

Report on the ad hoc re “Teuthonista” (SC2/WG2 N4081) held during the SC2/WG2 meeting at Helsinki, 2011 June 7/8

Michael Everson (tables), Karl Pentzlin (text)

The following issues regarding the characters proposed in SC2/WG2 N4081 were addressed and solved, resulting in the agreed character repertoire shown at the end of this document:

1. Character stacking and raised letters.

A feature of Teuthonista is the stacking of two characters to denote an intermediate sound, like

p over b: $\overset{p}{b}$ or: $\overset{p}{\underset{b}{}}$

(i.e., there are examples for using full size characters as well as smaller size characters).

There is consensus that both cases can be represented by combining small letters (like the COMBINING LATIN SMALL LETTER P in this example), which usually yields a glyph smaller than the base letter it is applied to.

Similar applies to raised letters. As an example, the "half geminate" which appears in some material as $\overset{t}{t}$ will be represented as $\overset{t}{t}$, using a pair of modifier letter + common letter.

2. Parenthesized diacritical marks and characters

There will be no combinations of diacritics or letters surrounded by small parentheses as atomic characters. Instead, there will be four diacritical marks:

COMBINING PARENTHESES ABOVE
COMBINING DOUBLE PARENTHESES ABOVE
COMBINING PARENTHESES OVERLAY
COMBINING PARENTHESES BELOW

While all combining marks above get the Canonical Combining Class value 230, all other combining marks (with the exception of COMBINING PARENTHESES OVERLAY, which gets "TBD" now as the value will be determined by the UTC) get the same Canonical Combining Class value 220 marking them as "below". This includes the centralization marks. Thus, it is ensured that all combining marks will be processed in the input order.

The parentheses will affect the character which precedes them, be it a diacritic or a base character. The exact placement (which is usually not symmetrical to the vertical axis of the base character e.g. in the case of an eng or a centralization mark) is to be determined by the font.

There will be two explanation paragraphs included into the text of the Standard.

One explains that the centralization strokes may be placed sideways to other combining marks being placed below the place letters (comparable of the handling of Vietnamese diacritics), in spite of having the same canonical combination class which otherwise denotes vertical stacking. The other one deals with the combining parenthesis pairs.

As an example, the Teuthonista combined letter:



is represented by the following sequence:

LATIN SMALL LETTER A
COMBINING STRONG CENTRALIZATION MARK BELOW
COMBINING PARENTHESES BELOW
COMBINING OPEN MARK BELOW
COMBINING PARENTHESES BELOW

3. Lenis f

The LATIN SMALL LETTER LENIS F as proposed in N4081 is a character different from the usual small f. While this has a bottom hook in italic fonts, the "lenis f" has not. As for the Unicode tables a Roman form is to be provided, this has to be invented and to be distinguishable from the normal f, by cutting its bottom and making it float above the baseline).

It is not adequate to use U+0192 LATIN SMALL LETTER F WITH HOOK for the "normal f" while using U+0066 LATIN SMALL LETTER LENIS F for the "ordinary f", as in cannot excluded that anybody inserts a citation from a corpus containing a "lenis f" into a text which already contains the letter f encoded as "usual f" from other sources, then being forced to re-encode all "f"s in their text.

4. Final Barline

There will be no new punctuation mark "Final Barline" encoded. Instead, U-1D102 MUSICAL SYMBOL FINAL BARLINE and similar symbols can be used in running text. Into fonts intended for linguistic use, such symbols can be included with a design compatible with the vertical lines encoded as punctuation marks.

5. Latin small letter g with low stroke

It is decided that the letter proposed in N4081 as LATIN LETTER SCRIPT G WITH LOW STROKE can be represented by U+01E5 LATIN SMALL LETTER G WITH STROKE, as the form of this letter in italic fonts usually resemble the required form.

6. The open mark

The "open mark below" is a character different from the ogonek, in spite of the superficial similarity especially in some inferior fonts used for some early Americanistic texts. As the geographical area of German dialectology largely overlaps with the area where Polish place names are used, which may contain an ogonek, both diacritical marks in fact occur in the same fields of databases. A unification would prevent the possibility to search selectively for open marks.

7. The "Latin small letter lunate epsilon"

The encoding of this letter, which is contrastively used to the open e in some Swiss material from the first decades of the 20th century, was deferred (and will possibly be dealt by later ballot comments or supplemental proposals). As long as Greek letters are used within IPA, U+03F5 GREEK LUNATE EPSILON SYMBOL can be considered as a replacement, although its glyph is inferior for this purpose (the letter used in the Swiss material in fact is a common "e" where the lower right part of the bowl is removed on the metal types).

8. Letters with high caron overlay

The encoding of the letters d/k/t with high caron overlay were deferred. It was pointed out that while these letters can be regarded as combinations of base letters with caron, such combinations are usually displayed by glyphs showing the caron besides the letter stem by a comma-like form. As the geographical area of German dialectology largely overlaps with the area where Czech place names are used, such place names can occur side by side with text using dialectal orthography, thus even if the letters are considered as combinations of base letter + caron, the proposed forms are to be coded separately even if considered as presentation forms.

As these letters are not used by the conventions applied to ongoing larger scientific projects, it was agreed to defer these letters to the encoding of Landsmålsalfabetet, which contains more such letters.

9. X-like letters

N4081 contains seven letters (at AB58...AB5E) which show x-like forms. As Teuthonista texts, as common for phonetics, usually are printed with italic types, the issue was raised that the straight glyphs provided as Roman types in N4081 are not appropriate. It was agreed that the glyphs will be replaced by ones which resemble more the italic forms.

As such a glyph redesign takes time, it was agreed that the ballot documents may show the glyphs shown in N4081, to be replaced later.

10. Roadmap

It was agreed that there is no longer a need to differentiate between the two blocks now in the Roadmap to the BMP:

Latin Extended-E: AB30...AB8F

Phonetic Extensions Extended B: AB90...ABBF

Thus, it was agreed to fuse them to a single block:

Latin Extended-E: AB30...ABBF

11. Agreed Repertoire

After having discussed the other letters and their shapes individually, the repertoire (containing a total of 85 characters) shown in the tables below was agreed upon.

It consists of:













































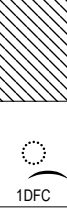















15 combining characters in the new block "Combining Diacritical Marks Extended";

14 combining characters in the block "Combining Diacritical Marks Supplement";

8 characters in the block "Latin Extended-D";

48 characters in the new block "Latin Extended-E".

	1AB	1AC	1AD	1AE	1AF
0	 1AB0				
1	 1AB1				
2	 1AB2				
3	 1AB3				
4	 1AB4				
5	 1AB5				
6	 1AB6				
7	 1AB7				
8	 1AB8				
9	 1AB9				
A	 1ABA				
B	 1ABB				
C	 1ABC				
D	 1ABD				
E	 1ABE				
F					

	1DC	1DD	1DE	1DF
0	 1DC0	 1DD0	 1DE0	 1DF0
1	 1DC1	 1DD1	 1DE1	 1DF1
2	 1DC2	 1DD2	 1DE2	 1DF2
3	 1DC3	 1DD3	 1DE3	 1DF3
4	 1DC4	 1DD4	 1DE4	 1DF4
5	 1DC5	 1DD5	 1DE5	
6	 1DC6	 1DD6	 1DE6	
7	 1DC7	 1DD7	 1DE7	
8	 1DC8	 1DD8	 1DE8	
9	 1DC9	 1DD9	 1DE9	
A	 1DCA	 1DDA	 1DEA	
B	 1DCB	 1ddb	 1DEB	
C	 1DCC	 1DDC	 1DEC	
D	 1DCD	 1DDD	 1DED	 1DFD
E	 1DCE	 1DDE	 1DEE	 1DFE
F	 1DCF	 1DDF	 1DEF	 1DFF

Used for Ancient Greek

These are used as editorial signs for Ancient Greek to indicate scribal deletion of erroneous accent marks.

- 1DC0 ◌̧ COMBINING DOTTED GRAVE ACCENT
→ 1FED ͂ greek dialytika and varia
- 1DC1 ◌̧ COMBINING DOTTED ACUTE ACCENT
→ 0344 ◌̧ combining greek dialytika tonos
→ 1FEE ͂ greek dialytika and oxia

Miscellaneous marks

- 1DC2 ◌̨ COMBINING SNAKE BELOW
- 1DC3 ◌̨ COMBINING SUSPENSION MARK
• Glagolitic
→ 0306 ◌̨ combining breve

Contour tone marks

- 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

Miscellaneous mark

- 1DCA ◌̨ COMBINING LATIN SMALL LETTER R BELOW

Contour tone marks

- 1DCB ◌̨̇ COMBINING BREVE-MACRON
• Lithuanian dialectology
- 1DCC ◌̨̇ COMBINING MACRON-BREVE
• Lithuanian dialectology

Double diacritic

- 1DCD ◌̨̇ COMBINING DOUBLE CIRCUMFLEX ABOVE

Medievalist additions

- 1DCE ◌̨ COMBINING OGONEK ABOVE
- 1DCF ◌̨ COMBINING ZIGZAG BELOW
- 1DD0 ◌̨ COMBINING IS BELOW
- 1DD1 ◌̨ COMBINING UR ABOVE
- 1DD2 ◌̨ COMBINING US ABOVE

Medieval superscript letter diacritics

- 1DD3 ◌̨ COMBINING LATIN SMALL LETTER FLATTENED OPEN A ABOVE
- 1DD4 ◌̨ COMBINING LATIN SMALL LETTER AE
- 1DD5 ◌̨ COMBINING LATIN SMALL LETTER AO
- 1DD6 ◌̨ COMBINING LATIN SMALL LETTER AV
- 1DD7 ◌̨ COMBINING LATIN SMALL LETTER C CEDILLA
- 1DD8 ◌̨ COMBINING LATIN SMALL LETTER INSULAR D
- 1DD9 ◌̨ COMBINING LATIN SMALL LETTER ETH
- 1DDA ◌̨ COMBINING LATIN SMALL LETTER G
- 1ddb ◌̨ COMBINING LATIN LETTER SMALL CAPITAL G
- 1DDC ◌̨ COMBINING LATIN SMALL LETTER K
- 1DDD ◌̨ COMBINING LATIN SMALL LETTER L
- 1DDE ◌̨ COMBINING LATIN LETTER SMALL CAPITAL L
- 1DDF ◌̨ COMBINING LATIN LETTER SMALL CAPITAL M
- 1DE0 ◌̨ COMBINING LATIN SMALL LETTER N
- 1DE1 ◌̨ COMBINING LATIN LETTER SMALL CAPITAL N
- 1DE2 ◌̨ COMBINING LATIN LETTER SMALL CAPITAL R
- 1DE3 ◌̨ COMBINING LATIN SMALL LETTER R ROTUNDA
- 1DE4 ◌̨ COMBINING LATIN SMALL LETTER S
- 1DE5 ◌̨ COMBINING LATIN SMALL LETTER LONG S
- 1DE6 ◌̨ COMBINING LATIN SMALL LETTER Z

Superscript letter diacritics for German dialectology

- 1DE7 ◌̨ COMBINING LATIN SMALL LETTER ALPHA
- 1DE8 ◌̨ COMBINING LATIN SMALL LETTER B
- 1DE9 ◌̨ COMBINING LATIN SMALL LETTER BETA
- 1DEA ◌̨ COMBINING LATIN SMALL LETTER SCHWA
- 1DEB ◌̨ COMBINING LATIN SMALL LETTER F
- 1DEC ◌̨ COMBINING LATIN SMALL LETTER L WITH DOUBLE MIDDLE TILDE
- 1DED ◌̨ COMBINING LATIN SMALL LETTER O WITH LIGHT CENTRALIZATION STROKE
- 1DEE ◌̨ COMBINING LATIN SMALL LETTER P
- 1DEF ◌̨ COMBINING LATIN SMALL LETTER ESH
- 1DF0 ◌̨ COMBINING LATIN SMALL LETTER U WITH LIGHT CENTRALIZATION STROKE
- 1DF1 ◌̨ COMBINING LATIN SMALL LETTER W
- 1DF2 ◌̨ COMBINING LATIN SMALL LETTER A WITH DIAERESIS
- 1DF3 ◌̨ COMBINING LATIN SMALL LETTER O WITH DIAERESIS
- 1DF4 ◌̨ COMBINING LATIN SMALL LETTER U WITH DIAERESIS

Additional mark

- 1DFC ◌̨ COMBINING DOTTED DOUBLE INVERTED BREVE BELOW
- 1DFD ◌̨ COMBINING ALMOST EQUAL TO BELOW

Additional marks for the Uralic Phonetic Alphabet

- 1DFE ◌̨ COMBINING LEFT ARROWHEAD ABOVE
- 1DFF ◌̨ COMBINING RIGHT ARROWHEAD AND DOWN ARROWHEAD BELOW

	A72	A73	A74	A75	A76	A77	A78	A79	A7A	A7B	A7C	A7D	A7E	A7F
0														
1														
2														
3														
4														
5														
6														
7														
8								ƒ A798						
9								f A799						
A								Ɔ A79A						
B								ɸ A79B						
C								Ɔ A79C						
D								ɸ A79D						
E								Ɔ A79E						
F								ɸ A79F						

Additional letters**Archaic letters for Ewe**

A798	ƒ	LATIN CAPITAL LETTER F WITH STROKE
A799	ƒ	LATIN SMALL LETTER F WITH STROKE

- old Ewe orthography
- also used in German dialectology

Archaic letters for Volapük

A79A	Ɔ	LATIN CAPITAL LETTER VOLAPUK AE
A79B	ɔ	LATIN SMALL LETTER VOLAPUK AE
A79C	Ɔ	LATIN CAPITAL LETTER VOLAPUK OE
A79D	ɔ	LATIN SMALL LETTER VOLAPUK OE
A79E	Ɔ	LATIN CAPITAL LETTER VOLAPUK UE
A79F	ɔ	LATIN SMALL LETTER VOLAPUK UE

	AB3	AB4	AB5	AB6	AB7	AB8	AB9	ABA	ABB
0	Ⓔ AB30	Ⓔ AB40	Ⓔ AB50						
1	Ⓔ AB31	Ⓔ AB41	Ⓔ AB51						
2	Ⓔ AB32	Ⓔ AB42	Ⓔ AB52						
3	Ⓔ AB33	Ⓔ AB43	Ⓔ AB53						
4	Ⓔ AB34	Ⓔ AB44	Ⓔ AB54						
5	Ⓔ AB35	Ⓔ AB45	Ⓔ AB55						
6	Ⓔ AB36	Ⓔ AB46	Ⓔ AB56						
7	Ⓔ AB37	Ⓔ AB47	Ⓔ AB57						
8	Ⓔ AB38	Ⓔ AB48	Ⓔ AB58						
9	Ⓔ AB39	Ⓔ AB49	Ⓔ AB59						
A	Ⓔ AB3A	Ⓔ AB4A	Ⓔ AB5A						
B	Ⓔ AB3B	Ⓔ AB4B	Ⓔ AB5B						
C	Ⓔ AB3C	Ⓔ AB4C	Ⓔ AB5C						
D	Ⓔ AB3D	Ⓔ AB4D	Ⓔ AB5D						
E	Ⓔ AB3E	Ⓔ AB4E	Ⓔ AB5E						
F	Ⓔ AB3F	Ⓔ AB4F	Ⓔ AB5F						

Letters for German dialectology

AB30	Ɑ	LATIN SMALL LETTER BARRED ALPHA
AB31	Ɱ	LATIN SMALL LETTER A REVERSED-SCHWA
AB32	Ɐ	LATIN SMALL LETTER BLACKLETTER E
AB33	Ɒ	LATIN SMALL LETTER BARRED E
AB34	ⱱ	LATIN SMALL LETTER E WITH FLOURISH
AB35	f	LATIN SMALL LETTER LENIS F → 0066 f latin small letter f
AB36	g	LATIN SMALL LETTER SCRIPT G WITH CROSSED-TAIL
AB37	ł	LATIN SMALL LETTER L WITH INVERTED LAZY S
AB38	ł̇	LATIN SMALL LETTER L WITH DOUBLE MIDDLE TILDE
AB39	ł̈	LATIN SMALL LETTER L WITH MIDDLE RING
AB3A	Ⱳ	LATIN SMALL LETTER M WITH CROSSED-TAIL
AB3B	ⱳ	LATIN SMALL LETTER N WITH CROSSED-TAIL
AB3C	ⱴ	LATIN SMALL LETTER ENG WITH CROSSED- TAIL
AB3D	o	LATIN SMALL LETTER BLACKLETTER O
AB3E	ȯ	LATIN SMALL LETTER BLACKLETTER O WITH STROKE
AB3F	ö	LATIN SMALL LETTER OPEN O WITH STROKE
AB40	œ	LATIN SMALL LETTER INVERTED OE = latin small letter o reversed-schwa
AB41	œ̈	LATIN SMALL LETTER TURNED OE WITH STROKE
AB42	œ̇	LATIN SMALL LETTER TURNED OE WITH HORIZONTAL STROKE
AB43	o̅	LATIN SMALL LETTER TURNED O OPEN-O
AB44	o̅̇	LATIN SMALL LETTER TURNED O OPEN-O WITH STROKE
AB45	ʀ	LATIN SMALL LETTER STIRRUP R
AB46	ʀ̇	LATIN LETTER SMALL CAPITAL R WITH RIGHT LEG
AB47	ṙ	LATIN SMALL LETTER R WITHOUT HANDLE
AB48	r̈	LATIN SMALL LETTER DOUBLE R
AB49	r̊	LATIN SMALL LETTER R WITH CROSSED-TAIL
AB4A	r̊̇	LATIN SMALL LETTER DOUBLE R WITH CROSSED-TAIL
AB4B	ʀ̇	LATIN SMALL LETTER SCRIPT R
AB4C	ʀ̇̈	LATIN SMALL LETTER SCRIPT R WITH RING
AB4D	ʃ	LATIN SMALL LETTER BASELINE ESH
AB4E	u̇	LATIN SMALL LETTER U WITH SHORT RIGHT LEG
AB4F	u̇̈	LATIN SMALL LETTER U BAR WITH SHORT RIGHT LEG
AB50	ui̇	LATIN SMALL LETTER UI
AB51	uï	LATIN SMALL LETTER TURNED UI
AB52	u̇̈	LATIN SMALL LETTER U WITH LEFT HOOK
AB53	χ	LATIN SMALL LETTER STRETCHED X → 03C7 χ greek small letter chi → A7AF latin small letter chi
AB54	χ̇	LATIN SMALL LETTER STRETCHED X WITH LOW RIGHT RING
AB55	χ̈	LATIN SMALL LETTER STRETCHED X WITH LOW LEFT SERIF
AB56	χ̉	LATIN SMALL LETTER X WITH LOW RIGHT RING
AB57	χ̊	LATIN SMALL LETTER X WITH LONG LEFT LEG
AB58	χ̊̇	LATIN SMALL LETTER X WITH LONG LEFT LEG AND LOW RIGHT RING
AB59	χ̊̈	LATIN SMALL LETTER X WITH LONG LEFT LEG WITH SERIF
AB5A	ẏ	LATIN SMALL LETTER Y WITH SHORT RIGHT LEG

Modifier letters for German dialectology

AB5B	ż	MODIFIER BREVE WITH INVERTED BREVE
AB5C	ḧ	MODIFIER LETTER SMALL HENG
AB5D	ł̈	MODIFIER LETTER SMALL L WITH DOUBLE MIDDLE TILDE
AB5E	ł̈̇	MODIFIER LETTER SMALL L WITH INVERTED LAZY S
AB5F	u̇̈	MODIFIER LETTER SMALL U WITH LEFT HOOK