SC2/WG2 N3070

Doc Type:	**Working Group Document**
Title:	**Proposal for Encoding 3 Additional Characters of the Uralic Phonetic Alphabet**
Source:	**Klaas Ruppel, Jack Rueter, Erkki Kolehmainen**
Status:	**Expert Contribution**
Date:	**2006-04-07**

As stated before (N2958) the Uralic Phonetic Alphabet is an open scheme with a large number of theoretically possible special characters and combinations of base chracters and diacritical marks. A fairly comprehensive survey of characters in UPA texts (done in spring 2005) brought at light very few characters not yet encoded. But even after that more characters came across. This is due to the open nature of the UPA scheme. Annex B gives an overview over the principles of UPA and possible characters, which are not yet encoded, but could theoretically be found in UPA texts.

The encoding of the UPA characters has a high priority in Uralistics. There are plans and ongoing work for the electronic publication of existing dictionaries, texts, theses and other material. New publications are electronic-based or available only electronically. Among others the following lexical projects ongoing in Finland can be mentioned:

1. The Dictionary of Finnish Dialects (The Research Insitute for the Languages of Finland - RILF): switching from traditional to fully implemented electronic publishing.

2. The Dictionary of Karelian (RILF, Finno-Ugrian Society): making the published dictionary electronically available.

3. The Nielsen Dictionary of Northern Sami (Department of Finno-Ugrian Studies at the Helsinki University): publishing of e renewed electronic version.

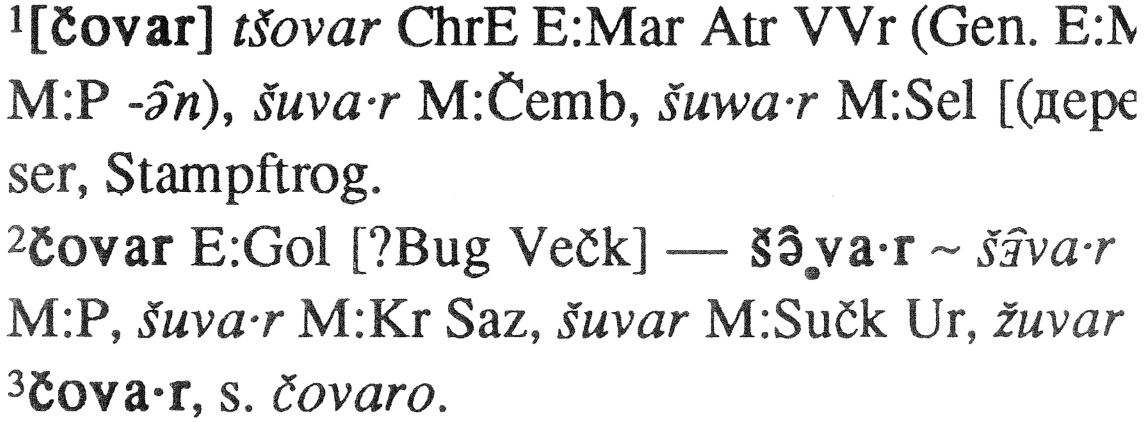
4. The Mordovian Dictionary (Finno-Ugrian Society, RILF): making the published dictionary electronically available.

5. Álgu - the Etymological Database for the Sami Languages (RILF): the web-based data base will be published by the end of the year.

In connection with the work on the Mordovian Dictionary the following characters not yet encoded were found:

1. LATIN LETTER SMALL CAPITAL INVERTED E

Example:

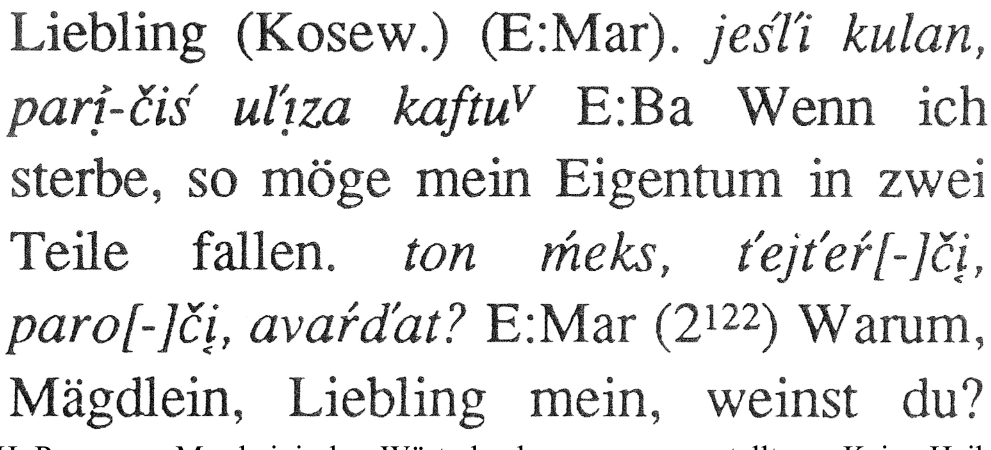


^{1}**[čovar]** *tšovar* ChrE E:Mar Atr VVr (Gen. E:M
M:P *-ə̂n*), *šuva·r* M:Čemb, *šuwa·r* M:Sel [(дере
ser, Stampftrog.
^{2}**čovar** E:Gol [?Bug Večk] — **šə̣̂va·r** ~ *šⱻ̂va·r*
M:P, *šuva·r* M:Kr Saz, *šuvar* M:Sučk Ur, *žuvar*
^{3}**čova·r**, s. *čovaro*.

H. Paasonens Mordwinisches Wörterbuch, zusammengestellt von Kaino Heikkilä. Band III. Suomalais-Ugrilainen Seura, Helsinki 1994. — Lexica Societatis Fenno-Ugricae XXIII, 3. P. 1549.

2. MODIFIER LETTER CAPITAL V

Example:



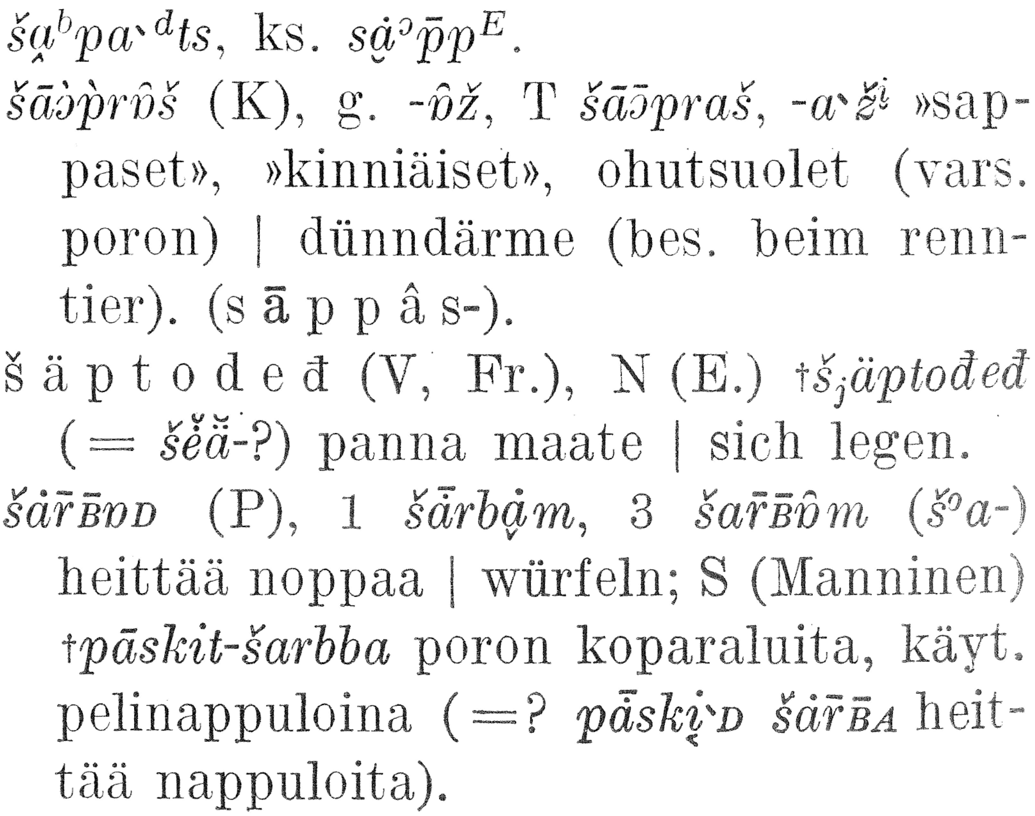
Liebling (Kosew.) (Œ:Mar). *jeśl'i kulan, pari̯-čiś ul'ı̣za kaftuⱽ* E:Ba Wenn ich sterbe, so möge mein Eigentum in zwei Teile fallen. *ton ḿeks, t'ejt'eŕ[-]či̯, paro[-]či̯, avaŕd'at?* E:Mar (2^{122}) Warum, Mägdlein, Liebling mein, weinst du?

H. Paasonens Mordwinisches Wörterbuch, zusammengestellt von Kaino Heikkilä. Band I. Suomalais-Ugrilainen Seura, Helsinki 1990. — Lexica Societatis Fenno-Ugricae XXIII, 1. P. XIV.

In connection with the Etymological Database for the Sami Languages (Álgu) the following character not yet encoded was found:

3. LATIN SUBSCRIPT SMALL LETTER J

Example:



ša̭^{b}pa`^{d}ts, ks. *sȧ̯^{ɔ}p̄p^{E}*.

šā̀ɔ̀prŵš (K), g. *-ŵž*, T *šāɔ̄praš*, *-a`ž̌^{i}* »sappaset», »kinniäiset», ohutsuolet (vars. poron) | dünndärme (bes. beim renntier). (s ā p p â s-).

š ä p t o d e đ (V, Fr.), N (E.) †*š$_{j}$äptođeđ* (= *šė̆ă-*?) panna maate | sich legen.

šǡr̄ʙᴅᴅ (P), 1 *šǡrbạ̌m*, 3 *šȧr̄ʙŵm* (*š^{o}a-*) heittää noppaa | würfeln; S (Manninen) †*pāskit-šarbba* poron koparaluita, käyt. pelinappuloina (=? *pǡskị̬`ᴅ šȧr̄ʙᴀ* heittää nappuloita).

T.I. Itkonen: Koltan- ja kuolanlapin sanakirja. Suomalais-Ugrilainen Seura, Helsinki 1958. — Lexica Societatis Fenno-Ugricae XV. P. 544.

We propose to add the 3 additional UPA character in the block Latin Extendend-C at the following code positions:

2C78	ⱻ	LATIN LETTER SMALL CAPITAL INVERTED E
2C79	ⱼ	LATIN SUBSCRIPT SMALL LETTER J
2C7A	ⱽ	MODIFIER LETTER CAPITAL V

Annex A: Proposal Summary Form

Annex B: Spotlight on UPA in the UCS

ISO/IEC JTC 1/SC 2/WG 2
PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646[1]

Please fill all the sections A, B and C below.

Please read Principles and Procedures Document (P & P) from http://www.dkuug.dk/JTC1/SC2/WG2/docs/principles.html for guidelines and details before filling this form.

Please ensure you are using the latest Form from http://www.dkuug.dk/JTC1/SC2/WG2/docs/summaryform.html.

See also http://www.dkuug.dk/JTC1/SC2/WG2/docs/roadmaps.html for latest *Roadmaps*.

A. Administrative

1. **Title:** *Proposal for Encoding 3 Additional Characters of the Uralic Phonetic Alphabet (UPA)*

2. Requester's name: *Klaas Ruppel, Jack Rueter, Erkki Kolehmainen*

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

4. Submission date: *2006-04-07*

5. Requester's reference (if applicable):

6. Choose one of the following:
This is a complete proposal: *YES*
(or) More information will be provided later:

B. Technical – General

1. Choose one of the following:
a. This proposal is for a new script (set of characters):
Proposed name of script:
b. The proposal is for addition of character(s) to an existing block: *YES*
Name of the existing block: *Phonetic Extensions / Latin Extended-C*

2. Number of characters in proposal: *3*

3. Proposed category (select one from below - see section 2.2 of P&P document):
A-Contemporary ____ B.1-Specialized (small collection) ____ B.2-Specialized (large collection) *X*
C-Major extinct ____ D-Attested extinct ____ E-Minor extinct ____
F-Archaic Hieroglyphic or Ideographic ____ G-Obscure or questionable usage symbols ____

4. Proposed Level of Implementation (1, 2 or 3) (see Annex K in P&P document): *3*
Is a rationale provided for the choice? *YES*
If Yes, reference: *UPA makes extensive use of Diacritical Marks*

5. Is a repertoire including character names provided? *YES*
a. If YES, are the names in accordance with the “character naming guidelines” in Annex L of P&P document? *YES*
b. Are the character shapes attached in a legible form suitable for review? *YES*

6. Who will provide the appropriate computerized font (ordered preference: True Type, or PostScript format) for publishing the standard? *Juhani Lehtiranta*
If available now, identify source(s) for the font (include address, e-mail, ftp-site, etc.) and indicate the tools used: *JL-types, jltypes@kolumbus.fi*

7. References:
a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided? *YES*
b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached? *YES*

8. Special encoding issues:
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)? *NO*

9. Additional Information:

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 http://www.unicode.org/Public/UNIDATA/UCD.html and associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for inclusion in the Unicode Standard.

[1] Form number: N3002-F (Original 1994-10-14; Revised 1995-01, 1995-04, 1996-04, 1996-08, 1999-03, 2001-05, 2001-09, 2003-11, 2005-01, 2005-09, 2005-10)

C. Technical - Justification

1. Has this proposal for addition of character(s) been submitted before? *NO*

If YES explain

2. 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*

If YES, with whom? *Juhani Lehtiranta, Álgu Project (RILF)*

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? *YES*

Reference: *this proposal*

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

Reference:

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

If YES, where? Reference: *Álgu Project (RILF), Finno-Ugrian Society*

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

If YES, is a rationale provided? *YES*

If YES, reference: *The characters are part of UPA*

7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)? *NO*

8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence? *NO*

If YES, is a rationale for its inclusion provided?

If YES, reference:

9. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters? *NO*

If YES, is a rationale for its inclusion provided?

If YES, reference:

10. Can any of the proposed character(s) be considered to be similar (in appearance or function) to an existing character? *NO*

If YES, is a rationale for its inclusion provided?

If YES, reference:

11. Does the proposal include use of combining characters and/or use of composite sequences? *NO*

If YES, is a rationale for such use provided?

If YES, reference:

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

If YES, reference:

12. Does the proposal contain characters with any special properties such as control function or similar semantics? *NO*

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

13. Does the proposal contain any Ideographic compatibility character(s)? *NO*

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

If YES, reference:

1	2	3	4	5	6	7	8	9	10	11	12	13
r			ɹ	ᴟ	ʀ	ʁ	ʳ	ʴ	ᴿ	ʶ	ᵣ	other

Spotlight on UPA in the UCS

The transcription scheme of the Uralic Phonetic Alphabet (UPA) is used for presenting linguistic data of languages belonging to the Uralic, Turkic and Mongolic language families. The principles of UPA can be described as following:

1. UPA data is traditionally presented in *Italics.*
2. Basically UPA makes use of Latin small letters with some letters borrowed mainly from the Greek and Cyrillic script.
3. Diacritical marks are used to specify the sound presented by the basic letters. Usually not more than two diacritics below and three diacritics above occur.
4. There is a number of features which can be expressed by using a derivate from the basic small letter:
   - 4.1 Reduceness is expressed by using an upside down turned letter. In some cases (when the upside down letter would not be correctly recognizable) a sideways letter is used. In some other cases an upside down letter is used for a different, not reduced sound, for which then sideways letter stands for reduceness, too.
   - 4.2 Small capitals express that a normally voiced sound is half-unvoiced and a normally unvoiced sound is half-voiced.
   - 4.3 Very short sounds are presented by superscript letters (basic letters, upside down turned letters, small capitals etc.).
   - 4.4 Side articulation is shown by subscript letters.

It is obvious that not all possible derivates (4.1–4.4) are sensible for each letter. However, at least in theoretic writings there can occur the need to refer to sounds not ever presented in any prior publication.

In the following overview derivates which are not foreseen to be usable are indicated with a dash.

NB. In connection with the original UPA proposal (which had its incarnation as the Phonetic Extensions block) we decided to use Superscript Capital letters instead of Superscript Small Caps, which would have followed the inner UPA logic. We concluded that the distinction between Capital letters and Small Caps in Superscript would be quite confusing and hardly recognizable in reality. We now notice that Superscript Small Caps have been added to the standard: 1DA6, 1DA7, 1DAB, 1DB0, 1DB8. Consequently, we need to provide specific guidance in order to avoid confusion among the users of UPA.

1	2	3	4	5	6	7	8	9	10	11	12	13
r			ɹ	ᴟ	ʀ	ᴚ	ʳ	ʴ	ᴿ	ᴚ	ᵣ	other

Matrix of the UPA scheme

v1.0 2006-04-07

1	2	3	4	5	6	7	8	9	10	11	12	13

1 Basic Character
2 Code Position
3 Name of Basic Character
4 Corresponding Turned Character
5 Corresponding Alternative Turned/Inverted Character
6 Corresponding Small Cap
7 Corresponding Turned/Inverted Small Cap
8 Corresponding Modifier Letter/Superscript
Corresponding Turned Modifier Letter/Superscript 9
Corresponding Cap Modifier Letter/Superscript 10
Corresponding Turned Capital Modifier Letter/Superscript 11
Corresponding Subscript 12
Other/Reference 13

a	0061	LATIN SMALL LETTER A	0250	-	1D00		1D43	1D44	1D2C		2090	
b	0062	LATIN SMALL LETTER B	-	-	0299		1D47	-	1D2E			→ 0180
c	0063	LATIN SMALL LETTER C	0254	1D12	1D04	1D10	1D9C	1D53				
d	0064	LATIN SMALL LETTER D	-	-	1D05		1D48	-	1D30	-		→ 0111
e	0065	LATIN SMALL LETTER E	0259		1D07	?2C78	1D49	1D4A	1D31	1D32	2091	2094

1	2	3	4	5	6	7	8	9	10	11	12	13
r			ɹ	ᴟ	ʀ	ʁ	ʳ	ʴ	ᴿ	ʶ	ᵣ	other

f	0066	LATIN SMALL LETTER F	-	-			1DA0	-				
g	0067	LATIN SMALL LETTER G	-	-	0262		1D4D	-	1D33			
h	0068	LATIN SMALL LETTER H			029C		02B0		1D34			→ 0266
i	0069	LATIN SMALL LETTER I	1D09	-	026A	-	2071	1D4E	1D35	-	1D62	
j	006A	LATIN SMALL LETTER J	-	-	1D0A	-	02B2	-	1D36	-	?2C79	
k	006B	LATIN SMALL LETTER K	029E	-	1D0B	-	1D4F		1D37	-		
l	006C	LATIN SMALL LETTER L	-	-	029F	-	02E1	-	1D38	-		→ 0142 019A 026B
m	006D	LATIN SMALL LETTER M	026F	1D1F	1D0D	-	1D50	1D5A	1D39	-		
n	006E	LATIN SMALL LETTER N	-	-	0274	1D0E	207F	-	1D3A	1D3B		→ 014B
o	006F	LATIN SMALL LETTER O	1D11	-	1D0F		1D52		1D3C		2092	→ 00F8 0275 1D16 1D17
p	0070	LATIN SMALL LETTER P	-	-	1D18	-	1D56	-	1D3E	-		
q	0071	LATIN SMALL LETTER Q	-	-	-	-		-	-	-		

1	2	3	4	5	6	7	8	9	10	11	12	13
r			ɹ	ᴟ	ʀ	ʁ	ʳ	ʴ	ᴿ	ʶ	ᵣ	other

r	0072	LATIN SMALL LETTER R	0279	-	0280	0281	02B3	02B4	1D3F	02B6	1D63	1D19 1D1A
s	0073	LATIN SMALL LETTER S	-	-		-	02E2	-		-		
t	0074	LATIN SMALL LETTER T	0287	-	1D1B	-	1D57		1D40	-		
u	0075	LATIN SMALL LETTER U	1D1D	-	1D1C		1D58	1D59	1D41		1D64	→ 00FC
v	0076	LATIN SMALL LETTER V	028C	-	1D20		1D5B	1DBA	?2C7A		1D65	
w	0077	LATIN SMALL LETTER W	-	-	1D21	-	02B7	-	1D42	-		
x	0078	LATIN SMALL LETTER X	-	-	-	-	02E3	-	-	-	2093	
y	0079	LATIN SMALL LETTER Y	028E	-	028F	-	02B8			-		
z	007A	LATIN SMALL LETTER Z	-	-	1D22	-	1DBB	-		-		
æ	00E6	LATIN SMALL LETTER AE	1D02		1D01			1D46	1D2D			
ø	00F8	LATIN SMALL LETTER O WITH STROKE	1D13	-								
ü	00FC	LATIN SMALL LETTER U WITH DIAERESIS	1D1E	-								
đ	0111	LATIN SMALL LETTER D WITH STROKE	-	-	1D06	-		-				
ł	0142	LATIN SMALL LETTER L WITH STROKE	-	-	1D0C	-		-		-		

1	2	3	4	5	6	7	8	9	10	11	12	13
r			ɹ	ᴟ	ʀ	ʁ	ʳ	ʴ	ᴿ	ʶ	ᵣ	other
ŋ	014B	LATIN SMALL LETTER ENG	-	-	-	-	1D51	-	-	-		
œ	0153	LATIN SMALL LIGATURE OE	1D14		0276							
ƀ	0180	LATIN SMALL LETTER B WITH STROKE	-	-	1D03	-		-	1D2F			
ƚ	019A	LATIN SMALL LETTER WITH BAR	-	-	-	-		-	-	-		
ɑ	0251	LATIN SMALL LETTER ALPHA	0252	-			1D45	1D9B				
ɛ	025B	LATIN SMALL LETTER OPEN E	1D08				1D4B	1D4C				
ɦ	0226	LATIN SMALL LETTER H WITH HOOK			-	-	02B1		-	-		
ɫ	026B	LATIN SMALL LETTER L WITH MIDDLE TILDE	-	-	-	-		-	-	-		
ɵ	0275	LATIN SMALL LETTER BARRED O	-	-		-	1DB1	-		-		
ʒ	0292	LATIN SMALL LETTER EZH	-	-	1D23	-	1DBE	-		-		
β	03B2	GREEK SMALL LETTER BETA	-	-			1D5D	-			1D66	
γ	03B3	GREEK SMALL LETTER GAMMA	-	-	1D26		1D5E	-			1D67	
δ	03B4	GREEK SMALL LETTER DELTA	-	-			1D5F	-				
λ	03BB	GREEK SMALL LETTER LAMDA	-	-	1D27	-		-		-		

1	2	3	4	5	6	7	8	9	10	11	12	13
r			ɹ	ᴟ	ʀ	ʁ	ʳ	ʴ	ᴿ	ʶ	ᵣ	other

π	03C0	GREEK SMALL LETTER PI	-	-	1D28	-		-		-		
ρ	03C1	GREEK SMALL LETTER RHO	-	-	1D29	-		-		-	1D68	
τ	03C4	GREEK SMALL LETTER TAU	-	-		-		-		-		
φ	03C6	GREEK SMALL LETTER PHI	-	-		-	1D60	-		-	1D69	
χ	03C7	GREEK SMALL LETTER CHI	-	-		-	1D61	-		-	1D6A	
ψ	03C8	GREEK SMALL LETTER PSI	-	-	1D2A	-		-		-		
ϑ	03D1	GREEK THETA SYMBOL	-	-		-		-		-		
л	043B	CYRILLIC SMALL LETTER EL	-	-	1D2B	-		-		-		
э	044D	CYRILLIC SMALL LETTER E	-	-		-		-		-		
ᴕ	1D15	LATIN LETTER SMALL CAPITAL OU	-	-	N/A	-		-	1D3D	-		
ᴖ	1D16	LATIN SMALL LETTER TOP HALF O	-	-	-	-	1D54	-	-	-		
ᴗ	1D17	LATIN SMALL LETTER BOTTOM HALF O	-	-	-	-	1D55	-	-	-		
ᴤ	1D24	LATIN LETTER VOICED LARYNGEAL SPIRANT	-	-	-	-		-	-	-		
ᴥ	1D25	LATIN LETTER AIN	-	-	-	-	1D5C	-	-	-		

1	2	3	4	5	6	7	8	9	10	11	12	13
r			ɹ	ɻ	ʀ	ʁ	ʳ	ʴ	ᴿ	ʶ	ᵣ	other

ᵼ	1D7C	LATIN SMALL LETTER IOTA WITH STROKE	-	-	-	-		-	-	-		
□	2C77	LATIN SMALL LETTER TAILLESS PHI	-	-	-	-		-	-	-		