To: LSA and UC Berkeley Communities
From: Deborah Anderson, UCB representative and LSA liaison
Date: 8 December 2004 (revised 10 January 2005)
Topic: Unicode Technical Meeting #101, 15 -18 November 2004, Cupertino, California

As the UC Berkeley representative and LSA liaison, I am most interested in the proposals for new characters and scripts that were discussed at the UTC, so these topics are the focus of this report. For the full minutes, readers should consult the "Unicode Technical Committee Minutes" web page (http://www.unicode.org/consortum/utc-minutes.html), where the minutes from this meeting will be posted several weeks hence.

***I. Proposals for New Scripts and Additional Characters***

A summary of the proposals and the UTC's decisions are listed below. As the proposals discussed below are made public, I will post the URLs on the SEI web page (www.linguistics.berkeley.edu/sei).

***A. Linguistics Characters***

Lorna Priest of SIL International submitted three proposals for additional linguistics characters. Most of the characters proposed are used in the orthographies of languages from Africa, Asia, Mexico, Central and South America. (For details on the proposed characters, with a description of their use and an image, see the appendix to this document.)

Two characters from these proposals were not approved by the UTC because there are already characters encoded that are very similar. The evidence did not adequately demonstrate that the proposed characters are used distinctively. The two problematical proposed characters were: the modifier straight letter apostrophe (used for a glottal stop, similar to ' APOSTROPHE U+0027) and the Latin small "at" sign (used for Arabic loanwords in an orthography for the Koalib language from the Sudan, similar to @ COMMERCIAL AT U+0040). The encoded characters do not have the "letter" character property assigned to them. The uppercase "at" was also not approved.

One character, LATIN CAPITAL LETTER L WITH BAR, was not approved because it has already been proposed (U+023D).

Of particular interest to linguists were the following phonetic symbols, which were approved (see glyphs in the Appendix):

Contour tone diacritics:

1DC4 COMBINING MACRON-ACUTE
1DC5 COMBINING GRAVE-MACRON
1DC6 COMBINING MACRON-GRAVE
1DC7 COMBINING ACUTE-MACRON
1DC8 COMBINING GRAVE-ACUTE-GRAVE
1DC9 COMBINING ACUTE-GRAVE-ACUTE

024F LATIN SMALL LETTER V WITH CURL, which is used by African linguists for a labiodental flap.

***B. Mathematical and Technical Symbols***

Asmus Freytag proposed 26 mathematical and technical symbols (L2/04-410). A number of the requested symbols come from the STIX project, which is reviewing mathematical and technical literature for needed characters. All 26 were approved.

A second proposal for 3 additional mathematical symbols by Murray Sargent (L2/04-411) was returned to the author for further research.

***C. Other Characters***

1. Proposal to add Invisible Letter to the UCS (WG2-N2822) by Michael Everson, Peter Constable, Rick McGowan and Ken Whistler
(http://std.dkuug.dk/jtc1/sc2/wg2/docs/n2822.pdf)

This character was proposed as a means of displaying combining marks in isolation, instead of space before a combining mark. Implementation of space + combining mark has proved problematic on a number of platforms and implementations. The solution proposed is an "invisible letter" with a letter property. There is potential use in a number of Indic scripts. Such a character could also be used to represent a missing or obscured base letter within the middle of a word in archaic documents.

This character was rejected by the UTC. The character U+00A0 NON-BREAKING SPACE (NBSP) has good semantics for such a use (though it does not have a letter property, nor was it originally intended for the purpose of displaying combination marks in isolation).

2. Proposal to Encode Aktieselskab by Andreas Stötzner (L2/04-394)
This character appears in Danish and Norwegian typography to indicate the legal status of a company; "Aktieselskab" means ‘joint-stock company’. The character was approved.

3. Proposal to Encode Capital Double S by Andreas Stötzner (L2/04-395)
This proposal requested the inclusion of an uppercase esszet. The character was rejected, as it was deemed to be a typographical issue.

4. Proposal to Encode Six Kannada characters by Michael Everson
(http://std.dkuug.dk/jtc1/sc2/wg2/docs/n2860.pdf)
Four of the proposed signs are used in Kannada to provide support for Sanskrit. The UTC agreed to not oppose these four if the issue comes up at the ISO WG2 meeting, but the UTC would oppose including the remaining two characters, the DANDA and the DOUBLE DANDA. The question of how to handle the DANDAs (i.e., whether they should be disunified from the Devanagari DANDAs) will be taken up in a Public Review Issue (http://www.unicode.org/review/).

5. Proposal to Encode Gujarati Signs Pao, Addho, and Pono by Manoj Jain,Government of India (L2/04-358)

This proposal requires additional information.

6. Two Greek proposals by Nick Nicholas
A preliminary proposal by Nick Nicholas to encode GREEK LETTER LOWERCASE HETA and CAPITAL HETA was briefly discussed. Additional information is needed on this proposal, including how other electronic text projects handle these characters.

Nick Nicholas submitted a second proposal for input from the UTC for six epigraphic characters (GREEK SMALL LETTER ARCHAIC SAMPI, GREEK LETTER CAPITAL SAMPI; GREEK LETTER SMALL TSAN, GREEK LETTER CAPITAL TSAN; GREEK LETTER SMALL CORINTHIAN EI, GREEK LETTER CAPITAL CORINTHIAN EI). There was some question about the eligibility of the characters, already raised in the document by Nick Nicholas. Additional evidence and documentation is required.

***D. Other Script Proposals***
1. Preliminary proposal for Balinese by Michael Everson
(http://std.dkuug.dk/jtc1/sc2/wg2/docs/n2856.pdf)
Further work is necessary on this proposal. The author, Michael Everson, will meet with the user community on a trip to Bali in January and expects to refine the proposal and include additional necessary information and documentation.

2. Preliminary proposal for Lanna by Martin Hosken, SIL (L2/04-351)
Additional information is needed on this proposal.

3. Proposals for Lepcha and Vedic accents (etc.) by the Government of India
These proposals will need additional work.

4. Phags-pa
Four documents were submitted by China regarding the Phags-pa script on 17 November 2004, including a new script proposal which differs from one already under ballot (contained in http://std.dkuug.dk/jtc1/sc2/wg2/docs/n2846.pdf). The four documents are:

a. http://std.dkuug.dk/jtc1/sc2/wg2/docs/N2869c.doc (Updated proposal to encode the Phags-pa Script) by G. Battsengel, MASM(Mongolian Agency for Standardization and Metrology)

b. http://std.dkuug.dk/jtc1/sc2/wg2/docs/n2869.pdf (Proposal to Encode the Phags-pa Script) by China and Mongolia

c. http://std.dkuug.dk/jtc1/sc2/wg2/docs/n2870.pdf (Summary of the Revised User's Agreement Related to Phags-pa Script) China National Information Technology Standardization Technical Committee, Mongolian Agency for Standardization & Metrology

d. http://std.dkuug.dk/jtc1/sc2/wg2/docs/n2871.pdf (Some Problems on the Encoding of Phags-pa Script) China National Information Technology Standardization Technical Committee, Mongolian Agency for Standardization & Metrology

***E. Other Character and Script Topics***

1. Holam Codepoint Change

A request by Peter Kirk (L2/04-346) to swap the codepoints for HEBREW POINT HOLAM HASER FOR VAV and HEBREW POINT QAMATS QATAN was approved. The new codepoint for HOLAM HASER FOR VAV will be U+05BA, and U+05C7 for QAMATS QATAN.

2. Glyph Change for CAPITAL LETTER Y WITH HOOK (L2/04-399 = N3768)

A proposal by Alain LaBonté requested the glyph for LATIN CAPITAL LETTER WITH HOOK be changed. The capital letter reference glyph for U+01B3 has the hook on the left, but examples cited in document L2/04-399 showed examples of it on the right, similar to that of the lowercase character (U+01B4). This character is used in some African scripts, and the error in the chart has cause confusion. The UTC approved the change.

3. A number of Indic script topics will be posted as Public Review Issues, including:

a. Disunification of the DANDAS
b. EYELASH-RA
c. 5 CHILLU characters for Malayalam

4. Saurashtra Line Breaking and Other Properties by Rick McGowan (L2/04-392)

A document by Rick McGowan addressed the line breaking and other properties for this script. The document will be sent for posting to WG2.

5. Cuneiform Properties by Rick McGowan (L2/04-394)

This document documented numeric and other properties for cuneiform. The information will be included in the data files for Unicode 5.0

***F. Collation***

Document L2/04-407 (=Public Review Issue #47) by Mark Davis was a request to change the default Unicode Collation Algorithm for 10 Latin characters (and their lowercases). There was some discussion whether such changes needed to be made to the default collation, or are better covered by individual tailoring. The request to change the default UCA for these 10 characters was approved.

***II. Future Meetings***

Future UTC (/L2) meetings will take place as follows:

7-10 Feb. 2005, Mountain View, CA
10-13 May 2005, Pleasanton, CA
16-19 August 2005, Redmond, WA
1-4 November 2005, San Jose, CA

The next ISO WG2 meeting will be held 24-28 January 2005 in Xiamen, China.
The following WG2 will be 1-5 August 2005, probably in Europe.

### *III. FPDAM1 and PDAM2 Comments*

The UTC meeting concluded with the reading of the comments that will accompany the positive vote on the Preliminary Draft Amendment 2 (PDAM2) and the Final Preliminary Draft Amendment 1 (FPDAM1). Both amendments include new characters and scripts for ISO/IEC 10646:2003. Comments on PDAM2 closed 23 November 2004, comments on FPDAM1 close 23 December. Errors should be reported promptly to one's national body representative.

FPDAM1 includes many new linguistic characters, particularly in the following blocks: Latin Extended-B, Combining Diacritical Marks, Phonetic Extensions, Phonetic Extension Supplement, Superscripts and Subscripts, and Modifier Tone Letters. It also includes new scripts (New Tai Lue, Buginese, Glagolitic, Coptic, Tifinagh, and Kharosthi). A version of the FPDAM is accessible
http://std.dkuug.dk/jtc1/sc2/wg2/docs/n2845.pdf

The Preliminary Draft Amendment 2 (PDAM2) contains scripts and characters that are in line for inclusion into Unicode 5.0. It is accessible at:
http://std.dkuug.dk/jtc1/sc2/wg2/docs/n2846.pdf

The disposition of comments will take place at the ISO WG2 meeting in Xiamen, China, from 24-28 January 2005. The FPDAM then goes to ISO for one final 2-month ballot, and then will be synchronized with Unicode 4.1, which is due to be released at the end of March.

(The WG2 document register is located at: http://std.dkuug.dk/jtc1/SC2/wg2/.)

## *Appendix*

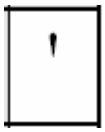
[Note: Tentative codepoint assignments are given below.]

**1. Proposal to encode additional Latin orthographic characters by Lorna Priest, SIL (L2/04-372) (1 November 2004)**

**UTC Decision on this proposal:** The MODIFIER LETTER STRAIGHT LETTER APOSTROPHE character was not approved at this time nor was LATIN CAPITAL LETTER L WITH BAR since it had been approved in an earlier proposal. All the other characters were accepted.

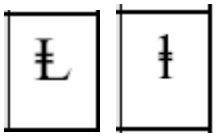
Characters from this proposal included:

MODIFIER LETTER STRAIGHT APOSTROPHE



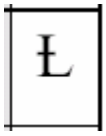
Comment: Used for a glottal stop in a variety of languages, including el chontal de la sierra de Oaxaca, Tlapaneco de Malinaltepec of Mexico, Ese Ejja of Bolivia, Maya Mopán of Guatemala, Dano (Upper Asano) of Papua New Guinea, and FraFra of Ghana.

2C60 LATIN CAPITAL L WITH DOUBLE BAR
2C61 LATIN SMALL LETTER L WITH DOUBLE-BAR



Comment: Both the capital and small letter L WITH DOUBLE BAR are used for a velar fricative lateral in the orthography of the Melpa and Nii languages of Papua New Guinea. Since the 1990s, the Nii have used the double-barred L, but originally used the one-bar L.

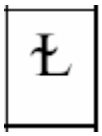
LATIN CAPITAL LETTER L WITH BAR



• lowercase is U+019A

Comment: This character had already been proposed and has tentatively been assigned the codepoint U+023D, so it was not approved.

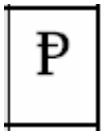
2C62 LATIN CAPITAL LETTER L WITH MIDDLE TILDE



• lowercase is U+026B

Comment: The Kobon language of Papua New Guinea uses the above symbols in its orthography: LATIN CAPITAL LETTER L WITH BAR for a retroflexed flapped lateral, and LATIN CAPITAL LETTER L WITH MIDDLE TILDE for an alveopalatal lateral. (LATIN CAPITAL LETTER L is used for an alveolar lateral.) The LATIN CAPITAL LETTER L WITH BAR has also been used in a nearby language, the Haruai.

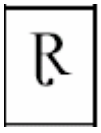
2C63 LATIN CAPITAL LETTER P WITH STROKE



• lowercase is in the Pipeline (proposed codepoint is U+1D7D)

Comment: used by both Letuama and Tanimuka of Columbia for a bilabial fricative.

2C64 LATIN CAPITAL LETTER R WITH TAIL

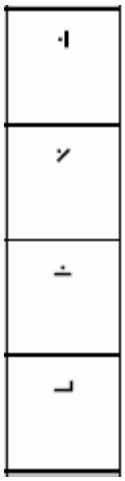


• lowercase is U+027D
Comment: CAPITAL LETTER R WITH TAIL is used in a number of Sudanese languages (Heiban, Koalib, Moro, and Otoro).

**2. Proposal to Encode Chinantec Tone Marks and Orthographic 'at' Characters by Lorna Priest, SIL (L2/04-349) (Date: 27 August 2004)**

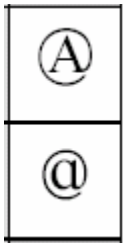
**UTC Decision on this proposal:** All the characters were approved except for the two "AT" signs. More evidence is needed for this character. Also, the UTC will not consider changing the properties for the current @ sign to a letter property.

A717 MODIFIER LETTER DOT VERTICAL BAR
A718 MODIFIER LETTER DOT SLASH
A719 MODIFIER LETTER DOT HORIZONTAL BAR
A71A MODIFIER LETTER LOWER RIGHT CORNER ANGLE



Comment: The above four characters are used orthographically for the Ozumacín Chinantec language of Mexico.

xx05 LATIN CAPITAL LETTER AT
xx06 LATIN SMALL LETTER AT

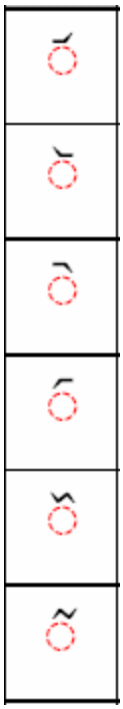


Comment: The LATIN CAPITAL LETTER AT and LATIN SMALL LETTER AT are used commonly for Arabic loan words in the orthography of the Koalib language of Sudan. The lowercase letter resembles closely the COMMERCIAL AT sign, which does not have a "letter" character property.

**3. Revised Proposal for Additional Latin Phonetic and Orthographic Characters by Lorna Priest (L2/04-348; replaces N2847; revised 23 August 2004)**

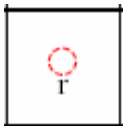
**UTC Decision on this proposal:** All the proposed characters were approved.

1DC4 COMBINING MACRON-ACUTE
1DC5 COMBINING GRAVE-MACRON
1DC6 COMBINING MACRON-GRAVE
1DC7 COMBINING ACUTE-MACRON
1DC8 COMBINING GRAVE-ACUTE-GRAVE
1DC9 COMBINING ACUTE-GRAVE-ACUTE



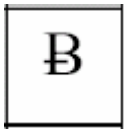
Comment: COMBINING MACRON-ACUTE, COMBINING GRAVE-MACRON, and COMBINING GRAVE-ACUTE-GRAVE are attested in the *IPA Handbook*. The COMBINING GRAVE-MACRON and COMBINING MACRON-GRAVE are used orthographically in the Bette language of Nigeria.

1DCA COMBINING LATIN SMALL LETTER R BELOW



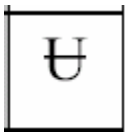
Comment: This character is used orthographically in four languages in Indonesia, namely Mongondow, Sangir, Siau and Talaud.

0242 LATIN CAPITAL LETTER B WITH STROKE



• lowercase is 0180
Comment: This character used for at least two languages of Vietnam (Jorai and Katu).

0243 LATIN CAPITAL LETTER U BAR



• lowercase is 0289 .

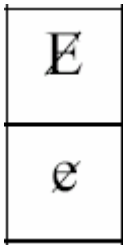
Comments: This character is used for a wide variety of languages, including Mesem and Melpa (languages from Papua New Guinea), Sayula Popoluca of Mexico, the Badwe'e language of Cameroon, the Budu language of Democratic Republic of Congo, Comanche, and Arhuaco of Colombia.

0244 LATIN CAPITAL LETTER TURNED V

• lowercase is 028C .

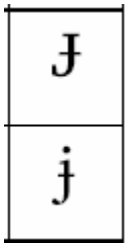
Comments: This character appears in the Nankina language of Papua New Guinea and North Tepehuan of Mexico.

0245 LATIN CAPITAL LETTER E WITH STROKE
0246 LATIN SMALL LETTER E WITH STROKE



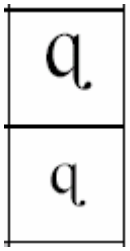
Comments: These characters are used orthographically in the Southeastern Tepehuan language of Mexico.

0247 LATIN CAPITAL LETTER J WITH STROKE
0248 LATIN SMALL LETTER J WITH STROKE



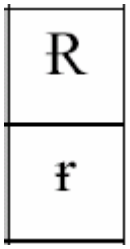
Comments: These characters are used in Arhuaco, a language in Columbia, for a voiced alveo-palatal affricate.

0249 LATIN CAPITAL LETTER SMALL Q WITH HOOK TAIL
024A LATIN SMALL LETTER Q WITH HOOK TAIL



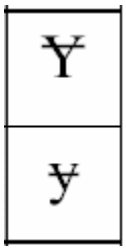
Comments: Lutheran missionaries in the 1930s (or 1940s) introduced these two characters for writing the Numanggang language of Papua New Guinea. In 2002, the community discontinued their use, though they are found in song books and liturgical materials. They are also used in literature of the Kâte language, which is distantly related to Numanggang.

024B LATIN CAPITAL LETTER R WITH STROKE
024C LATIN SMALL LETTER R WITH STROKE



Comments: LATIN SMALL LETTER R WITH STROKE and LATIN CAPITAL LETTER R WITH STROKE are used in the Kanuri orthography of Niger.

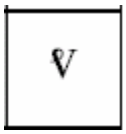
024D LATIN CAPITAL LETTER Y WITH STROKE
024E LATIN SMALL LETTER Y WITH STROKE



Comments: These characters are used to write the Lubuagan Kalinga language of the Philippines.

LATIN SMALL LETTER Y WITH STROKE and LATIN CAPITAL LETTER Y WITH STROKE are used orthographically in the Lubuagan Kalinga language of the Philippines.

024F LATIN SMALL LETTER V WITH CURL



Comments: This character is a phonetic symbol used to designate a labiodental flap. Though not approved by the IPA, it is widely used, especially among African linguists.