

# Tengwar *tehtar* and UCS encoding (Michael Everson 1998-03-03)

In ISO/IEC JTC1/SC2/WG2 N1641, the following text (which I have edited somewhat for clarity) is given regarding the proposed encoding for the Tengwar *tehtar*:

Non-spacing marks, generically called *tehtar* ‘signs’, indicate vowels or other modifications of consonantal letters. *Tehtar* are placed above or below consonants, or atop “carriers” when no consonant is present in the required position. The occurrence of a character in the *tehtar* range, depicted in the code table with relation to a dashed circle, constitutes an assertion that this character is intended to be applied via some process to the consonantal character that *precedes* it in the text stream. General rules for applying non-spacing marks are given in Section 2.5 of the Unicode Standard. In ISO 10646, Level 2 encoding is intended. See the remarks on Modes below.

The SHORT CARRIER  $\dot{\ }_1$  simply bears the vowel *tehta*; the LONG CARRIER  $\dot{\ }_j$  indicates that the vowel was long; this can also be done by doubling the vowel sign (so  $\acute{\ }_i = e$ ,  $\acute{\ }_j = \hat{e}$ ,  $\acute{\ }_i = \hat{e}$ ).

## Modes

The morphological structure of a language determines the “mode” in which the Tengwar script is used for it. For instance, the *tehtar* are placed above or below the *preceding* consonant in languages in which words tend to end in a vowel (*i.e.* with a CV structure); but they are placed above or below the *following* consonant in languages in which words tend to end in a consonant (*i.e.* with a CVC structure) – compare Quenya  $\acute{\ }_3$  *nelde* ‘three’,  $\acute{\ }_3$  *neltildi* ‘triangle’ with Sindarin  $\acute{\ }_3$  *neled* and  $\acute{\ }_3$  *nelthil* (the different use of  $\acute{\ }$  for *n* in Quenya and  $\acute{\ }$  for *n* in Sindarin (since  $\acute{\ } = nn$  in Sindarin) is irrelevant here). In accordance with UCS specifications, however, the *tehtar* are proposed to be encoded as non-spacing characters, and so must *follow* the consonant *over* which they appear. For Sindarin, this requires that the logical order of backing store does not reflect its true syllabic structure. For instance, the Quenya examples here are encoded  $\acute{\ }_3$  (NUUMEN-ACUTE-ALDA-ACUTE, **n-e-ld-e**), and  $\acute{\ }_3$  (NUUMEN-ACUTE-LAMBE-TINCO-AMATICSE-ALDA-AMATICSE, **n-e-l-t-i-ld-i**); the Sindarin examples are encoded  $\acute{\ }_3$  (NUUMEN-LAMBE-ACUTE-ANDO-ACUTE, **n-l-e-d-e**), and  $\acute{\ }_3$  (NUUMEN-LAMBE-ACUTE-THUULE-LAMBE-AMATICSE, **n-l-e-th-l-i**). English is generally written according to a Sindarin-type mode; Italian would be written according to a Quenya-type mode. This inconsistency of phonetic representation and encoding in the backing store is a function of the script's unique representation of modalities which must be reckoned with apart from the character set itself. Smart inputting methods, such as are used for some Brahmic scripts, could solve the problem for Sindarin-type mode inputting. In the mode of Beleriand, where the *tehtar* are not used, but full vowels, the Sindarin examples are written thus:  $\acute{\ }_3$  (OORE-YANTA-LAMBE-YANTA-ANDO, **n-e-l-e-d**) and  $\acute{\ }_3$  (OORE-YANTA-LAMBE-THUULE-SHORT CARRIER-LAMBE, **n-e-l-th-i-l**). Mapping software for conversion between *tehtar*-mode and Beleriand-mode Sindarin will be requisite.

Gary Roberts has criticized this proposal:

The issue concerns writing vowels before consonants when the vowels are pronounced after the consonants. The word ‘animal’ would be the code sequence (in presentation, vowels or *tehtar*, occur above the preceding consonant):  $\acute{\ }_3$  **n-a-m-i-l-a** given the ‘standard style’ to write English. Another style (commonly used for Quenya, but which can be used for English) would appear as  $\acute{\ }_3$  **^a-n-i-m-a-l** where the ^ indicates a carrier. Perhaps *tehtar* are not non-spacing marks at all, but are often written in the form of a ligature. So far as I know, English, or any other language *can* be written either with *tehtar* ligatured with the preceding

or following consonant. It would be unfortunate if the underlying form was different. Perhaps the mode of Belerian is an indication that the shape and positioning of *tehtar* is a rendering issue, and should not be codified.

It would, as Gary suggests, be possible to achieve this by a reanalysis of the *tehtar* and extensive use of ZWJ to achieve the ligatures. The examples given here would then be written thus:

<i>nelde</i>	նճ	ն + ZWJ + օ + Տ + ZWJ + օ	[n-e]-[ld-e]
<i>neltildi</i>	նճթճ	ն + ZWJ + օ + Շ + ք + ZWJ + օ + Տ + ZWJ + օ	[n-e]-l-[t-i]-[ld-i]
<i>neled</i>	նճը	ն + օ + ZWJ + Շ + օ + ZWJ + թ	n-[e-l]-[e-d]
<i>nelthil</i>	նճեի	ն + օ + ZWJ + Շ + ե + օ + ZWJ + Շ	n-[e-l]-[th-i]-l
<i>neled</i>	նաչաթ	ն + ա + Շ + ա + թ	n-e-l-e-d
<i>nelthil</i>	նաչեի	ն + ա + Շ + ե + ի + Շ	n-e-l-th-i-l
<i>animal</i>	անմա	յ + ZWJ + օ + ն + ZWJ + օ + մ + ZWJ + օ + Շ	[^a]-[n-i]-[m-a]-l
<i>animal</i>	անմա	օ + ZWJ + ն + օ + ZWJ + մ + օ + ZWJ + Շ	[a-n]-[i-m]-[a-l]
<i>animal</i>	անմա	ա + ն + ի + մ + ա + լ	a-n-i-m-a-l

An immediate advantage to be seen here is that the underlying syllabic structure of any language is preserved for most texts. What this would do to sorting operations should be looked at. Two disadvantages can be observed:

1. In the mode of Belerian, “a dot is sometimes placed over շ and յ to mark these as separate letters, and not a curl or a stem belonging to an adjacent letter.” (Lawrence J. Krieg, “The Tengwar of Fëanor”, in Jim Allan, ed. *An introduction to Elvish*. Frome: Bran’s Head, 1978, p. 236.) Such a practice would spoil the syllabic success attained by using the ZWJ (that is, the underlying encoding of շնմա would be [a-i]-n-[^i]-m-[a-i]-l).
2. Inputting is, in principle, made more complex for the user – or at least for the Quenya-writing user, since it has already been observed that, under the current proposal, Sindarin- or English-writing users will require special inputting software. Roberts’ “ligature” proposal could put all Tengwar users at the same “disadvantage”. It is rather an overuse of ZWJ, but such use would be unavoidable since mode is not algorithmically predictable.

It occurs to me that the principles discussed here may have some relevance to the planned encoding of the Pahawh Hmong script, which has some “nonlinear” features. However, Tengwar is the only script I know of to date in which the relationship of backing store and logical (phonetic) order is not clear for all usages of “nonspacing” marks.

Comparison of encodings:

	<b>Standard coding of vowels</b>	<b>ZWJ</b>
թոկիէն	t-o-l-k-i-^e-n (8 chars.)	[t-o]-l-[k-i]-[^e]-n (11 chars.)
թոկինէ	t-l-o-k-^i-n-e (8 chars.)	t-[o-l]-k-[^i]-[e-n] (11 chars.)
թոկաէն	t-l-o-k-^i-n-e (8 chars.)	t-[o-l]-k-[^i]-[e-n] (11 chars.)
թոկյէն	t-o-l-k-y-e-n (7 chars.)	t-o-l-[k-y]-e-n (8 chars.)
թոկյեն	t-o-l-k-y-e-n (7 chars.)	t-o-l-[k-y]-e-n (8 chars.)
թոկիէն	t-o-l-k-i-e-n (7 chars.)	t-o-l-k-i-e-n (7 chars.)
թոկիեն	t-o-l-k-i-e-n (7 chars.)	t-o-l-k-i-e-n (7 chars.)
Տիմարիլիօն	s-i-l-m-a-r-i-l-i-^o-n (13 chars.)	[s-i]-l-[m-a]-[r-i]-l-[l-i]-[^o]-n (18 chars.)
Տիմրալիօն	s-l-i-m-r-a-l-i-^i-n-o (12 chars.)	s-[i-l]-m-[a-r]-[i-l]-l-[i-^]-[o-n] (17 chars.)