Towards a Unicode Encoding of Teuthonista

Presentation held at the Workshop
"**Transcription Systems and UNICODE – Special Characters in a Standardized World**"
Vienna (Austria), 2008-11-05

Karl Pentzlin
http://pentzlin.com

Abstract *(added for the printed version)*

Teuthonista is a German dialect writing system employing numerous combinations of letters and diacritical marks stacked vertically or even side-/downwards to its symbols. Although this is not uncommon for some South Asian scripts already encoded in Unicode, this is uncommon for the Latin script in its extent it is used for Teuthonista.

Thus, a proposal to encode Teuthonista must convince the Unicode Consortium that Teuthonista is in fact a set of characters (i.e. a valid addition to the Latin script), rather than being a notation system like musical notes.

To accomplish this, there an exhaustive list of the building blocks of Teuthonista must be supplied, together with the rules how these building blocks are used to combine to the complex Teuthonista symbols.
Also, this must be done in a way that these building blocks and rules are compatible to the mechanisms defined in the Unicode standards, as well as to the unwritten practices of encoding.

Some special Teuthonista symbols are presented with analyses how such symbols can be decomposed into building blocks which can be encoded.

A preliminary character set based on the "Einführung in den Sprachatlas der Deutschen Schweiz" (R. Hotzenköcherle, Bern 1962) is presented informally, which can be used as base for further work on the Unicode encoding of Teuthonista.

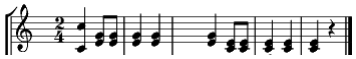
Unicode

- encodes characters

शुरुआत की थी जबकि हिन्दी विकिपीडिया की शुरुआत

A text, consisting of characters

- does not encode notational systems



Use of a notational system

(However, Unicode encodes building blocks for notational systems.)

Is Teuthonista a character set?

(rather than being a notational system)

(Of course, it is. But the Unicode Committee has to be convinced.)

Unicode

- encodes characters, not glyphs.
- Accented forms can be dynamically composed:

A + ◌̈ → Ä
0041 0308

Here two characters are used: A and ¨, to generate the “composed character” Ä.

- Characters have well-defined semantics:

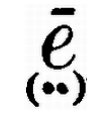
In the previous example, ¨ is a single character, with the semantic “diaeresis”, “trema”, or “umlaut”.
There is no recurrence to the „optical building elements“ (the single dots in this case).

The latter also implies:

- Teuthonista itself must be defined completely and exhaustively, in terms of characters, before getting encoded.

However, the phonetics which Teuthonista characters denote are not part of their definition when getting encoded in Unicode.

Thus, a typical composed Teuthonista character is ...



= LATIN SMALL LETTER E (the base letter)
+ COMBINING MACRON (the dash above of the e)
+ either (Solution A):
  - *COMBINING LEFT PARENTHESIS BELOW
  + COMBINING DIAERESIS BELOW
  + *COMBINING RIGHT PARENTHESIS BELOW
+ or (Solution B):
  - COMBINING DIAERESIS BELOW
  + *COMBINING PAIR OF PARENTHESES BELOW
+ or (Solution C):
  - *COMBINING PARENTHESIZED DIAERESIS BELOW

= LATIN SMALL LETTER E (the base letter)
+ COMBINING MACRON (the dash above of the e)
+ either (Solution A):
  - *COMBINING LEFT PARENTHESIS BELOW
  + COMBINING DIAERESIS BELOW
  + *COMBINING RIGHT PARENTHESIS BELOW
+ or (Solution B):
  - COMBINING DIAERESIS BELOW
  + *COMBINING PAIR OF PARENTHESES BELOW
+ or (Solution C):
  - *COMBINING PARENTHESIZED DIAERESIS BELOW

Drawbacks of Solution A:
- Parentheses occur in pairs only. Such pairs are a semantic unit in Teuthonista.

Drawbacks of Solution B:
- Font related: Fonts containing such characters must employ mechanisms for relative positioning of the parentheses and the enclosed base diacritics.
- Unicode related: There is no “combining class” to accommodate a diacritic like “pair of parentheses”.

Advantages of Solution C:
- Parenthesized diacritics are, as single entities, easily to handle on font design
- The members of the Teuthonista fourfold diacritic sequences (single in parens, single, double in parens, double) are treated equally as semantic units

Thus, Solution C is preferred when encoding Teuthonista.

Unicode

- Unicode characters represent plain text.

  Plain text e.g. is suited for storing into text databases, without the need to include any style, layout, or positioning information.

Not in the scope of Unicode are:

- any presentation and formatting issues
- any font and character style issues
- any layout issues.

This implies:

- All issues of vertical positioning must be handled by the definition of the characters themselves.

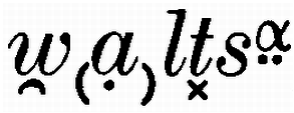
Unicode contains

- modifier letters (raised or lowered letters), like $^{a\ e\ x}\ _{a\ e\ x}$
- combining letters which act like accents placed above of a letter, like ẚ ḛ i̇ x̌

But what about these:

Unicode does not provide a mechanism to apply its combining characters (like the dieaeresis here) to a superscript letter.
Thus, the "superscript ü" must be encoded as a separate entity.

What is this? – It denotes a vowel pronounced between e and i, marked as a long vowel (by the macron) as a whole.
Thus, it is an "e" + "superscript i" + "macron". The macron is "stacked" upon the combination of "e" + "superscript i", which ***is*** the mechanism provided by Unicode. The macron retains its full length.
However, to get a "superscript i" without a dot, either the font must provide a special display for this combination, or a special "superscript dotless i" must be encoded.



– from [A] (see Bibliography), p.61

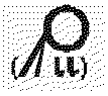
What about the diaeresis below the raised alpha?

- The raised alpha is a modifier letter
  (as any raised or lowered letter in Unicode is called).
- The character part where the diacritic applies is unambiguous.
  This differs from combining superscript letters, as for a dieaeresis on o + superscript u + diaeresis above *(where Unicode rules apply the diaeresis to the o, not to the superscript u)*.
- Therefore, the standard Unicode combining marks can be applied in such cases.
  It is the responsibility of the font or the rendering system to select the correct place and size for the diacritic.
- Thus, the last letter in the example above is:
  MODIFIER LETTER SMALL ALPHA + COMBINING DIAERESIS BELOW.
  In this case, nothing new is to be encoded.

These characters are taken from [C] (see bibliography).

They show a special diacritical mark, the “Teuthonista ray”, in the typical Teuthonista row of four: single in parens, single, double in parens, double.

Although this diacritical mark is left-attaching, it has to be entered *after* the base letter, as it is common for all diacritical marks. *(Such “logical order” being different from the “visual order” is common for South Asian scripts.)*



This special beast (also taken from [C]) shows dynamic rendering of character sequences, a feature also common for South Asian scripts.
It is: LATIN SMALL LETTER O
+ *DOUBLE PARENTHESIZED TEUTHONISTA RAY
+ *DOUBLE PARENTHESIZED TEUTHONISTA HOOK.
It requires to be done by the font or rendering system:
When a *DOUBLE PARENTHETIZED TEUTHONISTA RAY is followed by another double parenthesized diacritical mark to be placed below the base letter, they are displayed parenthesized as a whole.
(As this does not affect the semantics, this special rendering is optional.)

Bibliography

[A] Hotzenköcherle, Rudolf:
Einführung in den Sprachatlas der Deutschen Schweiz. Bern 1962
Band A: Zur Methodologie der Kleinraumatlanten

[B] Hotzenköcherle, Rudolf:
Einführung in den Sprachatlas der Deutschen Schweiz. Bern 1962
Band B: Fragebuch - Transkriptionsschlüssel - Aufnahmeprotokolle

[C] Reichel, Sibylle:
Handbuch zum Zeichensatz SMFTeuthonista. Erlangen 2003
http://www.sprachatlas.phil.uni-erlangen.de/materialien/Teuthonista_Handbuch.pdf

[M] Wiesinger, Peter:
Das phonetische Transkriptionssystem der Zeitschrift „Teuthonista“.
Zeitschrift für Mundartforschung, Wiesbaden 1964, pp.1-40

[S] Steger, Hugo; Schupp, Volker:
Einleitung zum Südwestdeutschen Sprachatlas. Marburg 1993

Appendix

Preliminary list of characters suitable for encoding found in the works listed in the Bibliography
For each character, an example is referenced in parentheses: letter referring to the Biography + page number

Diacritical Marks

PARENTHESIZED COMBINING DOT BELOW (A52)
PARENTHESIZED COMBINING DIAERESIS BELOW (A53)

PARENTHESIZED TEUTHONISTA HOOK (A63)
TEUTHONISTA HOOK (A52)
PARENTHESIZED DOUBLE TEUTHONISTA HOOK (A53)
DOUBLE TEUTHONISTA HOOK (A53)

PARENTHESIZED TEUTHONISTA RAY (C)
TEUTHONISTA RAY (C)
PARENTHESIZED DOUBLE TEUTHONISTA RAY (C)
DOUBLE TEUTHONISTA RAY (C)

PARENTHESIZED COMBINING DIAERESIS (B80)
PARENTHESIZED COMBINING TILDE (S70)
PARENTHESIZED COMBINING VERTICAL LINE BELOW (A68)

COMBINING TWOFOLD INVERTED BREVE ABOVE (B81)
COMBINING TWOFOLD INVERTED BREVE BELOW (A72)
COMBINING TWOFOLD BREVE BELOW (B92)
COMBINING DOUBLE INVERTED BREVE BELOW (B85)
COMBINING CIRCUMFLEX ACCENT PAIR ABOVE (B86)
COMBINING LEMNISCATE ABOVE (B85)
COMBINING GREATER THAN ABOVE (B88)
COMBINING DOWNWARDS ARROW ABOVE (M37)
COMBINING MULTIPLICATION SIGN BELOW (A72)
COMBINING WAVY LINE BELOW (A/XV) connects to left and right

Combining Letters above

COMBINING LATIN SMALL LETTER O UMLAUT (A66)
COMBINING LATIN SMALL LETTER U UMLAUT (B80)
COMBINING LATIN SMALL LETTER SCHWA
COMBINING LATIN SMALL LETTER B (A72)
COMBINING LATIN SMALL LETTER F (B89)
COMBINING LATIN SMALL LETTER SHARP S (B89)
COMBINING LATIN SMALL LETTER L WITH DOUBLE MIDDLE TILDE (B93)
COMBINING LATIN SMALL LETTER ESH (S73)
COMBINING LATIN SMALL LETTER O WITH TEUTHONISTA RAY (S68)
COMBINING LATIN SMALL LETTER U WITH TEUTHONISTA RAY (S68)

Latin letters

LATIN SMALL LETTER U WITH SHORT RIGHT STEM (A69)
LATIN SMALL LETTER U WITH LEFT CURL (B81)
LATIN SMALL LETTER E WITH FLOURISH (B83)
LATIN SMALL LETTER SHORT ESH (B89) (reduced to caps height)
LATIN CAPITAL LETTER R WITH RIGHT HOOK (B92)
LATIN SMALL LETTER R WITH RIGHT HOOK (due to stability policies)
LATIN SMALL LETTER L WITH MIDDLE LEMNISCATE (B93)
LATIN SMALL LETTER L WITH MIDDLE DOUBLE TILDE (B93)
LATIN SMALL LETTER X WITH TWO LONG LEGS (S73) (no chi! - straight when not italics)
LATIN SMALL LETTER X WITH TWO LONG LEGS AND RIGHT CURL (S73)
MODIFIER LETTER SMALL X WITH TWO LONG LEGS AND RIGHT CURL (S74)
LATIN SMALL LETTER X WITH LEFT LONG LEG (S73)
LATIN SMALL LETTER X WITH LEFT LONG LEG AND RIGHT CURL (S73)
LATIN SMALL LETTER X WITH AND RIGHT CURL (S73)
LATIN LETTER SMALL CAPITAL SCRIPT R (S76)

Punctuation marks

LEFT SMALL DOUBLE PARENTHESIS (A72, A/XV) marking standard language in dialectology
RIGHT SMALL DOUBLE PARENTHESIS (A72, A/XV)

Miscellaneous Technical

LARGE PLUS SIGN (A/XV) marking missing inquiry about a term in dialectology

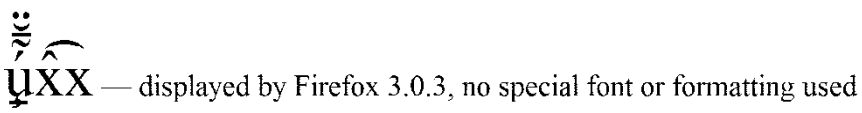
Appendix 2 *(added for the printed version)*

Some notes from the discussion held after the presentation at the Vienna workshop

1. As Unicode encodes characters, not glyphs, there cannot be assumed that the characters are displayed using a specific font style, as italics.
   As it is a common practice to typeset transcripted phrases in italics, there are character designs which base on their appearance in an italic font. An example was the "Bavarian f" presented at the workshop, which in italics look as an ordinary "f" having its descender cut off.
   However, it was not explained how this character has to look if a normal (i.e. not slanted or italicized) sans serif font is used:
   f – *f* = f – ?
   To get such a character to be encoded, such peculiarities must be solved.

2. It is not necessary to assign code points to combinations of a base letter with one or several diacritical marks.
   As it was shown, Unicode is capable to stack an unlimited number of diacritical marks above or below a base letter. The "rendering system" (i.e. the combination of the operating system, the text processing system, and the selected font) is responsible to display the result correctly, including issues like adjusting the line spacing correctly.
   However, it can be practical to design such combinations in a font, to ensure that a specific glyph design will be used when a special combination of base letter + diacritics is encountered, which is aesthetically superior to the display the rendering system would produce.
   Such fonts do not need to assign a code point to such a combined glyph. They can assign an internal handle like a "PostScript glyph name" to it which in turn is declared in the font to be used when a specific sequence of code points is encountered.
   Thus, such fonts do not even need to be "complete". If a diacritic combination is encountered for which there is no glyph in the font, the automatic rendering system will be active, producing a display which may aesthetically less pleasant, while staying semantically correct.

   It has to be admitted, however, that rendering systems which do diacritic stacking for Latin are hard to find today. Thus, having fonts containing glyphs for all diacritic combinations used within a project is an interim solution until the common rendering systems (text processing software as well as Internet access software) do as they are expected to do. There is at least progress in this direction, which can be seen on comparing the most recent versions of Internet Explorer or Firefox with older ones.


— displayed by Firefox 3.0.3, no special font or formatting used