIGN O, the consonant sign ◌ᩬ CONSONANT SIGN H *h*, and the two killers ◌᩵ PANGOLAT and ◌᩶ PANONGONAN are spacing marks. The characters ◌ᩮ VOWEL SIGN EE *e* and ◌ᩱ CONSONANT SIGN NG are non-spacing marks, the former drawn to the left side of the character and the latter to the right side. (When the two occur together on a consonant, there are two marks above: ᩁᩮᩱ *reng*; ᩁ RA + ◌ᩮ VOWEL SIGN EE + ◌ᩱ CONSONANT SIGN NG.) The character ◌ᩳ VOWEL SIGN U is placed under a consonant and somewhat to the right; it can ligate with its base consonant.

ᨠᩳ	u	=	ᨠ	a	+ ◌ᩳ -u		ᨠᩲᩳ	u	=	ᨠᩲ	S a + ◌ᩳ -u
ᨡᩳ	hu	=	ᨡ	ha	+ ◌ᩳ -u		ᨡᩲᩳ	hu	=	ᨡᩲ	S ha + ◌ᩳ -u
ᨢᩳ	hu	=	ᨢ	M ha	+ ◌ᩳ -u						
ᨻᩳ	bu	=	ᨻ	ba	+ ◌ᩳ -u		ᨾᩳ	bu*	=	ᨾ	K ba + ◌ᩳ -u
ᨸᩳ	pu	=	ᨸ	pa	+ ◌ᩳ -u		ᨸᩲᩳ	pu	=	ᨸᩲ	S pa + ◌ᩳ -u
ᨶᩳ	nu	=	ᨶ	na	+ ◌ᩳ -u		ᨶᩲᩳ	nu	=	ᨶᩲ	M na + ◌ᩳ -u
ᨿᩳ	wu	=	ᨿ	wa	+ ◌ᩳ -u		ᨿᩲᩳ	wu	=	ᨿᩲ	S wa + ◌ᩳ -u
ᩅᩳ	wu	=	ᩅ	P wa	+ ◌ᩳ -u						
ᨣᩳ	gu	=	ᨣ	ga	+ ◌ᩳ -u		ᨣᩲᩳ	gu	=	ᨣᩲ	S ga + ◌ᩳ -u
ᨩᩳ	ju	=	ᨩ	ja	+ ◌ᩳ -u		ᨯᩳ	du	=	ᨯ	da + ◌ᩳ -u
ᩁᩳ	ru	=	ᩁ	ra	+ ◌ᩳ -u		ᩁᩲᩳ	ru	=	ᩁᩲ	S ra + ◌ᩳ -u
ᨾᩳ	mu	=	ᨾ	ma	+ ◌ᩳ -u		ᨾᩲᩳ	mu	=	ᨾᩲ	S ma + ◌ᩳ -u
ᨲᩳ	tu	=	ᨲ	S ta	+ ◌ᩳ -u		ᨶᩳ	tu	=	ᨶ	N ta + ◌ᩳ -u
ᨧᩳ	su	=	ᨧ	sa	+ ◌ᩳ -u		ᨧᩲᩳ	su	=	ᨧᩲ	S sa + ◌ᩳ -u (Mandailing)
ᨧᩴ	su	=	ᨧᩲ	M sa	+ ◌ᩳ -u		ᨧᩲᩴ	su	=	ᨧᩲ	S sa + ◌ᩴ -u (Simalungun)
ᨿᩳ	yu	=	ᨿ	ya	+ ◌ᩳ -u		ᨿᩳ	yu	=	ᨿ	S ya + ◌ᩳ -u
ᨦᩳ	ngu	=	ᨦ	nga	+ ◌ᩳ -u						
ᩃᩳ	lu	=	ᩃ	la	+ ◌ᩳ -u		ᩃᩲᩳ	lu	=	ᩃᩲ	S la + ◌ᩳ -u
ᨬᩳ	nyu	=	ᨬ	nya	+ ◌ᩳ -u		ᨧᩳ	cu*	=	ᨧ	ca + ◌ᩳ -u
ᨶᩳ	ndu*	=	ᨶ	nda	+ ◌ᩳ -u		ᨻᩳ	mbu*	=	ᨻ	mba + ◌ᩳ -u

Note that the forms given with asterisks above do not occur since the letters are only used in Karo, which writes ᨾᩬ *bu*, ᨧᩬ *cu*, ᨶᩬ *ndu*, and ᨻᩬ *mbu*. Note too that while Mandailing may write ᨧᩴ for *su*, in Simalungun the ◌ᩳ VOWEL SIGN U vowel is not used with this letter. Instead the diacritic ◌ᩴ VOWEL SIGN U FOR SIMALUNGUN SA is used—only with this letter: ᨧᩴ. ◌ᩮ◌ᩥ◌ᩬ◌ᩯ◌ᩳ◌ᩴ◌ᩰ◌ᩱ◌ᩲ◌᩵◌᩶◌ᩲ◌ᩯ◌ᩦ◌ᩲ

The non-spacing consonant modifier ◌ᩲ TOMPI is used to change the value of ᨡ, ᨿ, or ᨢ (all *ha*) to *ka* as ᨠᩲ, ᨿᩲ, ᨢᩲ in Mandailing, and to change ᨧ, ᨿ, or ᨿ (all *sa*) to *ca* as ᨧᩲ, ᨿᩲ, ᨿᩲ in Mandailing. The consonant signs ◌ᩱ CONSONANT SIGN NG and ◌ᩲ CONSONANT SIGN H are usually rendered above the

spacing vowels ◌᯲ VOWEL SIGN E, ◌᯳ VOWEL SIGN I, ◌᯵ VOWEL SIGN KARO I, and ◌᯶ VOWEL SIGN O: as in —᯳ *ping*, —᯶ *pong*, —᯲ *pěh*, ᯉ᯵ *pih*.

The main peculiarity of Batak rendering has to do with the way vowel glyphs are re-ordered when the killer (PANGOLAT or PANONGONAN) is used to close the syllable by killing the inherent vowel of a final consonant. This re-ordering is entirely regular and there are no exceptions to it.

ᯖ—᯦	tap	=	ᯖ	ta			+ —	pa	+ ◌᯦	PANGOLAT
ᯖ—᯦᯲	těp	=	ᯖ	ta	+	◌᯲	-ě	+ —	pa	+ ◌᯦	PANGOLAT
ᯖ̄—᯦	tep	=	ᯖ	ta	+	◌̄	-e	+ —	pa	+ ◌᯦	PANGOLAT
ᯖ—᯦᯳	tip	=	ᯖ	ta	+	◌᯳	-i	+ —	pa	+ ◌᯦	PANGOLAT
ᯖ—᯶᯦	top	=	ᯖ	ta	+	◌᯶	-o	+ —	pa	+ ◌᯦	PANGOLAT
ᯖ—᯵᯦	tup	=	ᯖ	ta	+	◌᯵	-u	+ —	pa	+ ◌᯦	PANGOLAT

So although the backing store for *tip* is TA + I + PA + PANGOLAT, the display is not *ᯖ᯳—᯦ (which cannot occur) but rather ᯖ—᯦᯳. One way a font might implement this would be with a set of triplets, *Vowel + Consonant + Killer = glyph-CVK*. In the event that a visual order were entered in the text stream, an error state could be indicated with the retention of the dotted circle, thus:

ᯖ—᯦᯳	tip	=	ᯖ	ta	+	◌᯳	-i	+ —	pa	+ ◌᯦	PANGOLAT (correct)
ᯖ—᯳◌᯦	tapiK	=	ᯖ	ta	+	—	pa	+ ◌᯳	-i	+ ◌᯦	PANGOLAT (incorrect)

Another way of putting this is to say that the PANGOLAT cannot follow a VOWEL SIGN, but only a LETTER.

There are other ways in which a font might implement this behaviour; apparently the preferred method in the Uniscribe model could differ from the description above.

This regular re-ordering poses no significantly new architectural challenge for the Brahmic model; indeed glyph reordering in complex syllables in Tai Tham is far more complex. There are moreover a number of reasons for preferring logical order for Batak. Both open and closed syllables are very frequent in the languages which use Batak: —ᯛ᯶ᯗ᯦ᯂ᯳ᯩ᯦ *por-kis*, ᯀᯪ᯶ᯅᯀ᯶ᯗ᯦ᯂᯪ᯶᯦ *ma-no-ngos-kon*, ᯀᯪ᯦ᯅ—ᯖ᯶ᯗ᯦ᯂᯪ᯶᯦ *man-da-pot-kon*, ᯀᯛ᯶ᯗ᯦ᯂᯛ᯶ᯗ᯦ᯅ᯲ *mor-kor-ja*, ᯖᯛ᯵ᯗ᯦ᯂ *ta-rup-ku*. Phonetic syllable structure is easier to process, to sort, to search, if logical ordering is used, because these cannot be mis-identified as —ᯛ᯶ᯂᯀ᯳ *paro\kasi*, ᯀᯪ᯶ᯅᯀ᯶ᯂᯪ᯶ *manongaso\kano*, ᯀᯪᯅ—ᯖ᯶ᯂᯪ᯶ *mana\dapato\kano*, ᯀᯛ᯶ᯂᯛ᯶ᯗ᯲ *maro\karo\ja*, ᯖᯛ᯵ᯂ *tarapu\ku*—all of which have valid syllable structures. Moreover, like other languages of Indonesia, most speakers are literate in Bahasa Indonesian, and their experience with computing is with that language, which has an extremely phonetic orthography. Their expectation will be to input their language by sound. Similar discussion held with users of the Balinese and Javanese scripts likewise indicated that phonetic input was their expectation. Visual order in the UCS is used with Thai and Lao for reasons of legacy, and with Tai Tham because of its similarity to Thai. All other Brahmic scripts in the UCS use logical order, and Batak need be no exception.

5. Unification. Karo, Mandailing, Pakpak, Simalungun, and Toba each use the script in a different way. While language groups share most of their letters in common, sometimes a letter with a value in one language has a different value in another. The letter ᯀ, for instance is *nya* in Simalunge, Toba, and Mandailing, but *ca* in Karo; compare Latin *c*, which may be [k] or [s] or [ts] or [tʃ] or [dʒ] depending on language. This proposal encodes the superset of forms, regardless of pronunciation. There is a core of

common letters and a set of dialect-specific letters. In this way the encoding model for the Batak script is analogous to the model for Cyrillic, as opposed to the model for Old Italic.

6. Punctuation. Punctuation is not normally used, all letters simply running together, but a number of BINDU characters do exist and are occasionally used to disambiguate similar words or phrases. The BINDU PANGOLAT is trailing punctuation, following a word, surrounding the previous character somewhat.

The *bindu* apparently appears in several forms. The major mark used to begin texts is called the BINDU GODANG 'large bindu'. In letters written on bamboo, the BINDU PINARJOLMA 'human-being-shaped bindu' is used instead of the BINDU GODANG. There are many glyph variants of the bindu pinarjolma; when it is more snake-like than anthropomorphic, it is sometimes called *bindu pinarulok* 'snake-shaped bindu'. The actual length of the glyph for these marks is up to the font designer. It will readily be seen that the variation in the shapes of Batak punctuation is very free.

The minor mark used to begin paragraphs and stanzas is called the BINDU NA METEK 'small bindu'. It may have a number of variants such as BINDU PINARBORAS 'rice-shaped bindu', again used to separate sections of text. These marks can be written as large signs that physically separate sections of text, for instance by means of a long trailing line leading from them. A sign called BINDU JUDUL 'title bindu' is also sometimes used to separate a title from the main text which normally begins on the same line.

7. Collating order. The unified collation order is given below. For reference, the "alphabetical order" of each language is given subsequently

a > ^{S}*a* > *ha* > ^{S}*ha* > ^{M}*ha* > *ba* > ^{K}*ba* >

pa > ^{S}*pa* > *na* > ^{M}*na* > *wa* >> ^{P}*wa* > ^{S}*wa* >

ga > ^{S}*ga* > *ja* > *da* > *ra* > ^{S}*ra* > *ma* > ^{S}*ma* >

^{S}*ta* >> ^{N}*ta* > *sa* > ^{S}*sa* > ^{M}*sa* > *ya* > ^{S}*ya* >

nga > *la* > ^{S}*la* > *nya* >> *ca* > *nda* > *mba* > *i* > *u*

7.1. The Karo alphabet.

a, ha, ka, ba, pa, na, wa, ga, ja, da, ra, ma,

^{N}*ta, sa, ya, nga, la, / ca, nda, mba, i, u*

7.2. The Pakpak alphabet.

a, ha, ka, ba, pa, na, wa, ga, ja, da, ra, ma,

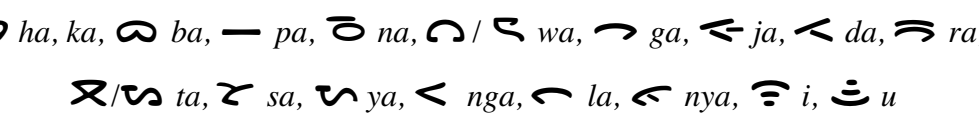
ta, sa, ca, ya, nga, la, i, u

7.3. The Simaluungun alphabet.

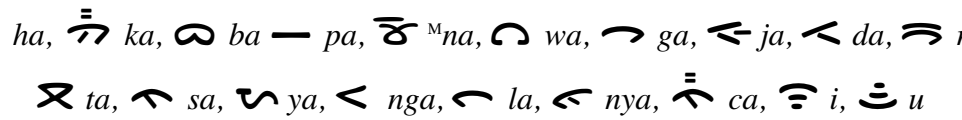
a, ha, ka, ba, pa, na, wa, ga, ja, da, ra, ma,

ta, sa, ya, nga, la, nya, i, u

7.4. The Toba alphabet.



7.5. The Mandailing alphabet.



8. Character names. The character names used follow Kozok 1999. Language identifiers are used to distinguish the characters in UCS terms; usually the language identifier chosen was SIMALUNGUN because Simalungun is the most common variant. It should be noted, however, that 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, defined as C(V(Cp|F)) where a consonant C may be followed by a vowel V which may be followed either by a killed consonant Cp or a final *-ng* or *-h* F. Batak punctuation marks can be expected to have behaviour similar to that of Devanagari DANDA.

10. Unicode Character Properties.

```
1BC0;BATAK LETTER A;Lo;0;L;;;;;N;;;;;
1BC1;BATAK LETTER SIMALUNGUN A;Lo;0;L;;;;;N;;;;;
1BC2;BATAK LETTER HA;Lo;0;L;;;;;N;;;;;
1BC3;BATAK LETTER SIMALUNGUN HA;Lo;0;L;;;;;N;;;;;
1BC4;BATAK LETTER MANDAILING HA;Lo;0;L;;;;;N;;;;;
1BC5;BATAK LETTER BA;Lo;0;L;;;;;N;;;;;
1BC6;BATAK LETTER KARO BA;Lo;0;L;;;;;N;;;;;
1BC7;BATAK LETTER PA;Lo;0;L;;;;;N;;;;;
1BC8;BATAK LETTER SIMALUNGUN PA;Lo;0;L;;;;;N;;;;;
1BC9;BATAK LETTER NA;Lo;0;L;;;;;N;;;;;
1BCA;BATAK LETTER MANDAILING NA;Lo;0;L;;;;;N;;;;;
1BCB;BATAK LETTER WA;Lo;0;L;;;;;N;;;;;
1BCC;BATAK LETTER SIMALUNGUN WA;Lo;0;L;;;;;N;;;;;
1BCD;BATAK LETTER PAKPAK WA;Lo;0;L;;;;;N;;;;;
1BCE;BATAK LETTER GA;Lo;0;L;;;;;N;;;;;
1BCF;BATAK LETTER SIMALUNGUN GA;Lo;0;L;;;;;N;;;;;
1BD0;BATAK LETTER JA;Lo;0;L;;;;;N;;;;;
1BD1;BATAK LETTER DA;Lo;0;L;;;;;N;;;;;
1BD2;BATAK LETTER RA;Lo;0;L;;;;;N;;;;;
1BD3;BATAK LETTER SIMALUNGUN RA;Lo;0;L;;;;;N;;;;;
1BD4;BATAK LETTER MA;Lo;0;L;;;;;N;;;;;
1BD5;BATAK LETTER SIMALUNGUN MA;Lo;0;L;;;;;N;;;;;
1BD6;BATAK LETTER SOUTHERN TA;Lo;0;L;;;;;N;;;;;
1BD7;BATAK LETTER NORTHERN TA;Lo;0;L;;;;;N;;;;;
1BD8;BATAK LETTER SA;Lo;0;L;;;;;N;;;;;
1BD9;BATAK LETTER SIMALUNGUN SA;Lo;0;L;;;;;N;;;;;
1BDA;BATAK LETTER MANDAILING SA;Lo;0;L;;;;;N;;;;;
1BDB;BATAK LETTER YA;Lo;0;L;;;;;N;;;;;
1BDC;BATAK LETTER SIMALUNGUN YA;Lo;0;L;;;;;N;;;;;
1BDD;BATAK LETTER NGA;Lo;0;L;;;;;N;;;;;
1BDE;BATAK LETTER LA;Lo;0;L;;;;;N;;;;;
1BDF;BATAK LETTER SIMALUNGUN LA;Lo;0;L;;;;;N;;;;;
1BE0;BATAK LETTER NYA;Lo;0;L;;;;;N;;;;;
1BE1;BATAK LETTER CA;Lo;0;L;;;;;N;;;;;
1BE2;BATAK LETTER NDA;Lo;0;L;;;;;N;;;;;
1BE3;BATAK LETTER MBA;Lo;0;L;;;;;N;;;;;
1BE4;BATAK LETTER I;Lo;0;L;;;;;N;;;;;
1BE5;BATAK LETTER U;Lo;0;L;;;;;N;;;;;
1BE6;BATAK SIGN TOMPI;Mn;7;NSM;;;;;N;;;;;
1BE7;BATAK VOWEL SIGN E;Mc;0;L;;;;;N;;;;;
1BE8;BATAK VOWEL SIGN PAKPAK E;Mn;0;NSM;;;;;N;;;;;
1BE9;BATAK VOWEL SIGN EE;Mn;0;NSM;;;;;N;;;;;
1BEA;BATAK VOWEL SIGN I;Mc;0;L;;;;;N;;;;;
1BEB;BATAK VOWEL SIGN KARO I;Mc;0;L;;;;;N;;;;;
1BEC;BATAK VOWEL SIGN O;Mc;0;L;;;;;N;;;;;
1BED;BATAK VOWEL SIGN KARO O;Mn;0;NSM;;;;;N;;;;;
1BEE;BATAK VOWEL SIGN U;Mn;0;NSM;;;;;N;;;;;
1BEF;BATAK VOWEL SIGN U FOR SIMALUNGUN SA;Mn;0;NSM;;;;;N;;;;;
1BF0;BATAK CONSONANT SIGN NG;Mn;0;NSM;;;;;N;;;;;
1BF1;BATAK CONSONANT SIGN H;Mn;0;NSM;;;;;N;;;;;
1BF2;BATAK PANGOLAT;Mn;9;L;;;;;N;;;;;
1BF3;BATAK PANONGONAN;Mn;9;L;;;;;N;;;;;
```

```
1BFA;BATAK SYMBOL BINDU GODANG;Po;0;L;;;;;N;;;;;
1BFB;BATAK SYMBOL BINDU PINARJOLMA;Po;0;L;;;;;N;;;;;
1BFC;BATAK SYMBOL BINDU NA METEK;Po;0;L;;;;;N;;;;;
1BFD;BATAK SYMBOL BINDU PINARBORAS;Po;0;L;;;;;N;;;;;
1BFE;BATAK SYMBOL BINDU JUDUL;Po;0;L;;;;;N;;;;;
1BFF;BATAK SYMBOL BINDU PANGOLAT;Po;0;L;;;;;N;;;;;
```

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

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

Kozok, Uli. 2004. *Reference list to the Batak-Dutch Dictionary by H. N. Van der Tuuk = Daftar rujukan untuk Kamus Batak-Belanda oleh H. N. Van der Tuuk*. Jakarta: Wedatama Widya Sastra. ISBN 979-3258-37-3

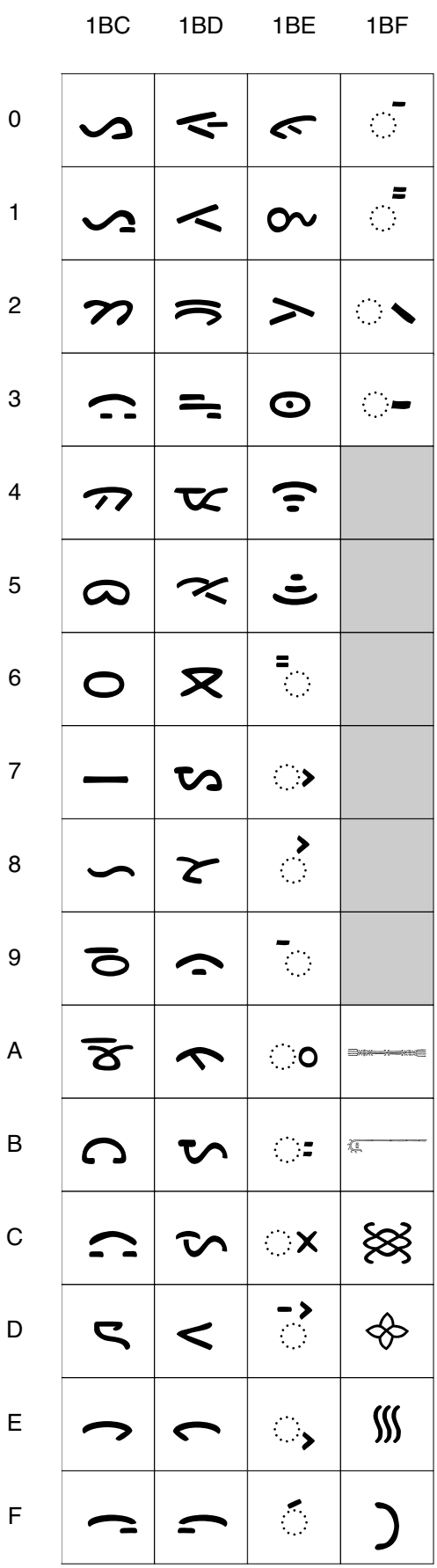
Meerwaldt, J. H. 1904. *Handleiding tot de beoefening der bataksche taal*. Leiden: E. J. Brill.

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

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

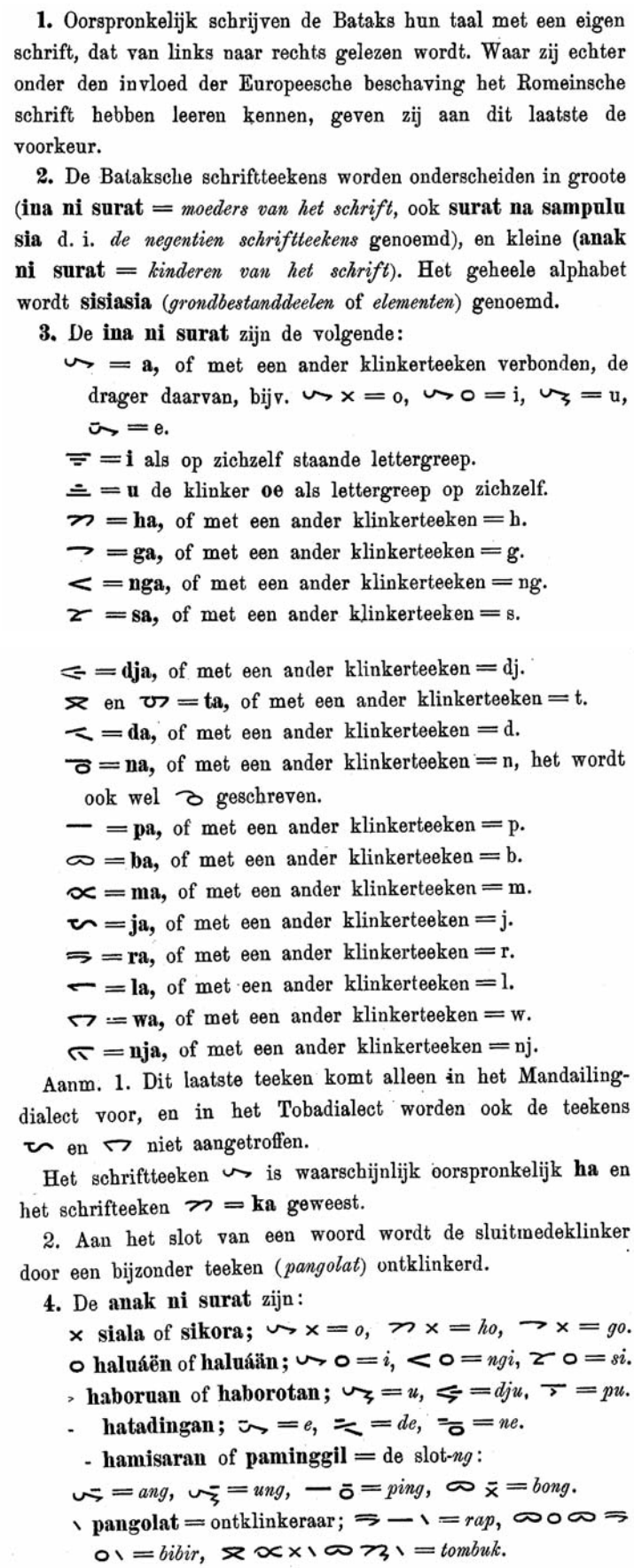
12. Acknowledgements. This project was made possible in part by a grant from the U.S. National Endowment for the Humanities, which funded the which funded the Universal Scripts Project (part of the Script Encoding Initiative at UC Berkeley) in respect of the Batak encoding. Any views, findings, conclusions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

Row 1B: BATAK DRAFT



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

Figures.



1. Oorspronkelijk schrijven de Bataks hun taal met een eigen schrift, dat van links naar rechts gelezen wordt. Waar zij echter onder den invloed der Europeesche beschaving het Romeinsche schrift hebben leeren kennen, geven zij aan dit laatste de voorkeur.

2. De Bataksche schriftteekens worden onderscheiden in groote (**ina ni surat** = *moeders van het schrift*, ook **surat na sampulu sia** d. i. *de negentien schriftteekens* genoemd), en kleine (**anak ni surat** = *kinderen van het schrift*). Het geheele alphabet wordt **sisiasia** (*grondbestanddeelen* of *elementen*) genoemd.

3. De **ina ni surat** zijn de volgende:

= **a,** of met een ander klinkerteeken verbonden, de drager daarvan, bijv. × = o, o = i, = u, = e.

= **i** als op zichzelf staande lettergreep.

= **u** de klinker **oe** als lettergreep op zichzelf.

= **ha,** of met een ander klinkerteeken = h.

= **ga,** of met een ander klinkerteeken = g.

= **nga,** of met een ander klinkerteeken = ng.

= **sa,** of met een ander klinkerteeken = s.

= **dja,** of met een ander klinkerteeken = dj.

en = **ta,** of met een ander klinkerteeken = t.

= **da,** of met een ander klinkerteeken = d.

= **na,** of met een ander klinkerteeken = n, het wordt ook wel geschreven.

= **pa,** of met een ander klinkerteeken = p.

= **ba,** of met een ander klinkerteeken = b.

= **ma,** of met een ander klinkerteeken = m.

= **ja,** of met een ander klinkerteeken = j.

= **ra,** of met een ander klinkerteeken = r.

= **la,** of met een ander klinkerteeken = l.

= **wa,** of met een ander klinkerteeken = w.

= **nja,** of met een ander klinkerteeken = nj.

Aanm. 1. Dit laatste teeken komt alleen in het Mandailing-dialect voor, en in het Tobadialect worden ook de teekens en niet aangetroffen.

Het schriftteeken is waarschijnlijk oorspronkelijk **ha** en het schrifteeken = **ka** geweest.

2. Aan het slot van een woord wordt de sluitmedeklinker door een bijzonder teeken (*pangolat*) ontklinkerd.

4. De **anak ni surat** zijn:

× **siala** of **sikora;** × = *o*, × = *ho*, × = *go*.

o **haluáën** of **haluáän;** o = *i*, o = *ngi*, o = *si*.

› **haboruan** of **haborotan;** = *u*, = *dju*, = *pu*.

- **hatadingan;** = *e*, = *de*, = *ne*.

- **hamisaran** of **paminggil** = de slot-*ng*:

= *ang*, = *ung*, = *ping*, = *bong*.

\ **pangolat** = ontklinkeraar; = *rap*, = *bibir*, = *tombuk*.

Figure 1. Description in Dutch of the Batak script.

Figure 2. Sample of Batak text on a sign for a hospital in Sumatra.

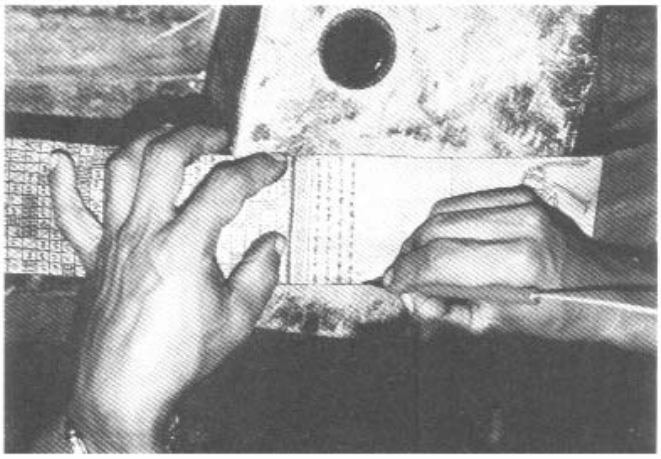


Figure 3. Photograph of a person writing of Batak text.
The hand position shows right-to-left directionality.

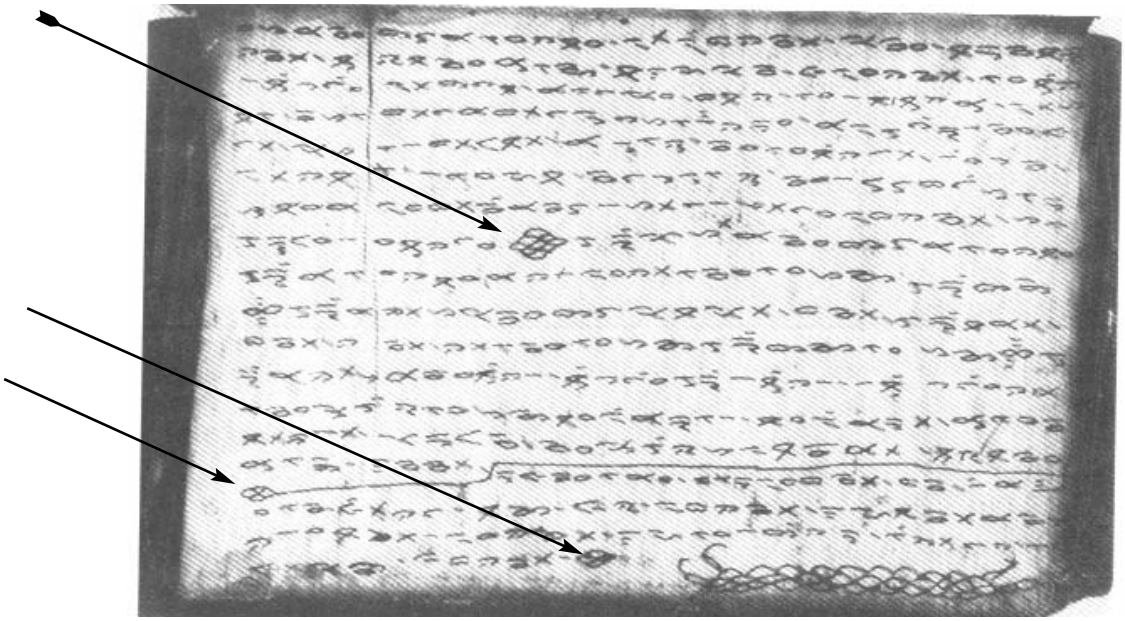


Figure 4. Sample of Batak text showing one example of BINDU NA METEK and two examples of BINDU PINARBORAS, one of which has a trailing line following from it. This kind of formatting would be achieved by a higher-level protocol in an encoded text.

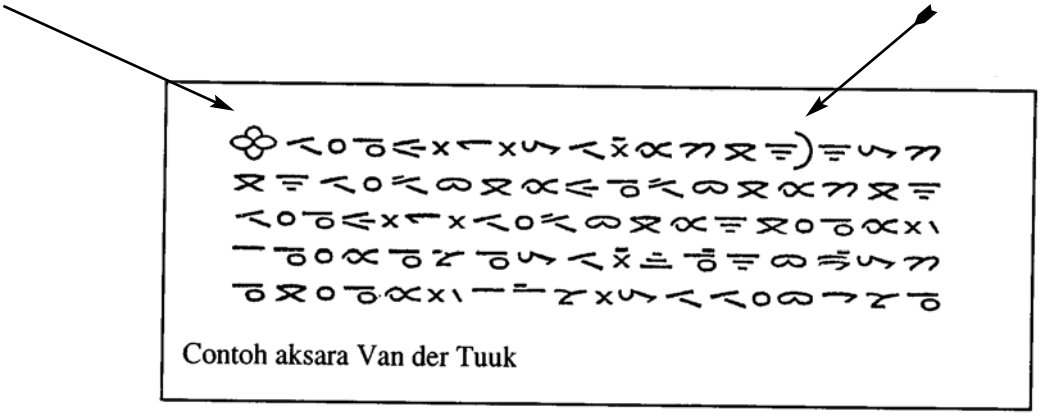


Figure 5. Sample of Batak text awr by van der Tuuk, showing BINDU PINARBORAS and BINDU PANGOLAT.

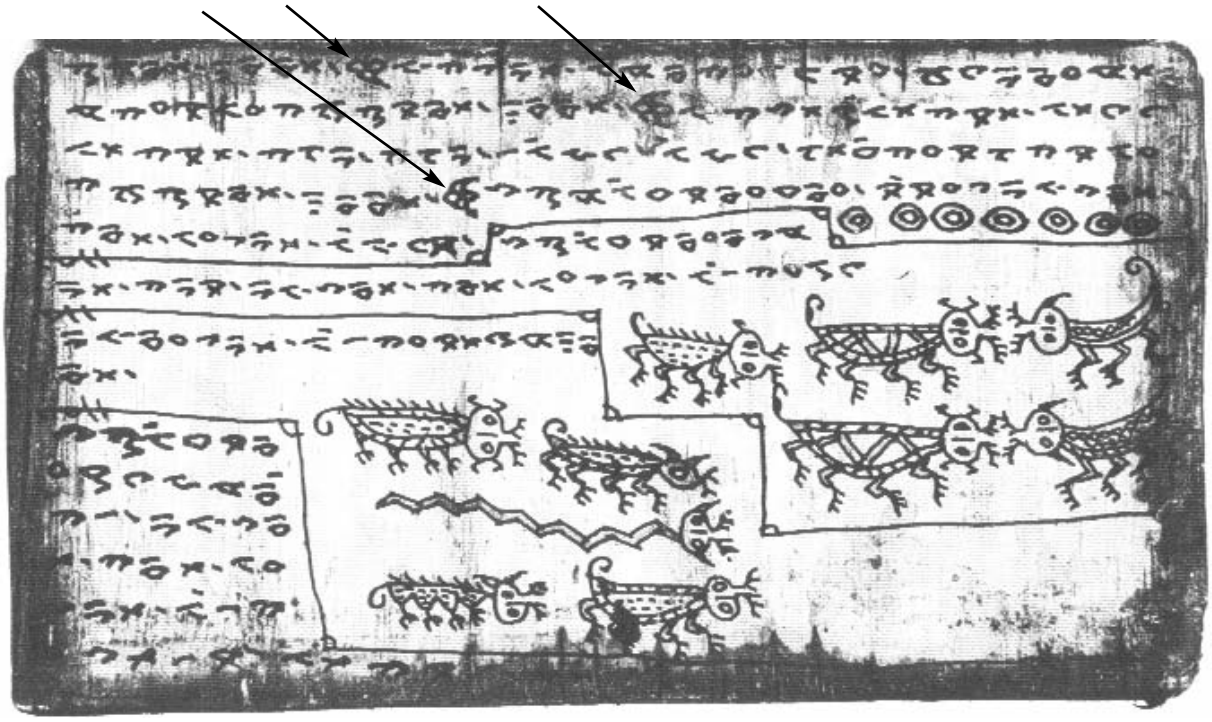


Figure 6. Sample of Batak text showing three examples of BINDU NA METEK.

Figure 7. Sample of Batak text showing BINDU GODANG in the first line.

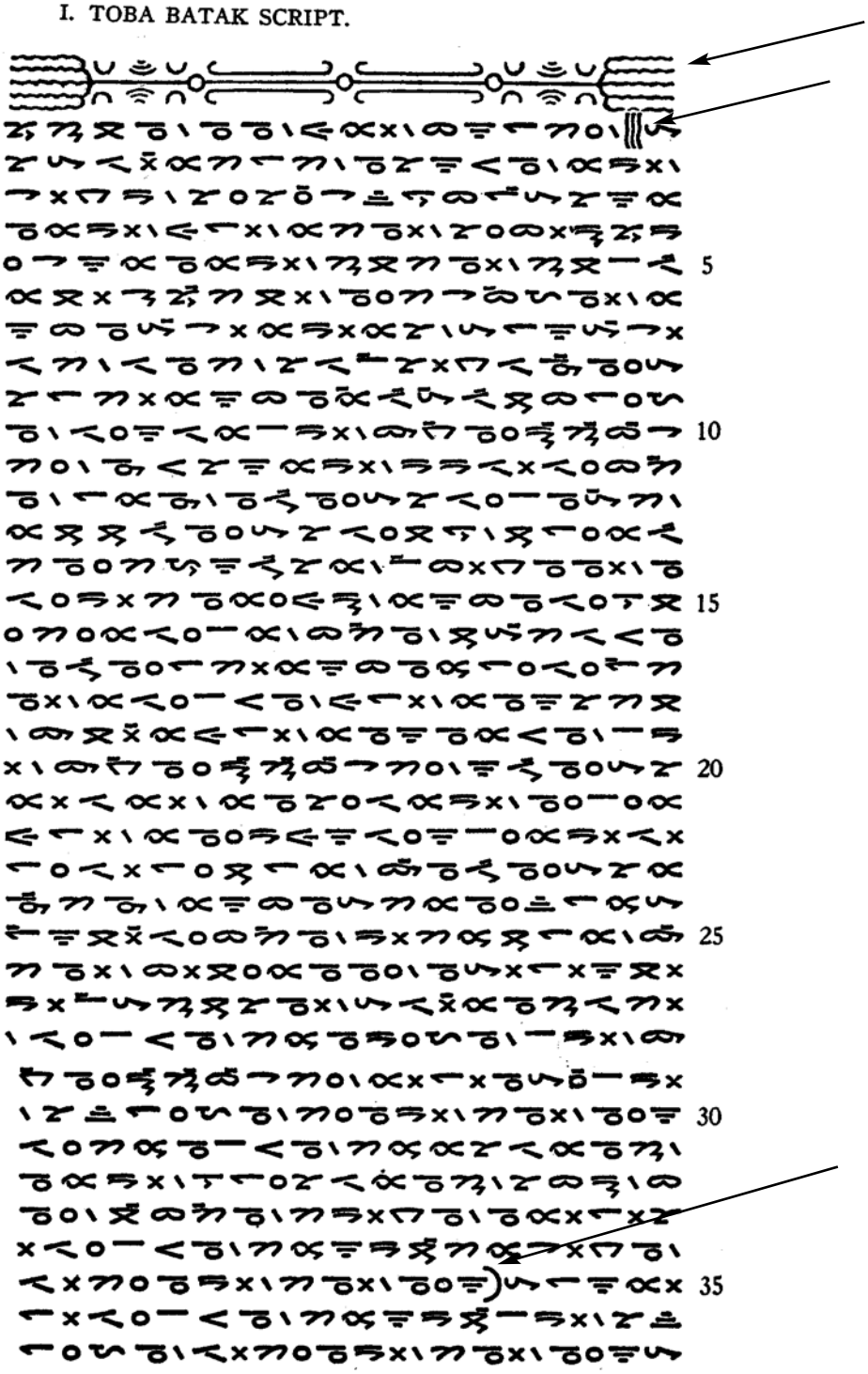


Figure 8. Sample of Toba Batak text set by van der Tuuk, showing BINDU GODANG, BINDU JUDUL, and BINDU PANGOLAT.

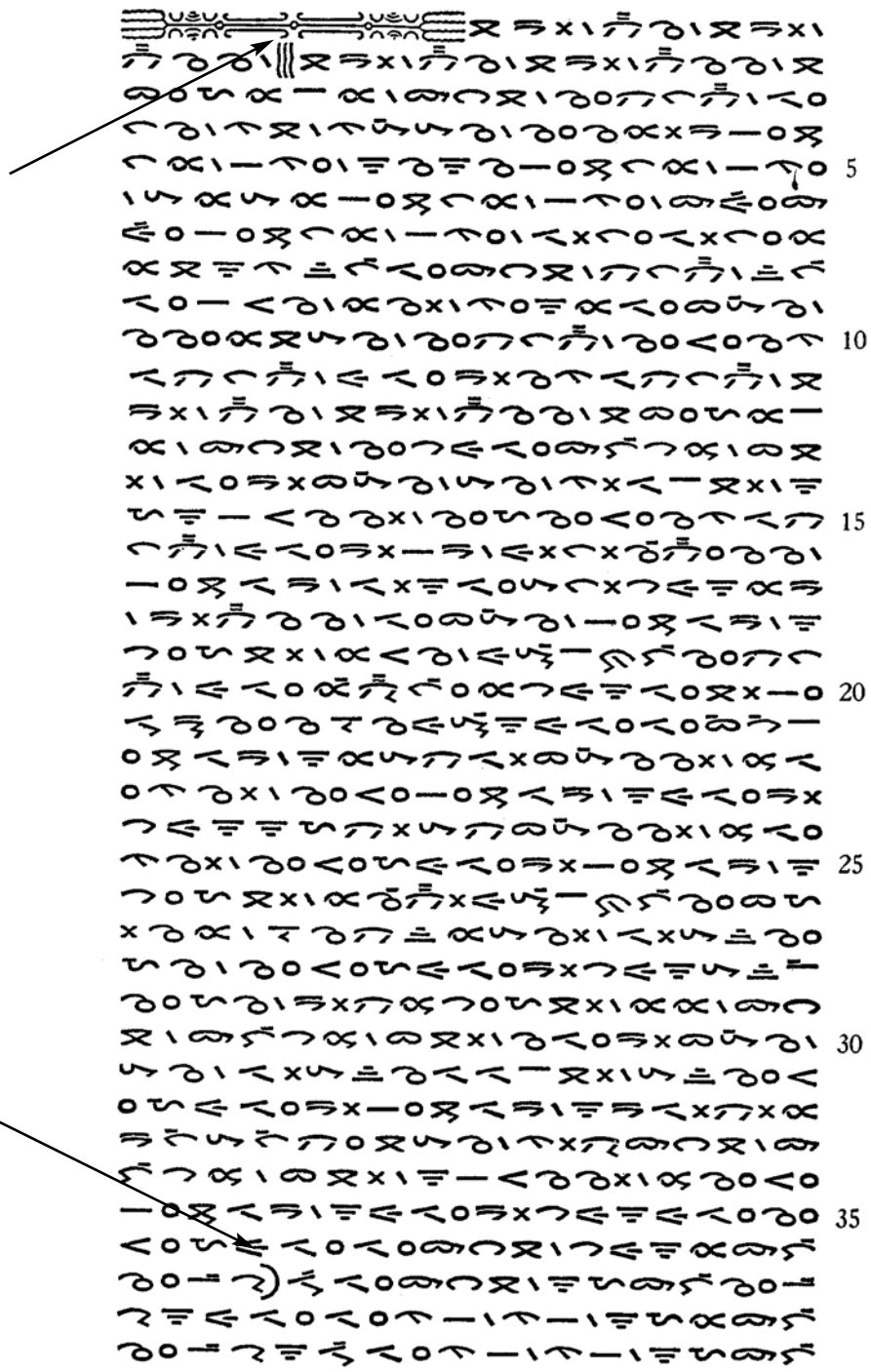


Figure 9. Sample of Mandailing Batak text showing BINDU GODANG, BINDU JUDUL, and BINDU PANGOLAT.

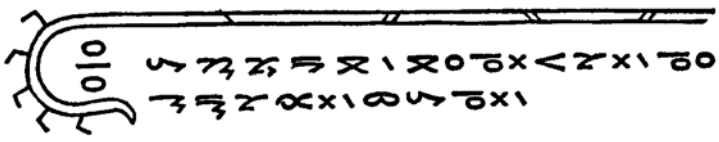


Figure 10. Sample of Batak text showing BINDU PINARJOLMA set as a kind of drop-cap with text nestled within it.

Figure 11. Sample of Batak text showing BINDU GODANG above and BINDU NA METEK in the centre.

Figure 12. Sample of Batak text showing two examples of BINDU PINARBORAS, one with a trailing line.

Figure 13. Sample of Batak text showing a number of examples of BINDU PINARJOLMA.

A. Administrative

1. Title

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

2. Requester's name

UC Berkeley Script Encoding Initiative (Universal Scripts Project); authors: Michael Everson and Uli Kozok

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

Liaison contribution.

4. Submission date

2008-10-07

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

58.

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. UTR#3, N3293R

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 northern 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?