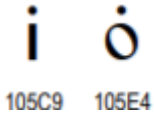
To: Script Ad Hoc/ UTC
From: Roozbeh Pournader (via Debbie Anderson)
Subject: Todhri encoding options
Date: 12 Feb. 2022

The following is an analysis of the options/models for Todhri, which was proposed in L2/20-188R2. The SAH had recommended approval in L2/20-250, but there remained one issue to be resolved, how to handle EI (which had been proposed at U+105C9) and U (proposed at U+105F4), with glyphs below.



(This document satisfies Action Item 165-A24.)

There are three options/models:

1. Encode the letters EI and U as characters, with no canonical decomposition.
2. Encode the letters EI and U as characters, with canonical decompositions to <I, COMBINING DOT ABOVE> and <O, COMBINING DOT ABOVE>.
3. Do not encode EI and U as characters, and represent them as the sequences <I, COMBINING DOT ABOVE> and <O, COMBINING DOT ABOVE>.

Pros and cons of each option are as follows:

Option	Pros	Cons
1	• No character in the script would have canonical decompositions, making some processes slightly simpler. • Collation of *clean* text would work as expected.	• *Do Not Use* tables would be needed, which are practically the same as equivalence, except that their data is typically unavailable in i18n libraries. This would cause “invisible” equivalence. • Some content creators will use U+0307 in the script anyway, causing multiple representation issues, including searching, matching, and collation issues.
2	• Alternative representations of the same text would be canonically equivalent. • Collation of *any* text would work as expected in UCA, since DUCET would automatically add contractions for the dotted letters. • *Do Not Use* tables would be avoided.	• Some characters would have canonical decompositions, adding slight complexity to the script. • U+0307 COMBINING DOT ABOVE would be used in a decomposition in a non-Latin script for the first time. (Note that it’s also used in the compatibility decomposition of U+02D9 DOT ABOVE.)

3	• Todhri text would have only one representation, with no need for canonical equivalence. • *Do Not Use* tables would be avoided.	• The two dotted letters would always be represented using two characters instead of just one. • Collation would need either tailoring or addition of contractions in DUCET.

In the authors' opinion, either of options 2 and 3 work for encoding the script, while option 1 is problematic for data processing. Considering that EI and U are thought as letters in the alphabet, option 2 appears to be more acceptable for the user community, with the additional benefit of collation working out of the box.

Note that Todhri already uses a series a series of diacritical marks from the U+0300..U+036F block, namely U+0301, U+0304, U+0311, and U+035E (see L2/20-188R2, p. 3). There are many scripts that use characters from the Combining Diacritical Marks block (U+0300..U+036F). A short overview gives us Latin, Greek, Cyrillic, Coptic, Syriac, Tifinagh, Tai Le, Old Permic, etc.

Of these, Latin, Greek, and Cyrillic are scripts that both use such characters, and include precomposed characters that include these combining marks in their decompositions. For example:

U+017C LATIN SMALL LETTER Z WITH DOT ABOVE ż ≡ <U+007A z, U+0307 ◌̇>
U+03AC GREEK SMALL LETTER ALPHA WITH TONOS ά ≡ <U+03B1 α, U+0301 ◌́>
U+04C2 CYRILLIC SMALL LETTER ZHE WITH BREVE ӂ ≡ <U+0436 ж, U+0306 ◌̆>