L2/18-314

TO: **UTC**
FROM: **Debbie Anderson, SEI, UC Berkeley**
SUBJECT: Mongolian Ad Hoc meeting summary
DATE: **22 September 2018 (revised 30 October 2018)**

Participants: Debbie Anderson, Peter Edberg, Liang Hai, Lisa Moore, Roozbeh Pournader, Michel Suignard, and Tex Texin

The Mongolian Ad Hoc met on 22 September 2018 to review Mongolian script topics, including review of two recently received documents from Badral Sanlig et al.: L2/18-293 "Solution for NNBSP issues" and L2/18-294 "Proposal to encode two Mongolian letters." The group met again on 12 October 2018, at which time Liang Hai provided further feedback on L2/18-293 and L2/18-294, which is also summarized here.

The two documents will require further detailed study by the Script Ad Hoc group and thorough discussion. Liang Hai will provide a further report on L2/18-293 and L2/18-294 to the Script Ad Hoc meeting on October 12, 2018. The recommendations from the Script Ad Hoc will be submitted for the Unicode Technical Committee meeting #158 in January 14-17, 2019. UTC #158 will take place before the next Mongolian Working Group meeting in Ulaanbaatar from April 3-5, 2019.

The following summarizes the main points raised during discussion and notes any follow-up actions.

1. Issues concerning NNBSP (U+202F NARROW NO BREAK SPACE)

- The NNBSP for Mongolian is not properly documented for implementers to use. The situation can be improved with better documentation.
- Features of NNBSP include:
  - it can "stretch" during justification
  - users don't have a consistent expectation for how NNBSP should behave (i.e., they don't select NNBSP primarily for its line breaking and word breaking behavior; instead, users are forced to use NNBSP to form separated suffixes because of its special shaping function)
  - it has special contextual shaping properties (i.e., it triggers the required special shapes of the Mongolian characters that follow it), which differs from the use of NNBSP in European typography
  - it is used primarily for Hudum and is required by only one separated suffix in Todo, Manchu, and Sibe, respectively.
  - when it appears at a line break, it needs to "disappear" into the margin, while the characters appearing after the line break still need to be triggered for the special shaping.
  - when it appears between a non-Mongolian character and a sequence of Mongolian characters, the characters after NNBSP also need to be triggered for the special shaping (which often fails in current shaping engines).
- While users don't have a consistent expectation for NNBSP about its word and line breaking behavior, retaining its current behavior is still deemed the most appropriate approach, as it reflects the most acceptable behavior for the majority of users, although this behavior may not be ideal. In short, NNBSP's properties Line_Break = GL and Word_Break = ExtendNumLet do not need to change.

    - Keeping the property values intact will give some users, the prescriptivists, ideal behavior (given proper implementation), but others, who don't need NNBSP to prevent line or word breaking, will get over-restrictive, but acceptable behavior. If, on the other hand, the values are changed to be breakable, the prescriptivists will find NNBSP's behavior unacceptable, while other users will get their preferred breaking behavior, though such behavior is likely not critical for them. Note that the default algorithms from UAX #14 "Unicode Line Breaking Algorithm" and UAX #29 "Unicode Text Segmentation" are not meant to fulfill all use cases.
- To reflect the usage of NNBSP in both Latin and Mongolian text, NNBSP should have Script_Extensions = Latn Mong. (Note that the Script_Extensions value is an unordered list; the order of the scripts is not significant.) Clear documentation of criteria for curating the Script_Extensions property and the intended usage of the property should be provided in the future.
- NNBSP's Script = Common should be retained.

Follow-up actions:

- Roozbeh Pournader and Liang Hai will provide wording for UAX #14 Unicode Line-Breaking Algorithm.
- Roozbeh Pournader will add Latin and Mongolian in ScriptExtensions.txt for U+202F.

Further Recommendations from Liang Hai:

- Contact should be made with Microsoft on two issues: (1) the implementation of NNBSP in MS Word for Office, so it is conformant with UAX #29's latest Word_Break = ExtendNumLet and (2) improving Mongolian spell checking in Word.
- In order to help achieve an agreement on the special shaping effect of NNBSP, the authors of L2/18-293 should compare lists of structures requiring NNBSP amongst various standards and specifications, including the *Users' Convention* (TR #170 and MNS 4932: 2000), China's unilaterally modified *Users' Convention* (Prof. Quejingzhabu's personal publication 蒙古文编码 and GB/T 26226-2010), and other specs, such as Prof. Quejingzhabu's *Specification 9* and the EAC Project Standard.

2. Proposal to encode two new characters (KE and GE)

It was noted that the inconsistent shaping between implementations is more the result of the lack of a proper shaping specification, rather than encoding level issue (that is, encoding characters that are not ideal candidates for unification will complicate the required shaping logic).

Encoding new characters typically takes several years before they are supported in software and appear in fonts. In addition, use of the legacy encoded characters typically does not really stop, but more often, old encodings continue to be the primary representations. As a result, implementations will still need to support the old encodings with the aid of a proper shaping specification (which itself can already resolve the problem).

Attempting to alter the encoding-level aspect will not solve the problem, unless the urgently needed shaping specification becomes available. Eventually it would be the shaping specification that would actually solve the problem, while the change in the encoded characters could actually be harmful, because:

- encoding new characters still doesn't solve the problem of a lack of a shaping specification
- the new characters complicate the situation for implementers, since they have to wait for the new characters and then will have to support both the new characters and the old encoding.

To improve the current situation, the following minimal activities should be done:

- clarify how the existing characters should be rendered
- include a warning to implementers about use of U+182C MONGOLIAN QA and U+U+182D MONGOLIAN GA.

Further Recommendations from Liang Hai:

- The feminine forms of U+182C MONGOLIAN LETTER QA and U+182D MONGOLIAN LETTER GA should not be disunified (either to new characters or reused characters) from the existing two characters.
- A complete shaping specification (which will affect the recommended text representation logic) should be drafted by the Unicode experts and receive the community's feedback.
- An example of an OpenType implementation should be made available for implementers to refer to, since certain complicated contextual mechanisms are hard to implement with only basic OpenType knowledge.
- Once the Mongolian encoding is stabilized in the future, the authors should help LibreOffice and other software packages correct their strings for previewing Mongolian fonts.