L2/18-107

MWG/2-N13

Title: Summary of suggestions and proposals for improvements to the Mongolian phonetic model
Author: Roozbeh Pournader (WhatsApp/Facebook)
Date: 2018-04-05

This document briefly lists all the proposals related to the phonetic model raised at the Mongolian Working Group 2 meeting in April 2018 in San José, California. Higher priorities for the participating experts are specified and highlighted.

Proposals and suggestions from Badral Sanlig (main document MWG/2-N3R)

- The feminine forms of MONGOLIAN LETTER QA and MONGOLIAN LETTER GA either should be separately encoded as separate characters, or as fixed variants with FVS1. **(Priority 1.)**
- We should fix the NNBSP character as soon as possible or encode a new suffix connector for Mongolian. **(Priority 2.)**
- Stylistic and historical characters and variation sequences in the Mongolian block need to be cleaned up and potentially reorganized into the Ali Gali section of the Mongolian block. **(Priority 3.)**
- Font rules for Mongolian fonts (OTF) need to become simplified and standardized and made public.
- Different implementations of Mongolian script need to converge and get unified.
- Rendering rules for Mongolian in text rendering engines should be standardized.
- Positional mismatches, between the name of Mongolian variation sequences and default forms in Unicode and their actual cursive joining positional variants, need to be fixed.
- The number of Free Variation Selectors should be reduced to just one, at least in input.
- A font mechanism may need to be added to make the users aware of incorrect use of Free Variations Selectors, like typing more than one of them consecutively.
- New characters may need to be encoded for quotation marks.
  Note: Other participants at the meeting suggested using U+00AB LEFT-POINTING DOUBLE ANGLE QUOTATION MARK and U+00BB RIGHT-POINTING DOUBLE ANGLE QUOTATION MARK.
- A single-dot abbreviation sign may need to be added for Mongolian to help with name and surname abbreviations.
  Note: other participants at the meeting suggested that instead of encoding a new character, we may be able to solve this in fonts, by adjusting and reducing the left-side and right-side bearings of U+1802 MONGOLIAN COMMA (top-side and bottom-side in vertical writing) or other punctuation such as U+00B7 MIDDLE DOT, and using an additional space character before or after the normal full stop, as needed.
- A compound word joiner character may need to be added for writing compound words. This character would have a visual effect on the previous and following syllables by modifying the shaping and vowel harmony context.

- There are missing forms in Hudum, Ali Gali, and Todo that may need to be encoded as new characters or variations sequences.
- The Mongolian script, as presently encoded, may be unsuitable for internationalized domain names due to security issues.

Proposals and suggestions from Jirimutu, Siqin, Bao Haishan, Burigudu, and Menghejiya (main document MWG/2-N1)

- Because NNBSP, as a crucial shaping control character for Mongolian, is fragile in text rendering, especially when there is font fallback and there's ambiguity in how text should be segmented into script runs and font runs:
    - Change behavior of NNBSP to not have any contextual effect on succeeding letters. Instead, use Free Variation Selectors to control the special shaping of suffixes.
    - Allow the use of normal space as a suffix joiner (when NNBSP is used today) for cases where it's allowable to break the line at the position.

Proposals and suggestions from Menghejiya (main document MWG/2-N7)

- Style variants should be implemented using fonts, not variations sequences. Texts that want to display such style variants would have the same encoding, but would use a modern or historic font depending on the desired display.
- Unify all variation sequences across different models and implementations:
    - A few positional variants (which are different than style variants and can happen in the same style) may need to be added.
    - Various differences in submodels and implementations should be fixed and the implementations should be unified and use the same submodel.
      Note: submodel here means different models used for implementing the phonetic model by different implementations.
    - All positional variants (especially isolated variants) need to be defined, specified, and shown in the standard.
    - When Free Variation Selectors are used, the effect of the Free Variation Selectors on a base letter should not depend on the context. The shape of a Mongolian variation sequence may only depend on its base letter, its variation selector, and if the preceding and succeeding characters are joining or non-joining.
    - All positional variants of a letter, including its default form, should have a corresponding variations sequence. (Presently, the default form of a letter in a position does not have a variations sequence specified for it.) **(Priority 1.)**
- There are various problems with the NNBSP. We should either **(Priority 2)**:
  A) Add a new Mongolian Suffix Connector character in the Mongolian block; or
  B) Specify that NNBSP does not affect Mongolian shaping, and its only role is to make sure there is no word break at that position.

- A fourth Free Variation Sequence needs to be added for a dotless medial form of GA and potentially other positional variants for other letters, including encoding a new Free Variation Selector 4. **(Priority 1.1, a corollary to Priority 1.)**
- We need a control character for **(Priority 3)**:
  1. For limiting the effect of feminine or masculine vowels in a word when the word has two or more roots of different genders.
  2. For breaking automatic ligatures.
- The behavior and properties of U+180E MONGOLIAN VOWEL SEPARATOR need to be clarified and potentially modified to match its usage and effects in the Mongolian script.

Suggestions and proposals by Enkhdalai Baatar (main document MWG/2-N9)

- A new character should be added for the feminine form of QA and GA, disunified from the currently encoded QA and GA. This would help reduce the number of rules needed for shaping Mongolian. **(Priority 1.)**
- Stylistic and historical variants should be implemented using fonts, not variations sequences. Unicode encoding should not include style variants, they should be removed so the rules in Unicode would be simplified and shortened. **(Priority 2.)**
- MVS and the three Free Variation Selectors should be replaced by just one Free Variation Selector (with potentially two or more of it used in a row). **(Priority 3.)**
- Using the Mongolian script block in Unicode should become easier for end users and supporting it should become easier for font and application developers.
- For proofreading fonts, diacritic changes could be made to the o/u/oe/ue vowels to differentiate their pronunciation, for example, a horizontal or vertical stroke, or a dot in the middle of their loop.
- Page scrolling and <textarea> and <textview> scrolling should be handled separately and automatically, so they could be different for horizontal and vertical text.
- When right-to-left text is inserted in Mongolian vertical text, the right-to-left text should go top-to-bottom, and not bottom-to-top.

Other new information based on discussions by participating experts

- U+1806 MONGOLIAN TODO SOFT HYPHEN:
  - The usage of this character is not limited to the Todo script. It is also used for Hudum in Mongolia to separate compound words.
- U+202F NARROW NO-BREAK SPACE (also known as suffix connector):
  - The width of the character is not always less than a normal space. The width is a typographical choice depending on the style.
  - In some use cases in Mongolian, including in book titles, store signs, etc, it is valid to break the line at this character. When and if the break happens, the suffix connector disappears at the line break the same way that in Western typography, a space disappears if a line is broken at it.
  - Because of justification, NNBSP may also need to get expanded if the spaces in the line are stretched (currently its width is fixed in most implementations).

    - Some implementations continue to either replace NNBSP with a normal space, or handle it incorrectly. Sometimes the mishandling happens when NNBSP is both preceded and followed by Mongolian text, which is easier to fix. But sometimes on one side of NNBSP there are digits or non-Mongolian text, which is harder to fix.
- Parallel models:
    - Based on presentations and comments by Liang Jinbao, Jirimutu, Liang Hai, and others, it is clear that the phonetic model is better than the graphetic model for some use cases, while the graphetic model is better than the phonetic model for some other use cases. We should consider pursuing two parallel models for Mongolian script in Unicode, similar to how Korean in Hangul script is encoded, so that different applications could choose the best model for their use case or their users. Similar to Hangul, algorithms for converting between the two models could be specified to reduce ambiguity and disparity between implementations. The graphetic model could be encoded in a separate block, after considering other languages written in the Mongolian script. Fixing the phonetic model in the Unicode Standard should remain a high priority.