Universal Multiple-Octet Coded Character Set
International Organization for Standardization
Международная организация по стандартизации

Doc Type: **Working Group Document**
Title: **Justification for placing the Uralic Phonetic Alphabet in the BMP**
Source: **Michael Everson, Erkki Kolehmainen, Klaas Ruppel, Trond Trosterud**
Status: **Expert contribution**
Date: **2002-05-21**

In N2419, the plain-text requirement for UPA is stated clearly:

> **Requirement for plain-text representation.** Because of the precise nature of the linguistic data described by this system, we consider it absolutely necessary that the UPA be fully representable in plain text in the UCS. In Finland the tendency is being established, that scientific work is published electronically. (An article on this subject was published several years ago in Finland: "Yliopistot julkaisevat opinnäytteitä verkossa" ('Universities publish doctoral theses on the net') by Tuomas Hyytinen in the newspaper *Helsingin Sanomat*, 1999-06-23, page A 10. Instructions (in Finnish) for publishing academic dissertations and others on paper and on the net given by the University of Helsinki can be found at http://ethesis.helsinki.fi/english.html.). For the electronic publication of Uralic science, the encoding of UPA characters in the BMP is a *conditio sine qua non*.

This requirement remains paramount. For more than a century, texts in the UPA have been published in a very accurate way. Since the 1980s, information technology has been used increasingly for this work. In the beginning the accuracy was very poor, but nowadays UPA texts can be published using IT with great accuracy. Print publication is not the issue, of course. The point of *encoding* the UPA is *processing*. Modern linguistics have reached the Uralicists, but the application of modern methods and computing is impossible as long as the material can not be encoded in a comprehensive and accurate way. Small caps phonetic characters are a vital, substantial, and meaningful part of the UPA, as are superscript and subscript characters; styled text cannot preserve the distinctions required.

At the same time, the Uralicist scientific community needs timely solutions. The repertoire of UPA is not so large that it would have to be split on different planes; indeed this would make implementation quite difficult at present. UPA texts are natural language texts, transcribed in a scientific way. In the Research Institute for the Languages of Finland, there are several projects in the pipeline which can be properly realized if only UPA is encoded. Elsewhere there are temporary UPA solutions making use the Private Zone for own encodings, but serious scientific work cannot be done on a global level in this manner. Those solutions are meant for temporary use until UPA is encoded properly. An example of the latter: <http://www.uni-koblenz.de/~uedb/uedb_aktuell/index.html>.

The requirement for BMP encoding of UPA character is also stated in N2419:

> **BMP encoding of UPA characters.** The UPA makes use of some 258 characters (see pp 19-28 below). The BMP contains 125 of these already, and we propose the addition of the 133 missing characters also to the BMP for simplicity's sake; it is logical to keep all phonetics characters in the BMP…. While it is commonplace to assume that characters proposed for the UCS are intended to be for plain text support, we state this requirement explicitly here because, of the 108 alphabetic characters proposed (8 combining diacritics, 17 "other" modifier letters, and 2 general punctuation marks are also proposed), 14 of them are "regular" letters, 34 of them are "small capital" letters, 49 of them are "superscript" modifier letters, and 9 of them are "subscript" modifier letters. An argument could be presented that fancy text can represent these – but the plain-text requirement for sorting and searching is paramount for the UPA (as it is for the IPA). Samples with near-minimal pairs may be found in Figs. 88, 89, 90, and 91. Further, the UCS already contains 8 IPA "small capital" and 6 IPA "superscript" modifier letters which are also used in the UPA.

It has been suggested that the small capital, superscript, and subscript UPA characters be encoded in the SMP instead of the BMP, but the arguments presented for preferring the SMP are not convincing, and work to the detriment of early implementation of these characters. Basically, the argument is that these characters are analogous to the mathematic alphabets which were placed in the SMP to "prevent people from misusing the characters for styled text". But the UPA characters are ordinary Latin letters, *intended* to be used in running text to write *words*, while the mathematical characters are *not*. Encoding and implementation of the UPA is needed *now* for serious scientific work. It would be a pity if this task would be made more difficult because of the assumed possibility that someone would use these encoded characters for some other purpose.

Here are some further arguments against encoding any UPA characters in the SMP and all of them in the BMP:

1. There are already small capital letters, subscripts, and supersripts in the BMP (used for *phonetic* purposes) and no one is making such misuse of *them*.

2. The proposal to encode in the SMP begs the question, without proof, as to why would anyone go to the trouble of making such substitutions just to produce text (without any particular processing requirements) when *everyone* already has wordprocessors which subscript, superscript, and small-capitalize the regular alphabet far more easily.

3. Unlike the mathematical alphabet letters, the UPA characters are not neatly laid out in alphabetical order making it easy to find and input them.

4. Unlike mathematical alphabet letters, which are used singly or in pairs or triplets with numbers and symbols and all kinds of special formatting, UPA letters are used, and are expected to be used, as *ordinary* Latin letters forming *words* in *ordinary* Latin text.

5. Putting UPA characters in the SMP would inconvenience (greatly) the primary user group, and would not deter people *really* wanting to misuse the characters in any way. SMP implementation is at its infancy and input, font encoding, and other aspects of processing are *not* at all widely implemented on any platform at present.

6. It is our view that all letters belonging to the Latin alphabet intended for use to represent natural language (whether in phonetic notation or not) should be encoded in the BMP. A block of Latin phonetic characters was roadmapped for, intended for these UPA characters as well as others. No artificial distinction between "styled clones" and other kinds of letters should be entertained so as to divide the UPA and force an artificial and unnecessary separation of these letters between two coded planes in the UCS.

Feedback on N2419 suggested that some Greek and Cyrillic charcters used in the UPA should be placed in the Phonetic Extensions block rather than in the Greek and Cyrillic blocks. We give below a proposed reorganization.

Annex A. Proposed character names

Modifier letters

02EF MODIFIER LETTER LOW DOWN ARROWHEAD
02F0 MODIFIER LETTER LOW UP ARROWHEAD
02F1 MODIFIER LETTER LOW LEFT ARROWHEAD
02F2 MODIFIER LETTER LOW RIGHT ARROWHEAD
02F3 MODIFIER LETTER LOW RING
02F4 MODIFIER LETTER MIDDLE GRAVE ACCENT
02F5 MODIFIER LETTER MIDDLE DOUBLE GRAVE ACCENT
02F6 MODIFIER LETTER MIDDLE DOUBLE ACUTE ACCENT
02F7 MODIFIER LETTER LOW TILDE
02F8 MODIFIER LETTER RAISED COLON
02F9 MODIFIER LETTER BEGIN HIGH TONE
02FA MODIFIER LETTER END HIGH TONE
02FB MODIFIER LETTER BEGIN LOW TONE
02FC MODIFIER LETTER END LOW TONE
02FD MODIFIER LETTER SHELF
02FE MODIFIER LETTER OPEN SHELF
02FF MODIFIER LETTER LOW LEFT ARROW

Combining diacritical marks

0350 COMBINING RIGHT ARROWHEAD ABOVE
0351 COMBINING LEFT HALF RING ABOVE
0352 COMBINING FERMATA
0353 COMBINING X BELOW
0354 COMBINING LEFT ARROWHEAD BELOW
0355 COMBINING RIGHT ARROWHEAD BELOW
0356 COMBINING RIGHT ARROWHEAD AND UP ARROWHEAD BELOW

Latin letters

1D00 LATIN LETTER SMALL CAPITAL A
1D01 LATIN LETTER SMALL CAPITAL AE
1D02 LATIN SMALL LETTER TURNED AE
- glyph can also have sideways orientation

1D03 LATIN LETTER SMALL CAPITAL BARRED B
1D04 LATIN LETTER SMALL CAPITAL C
1D05 LATIN LETTER SMALL CAPITAL D
1D06 LATIN LETTER SMALL CAPITAL ETH
1D07 LATIN LETTER SMALL CAPITAL E
1D08 LATIN SMALL LETTER TURNED OPEN E
1D09 LATIN SMALL LETTER TURNED I
1D0A LATIN LETTER SMALL CAPITAL J
1D0B LATIN LETTER SMALL CAPITAL K
1D0C LATIN LETTER SMALL CAPITAL L WITH STROKE
1D0D LATIN LETTER SMALL CAPITAL M
1D0E LATIN LETTER SMALL CAPITAL REVERSED N
1D0F LATIN LETTER SMALL CAPITAL O
1D10 LATIN LETTER SMALL CAPITAL OPEN O
1D11 LATIN SMALL LETTER SIDEWAYS O
1D12 LATIN SMALL LETTER SIDEWAYS OPEN O
1D13 LATIN SMALL LETTER SIDEWAYS O WITH STROKE
1D14 LATIN SMALL LETTER TURNED OE
- glyph can also have sideways orientation

1D15 LATIN LETTER SMALL CAPITAL OU
1D16 LATIN SMALL LETTER TOP HALF O
1D17 LATIN SMALL LETTER BOTTOM HALF O
1D18 LATIN LETTER SMALL CAPITAL P
- represents a semi-voiced [p]

1D19 LATIN LETTER SMALL CAPITAL REVERSED R
1D1A LATIN LETTER SMALL CAPITAL TURNED R
1D1B LATIN LETTER SMALL CAPITAL T
1D1C LATIN LETTER SMALL CAPITAL U
1D1D LATIN SMALL LETTER SIDEWAYS U
1D1E LATIN SMALL LETTER SIDEWAYS DIAERESIZED U
• glyph can also have turned orientation
1D1F LATIN SMALL LETTER SIDEWAYS TURNED M
1D20 LATIN LETTER SMALL CAPITAL V
1D21 LATIN LETTER SMALL CAPITAL W
1D22 LATIN LETTER SMALL CAPITAL Z
1D23 LATIN LETTER SMALL CAPITAL EZH
1D24 LATIN LETTER VOICED LARYNGEAL SPIRANT
1D25 LATIN LETTER AIN

Greek letters

1D26 GREEK LETTER SMALL CAPITAL GAMMA
1D27 GREEK LETTER SMALL CAPITAL LAMDA
1D28 GREEK LETTER SMALL CAPITAL PI
1D29 GREEK LETTER SMALL CAPITAL RHO
• represents a voiceless uvular trill
1D2A GREEK LETTER SMALL CAPITAL PSI

Cyrillic letter

1D2B CYRILLIC LETTER SMALL CAPITAL EL
• in italic style, the glyph is obliqued, not italicized

Latin modifier letters

1D2C MODIFIER LETTER CAPITAL A
1D2D MODIFIER LETTER CAPITAL AE
1D2E MODIFIER LETTER CAPITAL B
1D2F MODIFIER LETTER CAPITAL BARRED B
1D30 MODIFIER LETTER CAPITAL D
1D31 MODIFIER LETTER CAPITAL E
1D32 MODIFIER LETTER CAPITAL REVERSED E
1D33 MODIFIER LETTER CAPITAL G
1D34 MODIFIER LETTER CAPITAL H
1D35 MODIFIER LETTER CAPITAL I
1D36 MODIFIER LETTER CAPITAL J
1D37 MODIFIER LETTER CAPITAL K
1D38 MODIFIER LETTER CAPITAL L
1D39 MODIFIER LETTER CAPITAL M
1D3A MODIFIER LETTER CAPITAL N
1D3B MODIFIER LETTER CAPITAL REVERSED N
1D3C MODIFIER LETTER CAPITAL O
1D3D MODIFIER LETTER CAPITAL OU
1D3E MODIFIER LETTER CAPITAL P
1D3F MODIFIER LETTER CAPITAL R
1D40 MODIFIER LETTER CAPITAL T
1D41 MODIFIER LETTER CAPITAL U
1D42 MODIFIER LETTER CAPITAL W
1D43 MODIFIER LETTER SMALL A
1D44 MODIFIER LETTER SMALL TURNED A
1D45 MODIFIER LETTER SMALL ALPHA
1D46 MODIFIER LETTER SMALL TURNED AE
1D47 MODIFIER LETTER SMALL B
1D48 MODIFIER LETTER SMALL D
1D49 MODIFIER LETTER SMALL E
1D4A MODIFIER LETTER SMALL SCHWA
1D4B MODIFIER LETTER SMALL OPEN E

1D4C MODIFIER LETTER SMALL TURNED OPEN E
1D4D MODIFIER LETTER SMALL G
1D4E MODIFIER LETTER SMALL TURNED I
1D4F MODIFIER LETTER SMALL K
1D50 MODIFIER LETTER SMALL M
1D51 MODIFIER LETTER SMALL ENG
1D52 MODIFIER LETTER SMALL O
1D53 MODIFIER LETTER SMALL OPEN O
1D54 MODIFIER LETTER SMALL TOP HALF O
1D55 MODIFIER LETTER SMALL BOTTOM HALF O
1D56 MODIFIER LETTER SMALL P
1D57 MODIFIER LETTER SMALL T
1D58 MODIFIER LETTER SMALL U
1D59 MODIFIER LETTER SMALL SIDEWAYS U
1D5A MODIFIER LETTER SMALL TURNED M
1D5B MODIFIER LETTER SMALL V
1D5C MODIFIER LETTER SMALL AIN

Greek modifier letters

1D5D MODIFIER LETTER SMALL BETA
1D5E MODIFIER LETTER SMALL GREEK GAMMA
1D5F MODIFIER LETTER SMALL DELTA
1D60 MODIFIER LETTER SMALL GREEK PHI
1D61 MODIFIER LETTER SMALL CHI

Latin subscript letters

1D62 LATIN SUBSCRIPT SMALL LETTER I
1D63 LATIN SUBSCRIPT SMALL LETTER R
1D64 LATIN SUBSCRIPT SMALL LETTER U
1D65 LATIN SUBSCRIPT SMALL LETTER V

Greek subscript letters

1D66 GREEK SUBSCRIPT SMALL LETTER BETA
1D67 GREEK SUBSCRIPT SMALL LETTER GAMMA
1D68 GREEK SUBSCRIPT SMALL LETTER RHO
1D69 GREEK SUBSCRIPT SMALL LETTER PHI
1D6A GREEK SUBSCRIPT SMALL LETTER CHI

General punctuation

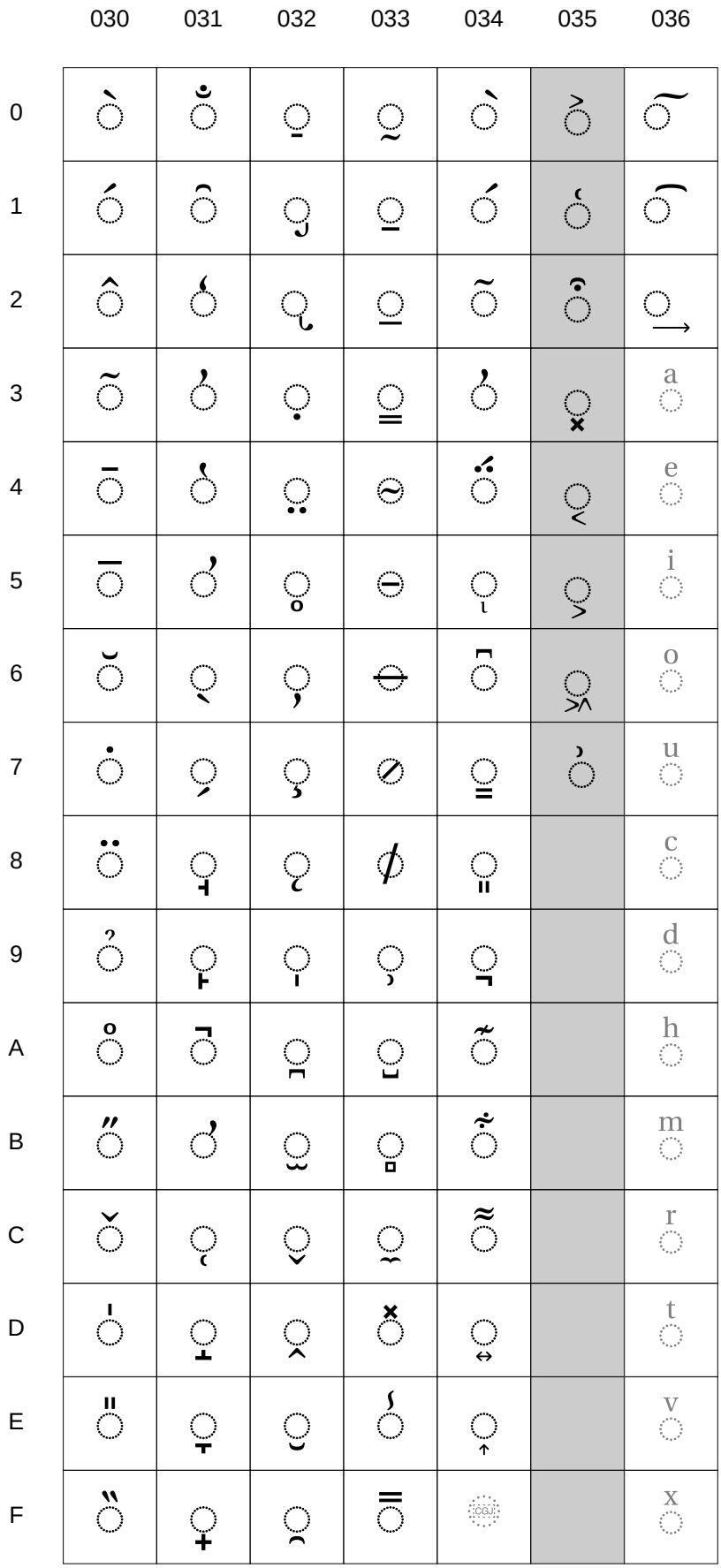
2053 SWUNG DASH
2054 INVERTED UNDERTIE

Row 02: SPACING MODIFIER LETTERS

	02B	02C	02D	02E	02F
0	ʰ	ʲ	˭	ˠ	˄
1	ʱ	˹	ˮ	ˡ	˂
2	ʲ	˂	˯	ˢ	˃
3	ʳ	˃	˰	ˣ	˅
4	ʴ	˄	˱	˺	ˊ
5	ʵ	˅	˲	˻	ˋ
6	ʶ	ˆ	˳	˼	ˎ
7	ʷ	ˇ	˴	˽	˜
8	ʸ	ˉ	˵	˾	˸
9	ʹ	ˍ	˶	˿	˩
A	ʺ	ˊ	˷	ˀ	˥
B	ʻ	ˋ	˸	ˁ	˧
C	ʼ	ˉ	˜	˅	˨
D	ʽ	ˍ	˝	˭	˫
E	ʾ	ˊ	˞	ˮ	ˬ
F	ʿ	ˋ	˟	˅	←

G = 00
P = 00

Row 03: COMBINING DIACRITICAL MARKS



G = 00
P = 00

Row 1D: PHONETIC EXTENSIONS

	1D0	1D1	1D2	1D3	1D4	1D5	1D6	1D7
0	A	Ɔ	V	D	T	m	φ	
1	Æ	ɔ	W	E	U	ŋ	χ	
2	ᴂ	ɔ	Z	Ǝ	W	o	i	
3	ʙ	ɵ	Ʒ	G	a	ɔ	r	
4	C	ᴂ	ƨ	H	ɐ	ɑ	u	
5	D	ɤ	ᴥ	I	ɑ	ʊ	v	
6	Đ	ɑ	Γ	J	æ	p	β	
7	E	ʊ	Λ	K	b	t	γ	
8	Ɜ	P	Π	L	d	u	ϱ	
9	Ɨ	Я	P	M	e	ɪ	φ	
A	J	ʁ	Ψ	N	ə	ɯ	χ	
B	K	T	Л	И	ɛ	v		
C	Ł	U	A	O	ɜ	ᴥ		
D	M	ɪ	Æ	ɤ	g	β		
E	И	ɨ	B	P	ɨ	γ		
F	O	ɯ	ʁ	R	k	δ		

G = 00
P = 00

Row 20: GENERAL PUNCTUATION

	200	201	202	203	204	205	206
0		‒	†	‰	⁀	⁀	
1		‒	‡	‱	⁁	⁂	
2		‒	•	′	⁂	⁒	
3		―	‣	″	⁃	⁓	
4		―	‧	‴	⁄	⁀	
5		―	‥	‵	⁅		
6		‖	…	‶	⁆		
7		‗	‧	‷	⁇	⁗	
8		‘		‸	⁈		
9		’		‹	⁉		
A		‚		›	⁊		
B		‛		※	⁋		
C		“		‼	⁌		
D		”		⁋	⁍		
E		„		⁻	⁎		
F		‟		⁔	⁏		

G = 00
P = 00