

Comparison of Kinzi Proposals

Two solutions have been proposed for distinguishing between Kinzi and YRWH medials attached to Myanmar letter Nga. This document compares the advantages and disadvantages of the two proposals. It is important that a firm decision is made as to which proposal to go with before Myanmar Unicode software is released to the wider community.

Proposal A was made by Martin Hoskens and Maung Tun Tun Lwin in Unicode Technical Note 11.

Proposal B was made by the Myanmar Language Commission and NLP Research Lab.

These cases are summarized in the table below for the case of U+101D, but also apply if U+101A, U+101B or U+101F is substituted for U+101D.

<i>Glyphs</i>	<i>Proposal A: UTN 11</i>	<i>Proposal B: NLP</i>
ꯀꯂ	U+1004 U+1039 U+101D U+1031	U+1004 U+1039 <i>U+200D</i> U+101D U+1031
ꯀꯃ	U+1004 <i>U+200D</i> U+1039 U+101D U+1031	U+1004 U+1039 U+101D U+1031

A variety of points have been made about the relative advantages and disadvantages. These are presented in the table below. Note the proposals suggest different renderings for the sequence U+1004 U+1039 U+101D U+1031 and its equivalents for the other medials, so it is essential to resolve this.

	<i>Issue</i>	<i>Proposal A: UTN 11</i>	<i>Proposal B: NLP</i>
1	File size – kinzi is rare	This may take more space, given kinzi’s frequency compared to the medials used with nga, however, this factor is negligible on modern computers.	Perhaps slightly reduced file sized, but negligible, especially if compression is used. This argument has already been ignored for the case U+1004 U+1039 U+200C
2	Collation rules	In glibc locale 10 extra rules required – but this is negligible compared to the total number of rules required for Myanmar. ¹	No change to normal rules.
3	Unicode 4 compliance Note: UTN11 is only a technical note, it is not part of the standard.	Kinzi – always compliant. YRWH medials – deviate in special cases as specified in UTN 11.	Kinzi – deviates in special cases. YRWH medials – always compliant. UTN 11 – non-compliant.
4	Searching for Kinzi	Search for U+1004 U+1039 ![U+200C] where ! indicates a logical NOT This is a simpler than case B because U+200D will always be inserted before a medial if it follows U+1004 when used as a base. In practice you would normally search for a complete word, so this search will be rare.	Search for U+1004 U+1039 ![U+200C U+101A U+101B U+101C U+101D] where ! indicates a logical NOT and [...] matches any one of the characters inside the bracket. This is more complicated than A because of the medial forms. The multiple NOT is difficult for a user to specify.
5	Searching for Medial	e.g. medial U+101F ![U+1004] U+1039 U+101F This correctly finds the medial, but is slightly more complicated. Since Kinzi on YRWH is very rare the number of false positives without specifying ![U+1004] is very small (c.f. issue 1).	e.g. medial U+101F U+1039 U+101F This is slightly easier.

¹ Arguably these rules should be present even for proposal B so that standalone medials could be sorted.

	<i>Issue</i>	<i>Proposal A: UTN 11</i>	<i>Proposal B: NLP</i>
6	<p>Old Myanmar – with the example given: င</p> <p>There are 2 possibilities:</p> <p>a) င is a medial with င as the base consonant</p> <p>b) င is attached to another base consonant (equivalent to -ငင) with င as the main consonant for a second syllable.</p> <p>This distinction is crucial for collation. It would be helpful to see some examples in real words.</p>	<p>Medial case: dealt with using U+1004 U+200D U+1039 U+101C</p> <p>Stacked case: Unspecified. Since this would normally give Kinzi when stacked, the NLP suggestion seems a sensible extension in this case.</p>	<p>Medial case: Unspecified, but since it does occur in medial form in င (drop) this must be representable somehow.² It must be distinguished from the normal stacked case for collation.</p> <p>Stacked case: U+1004 U+1039 U+101C</p>
7	<p>Rendering complexity – crude estimate of number of context specific rules required.</p> <p>This comparison is too simplistic: it is implementation dependent.³</p>	<p>Kinzi – one rule</p> <p>YRWH – two rules each</p> <p>Total: 9 rules</p> <p>Note: if character classes are used then this is only 3 rules.</p>	<p>Kinzi – two rules</p> <p>YRWH – one rule</p> <p>Total: 3 rules</p> <p>Fewer rules if character classes not supported.</p>
8	<p>Input Method complexity – assuming Kinzi is typed after the consonant it sits on top of.</p> <p>This relative complexity is implementation dependent.⁴</p>	<p>Context length 2 required for determining YRWH medial. i.e. after the sequence</p> <p>U+1004 [U+1031]?</p> <p>insert U+200D U+1039 [U+101A U+101B U+101D U+101F] before U+1031</p> <p>where ? means the character occurs 0 or 1 times.</p> <p>(Note: the medials already need a much longer context than to support reordering of multiple medials)</p>	<p>Context length 2 required for determining which Kinzi to use.</p> <p>After [U+101A U+101B U+101D U+101F] [U+1031]? substitute U+1004 U+1039 U+200D before the sequence.</p> <p>(Note the substitution inserts 2 characters back, not 1, but this complexity is already necessary for Myanmar Input Methods).</p>
9	<p>Implied syllable breaking – this is the argument used in UTN11.</p>	<p>ZWJ signifies that the medial is linked to the base consonant within the syllable.</p>	<p>ZWJ appears (incorrectly) to join Kinzi with the consonant of the next syllable.</p>

² See example in UTN 11.

³ Implementation example using Graphite for rendering Padauk:

Proposal A: 200D just prevents the Kinzi rule from matching, no extra rule is required with 200D. The “Take Kinzi” character class has 4 extra characters compared to proposal B.

Proposal B: One extra rule is required, so it could be argued that A is marginally easier to implement.

⁴ For the Keyman Input Method (which uses character classes) two rules are needed to implement both proposals so there is no difference in complexity. You can switch between the two proposals by commenting out 2 lines and uncommenting the other 2.

Summary

1,2,7,8 are all implementation issues. Both Kinzi proposals are implementable using the current Myanmar Unicode technology and neither method significantly adds to the level of complexity already required. Proposal B has a marginal advantage in 1 and 2. The relative complexity for 7 and 8 is implementation dependent – see footnotes for example.

9 is valid for proposal A, but technical issues rather than linguistic arguments should be the deciding factor.

6 is an issue for old Myanmar. Perhaps neither proposal addresses all possible combinations. It is important that old Myanmar is unambiguously representable in Unicode, so it needs to be clarified. However, it does not affect current use so it should not be the deciding factor.

4 and 5 are usability issues that affect whether a normal user gets what he “expects”. Anything that is done in the underlying Input Method or Renderer is hidden from the user. However, searching is one area where the user is in control and software can’t do much to help and so should have a higher priority than implementation complexity. In these areas proposal A has an advantage in 4 and proposal B has the advantage in 5. However, if you take the easier search string (i.e. A’s for 4 and B’s for 5) you get fewer false positives with A (since Kinzi is rare - see issue 1). Neither search is likely to be very common, but A has an advantage here. Most users don’t know about logical ORs and NOTs.

3 is by far the most important issue. At the moment A has the clear advantage, since it is already a Unicode technical note. If proposal B is to be adopted, then a new technical note or other official Unicode documentation is necessary to clarify the situation. For this to be accepted proposal B must have persuasive arguments in its favour. If the comparison presented here is correct, then it seems unlikely that the arguments meet that criteria.

6th December 2004

Keith Stribley

References

Unicode Standard, Version 4.0, Unicode Consortium, 2003, Addison-Wesley.

Representing Myanmar in Unicode: Details and Examples, 2003, Martin Hosken and Maung Tun Tun Lwin.

Kinzi Issue: Technical Adjustment, 2004, Myanmar Language Commission and Myanmar NLP Research Team.