

Comments on L2/18-293 and L2/18-294

Liang Hai (梁海) lianghai@gmail.com

14 October 2018

Reviewed documents

- **L2/18-293 Solution for NNBSF Issues** (Badral Sanlig, Munkh-Uchral Enkhtur)
- **L2/18-294 Proposal to encode two Mongolian letters** (Badral Sanlig, Jamiyansuren Togoobat, Munkh-Uchral Enkhtur)

Comments on L2/18-293

1. Background: NNBSF is a character loaded with multiple purposes for encoding Mongolian enclitics:
 - It is a whitespace that separates a word/enclitic and an enclitic, and superficially appears to have complicated cursive-joining effects.
 - It prevents both line breaking and word breaking, because grammatically enclitics are considered part of the preceding word.
 - It triggers special shaping of its following Mongolian characters, for certain enclitics that are written in special ways.
2. The problem is, as introduced in the proposal, because such a functionally overloaded character lacks a complete specification in any of the various standards, the intended behavior is not supported well in platforms and applications.
 - Due to misunderstandings about how text shaping works (the authors thought NNBSF's shaping effect fails because it's replaced by an ordinary space in shaping engines), NNBSF's eternal problem for script run segmentation (as a character essential in Mongolian shaping, it's got `Script = Common`) is not discussed in the document.
 - NNBSF does get replaced by U+0020 SPACE though, but this is largely because unexpected normalization (NNBSF has a compatibility decomposition of U+0020 SPACE), not because it has `gc = Zs` (Space_Separator).

3. The authors put an emphasis particularly on the poorly supported behavior of preventing word breaking.
 - It's claimed that, in order to enable spell-checking for Mongolian, Office Word must be able to recognize a stem word and its following enclitics as a single word. It's not clear in the proposal whether it's Word that doesn't provide multi-word string as a spell checking context or it's the authors' spell-checking solution that doesn't deal with multi-word string.
 - Also, authors report that NNBSF still doesn't work well in Microsoft Word 2016, but as it's reported in [L2/17-036](#)'s Appendix III, "... there was a fix applied late summer 2016 which did fix the word-count feature under Microsoft Word 2016", Word 2016 should already be able to recognize NNBSF as a word-joining character. More background research is needed.
 - An interesting fact presented by the authors: Spell checking is crucial for the Mongolian encoding because the encoding was so inappropriately designed to allow serious visual ambiguity in text, while in order to implement spell checking it's probably important to improve the NNBSF situation.

4. In order to reliably prevent word breaking while mostly keeping NNBSF's current behavior, the authors propose to encode a new character Mongolian Suffix Connector (MSC) at 0x180F to replace NNBSF.
 - Due to some internal inconsistencies in the proposal though, it's not clear if the authors want MSC to allow or prevent line breaking.
 - The authors argued from a grammatical point of view that it's a strong requirement of the Mongolian script to have a character that reliably behaves like how NNBSF was intended to behave (especially, to prevent word breaking although appears to be a whitespace).
 - However grammatical considerations are largely beyond a text encoding's concerns.
 - Although NNBSF was originally proposed from a grammatical point of view (and the shaping behavior was considered as a natural result of such a structure), it should be understood that NNBSF's de facto major duty today is to trigger correct shaping for enclitics.

5. As the authors consider NNBSF's deficiency is an encoding level problem, It's not thoroughly discussed how NNBSF's special shaping effect and other abilities can be clarified and standardized with better documentation.
 - However, since UAX #29 already specifies `Word_Break = ExtendNumLet` for

NNBSP (and the authors focus mostly on the non-word-breaking requirement), it's not clear what encoding (or character properties and algorithms) level problems still present.

6. A very informative list of enclitics is provided in the appendix. But there isn't a comparison study of the similar lists from various standards.
7. Note that the current Project Standard promoted by the Ethnic Affairs Committee of Inner Mongolia specifies both NNBSP and FVSes can be used for representing enclitics.
 - Although a preference is not explicitly specified in the Project Standard, in all the digitalized text commissioned by the EAC, enclitics are always encoded with FVSes and are usually also accompanied by NNBSP for preventing line breaking.

Comments on L2/18-294

1. The status quo of the KE–GE shaping problem is introduced in the document:
 - Inconsistent shaping between fonts.
 - Some forms are not included in the Unicode Standard's variant set. (Because by the encoding model's original design those forms are handled directly by the bowed-consonant ligation mechanism.)
 - The shaping relies on complicated and underspecified contexts.
 - Implementers are not able to implement the desired behavior with a reasonable number of rules.
 - Implementers have been abusing contextual rules in order to cover more words without requiring FVSes.
2. The authors' analysis is inappropriate.
 - The encoding is correctly recognized by the authors as the fundamental cause of the complicated shaping logic.
 - But the authors didn't realize that the lack of a standardized shaping specification is actually the direct cause.
 - The authors tried to use the Unicode Standard as the standard reference for text representation, then found most fonts today don't conform with the Unicode Standard. However, since the Unicode Standard has never been an appropriate complete guidance for implementing the Mongolian script, it is

exaggerating the incompatibility issue to exhibit how fonts don't match the Unicode Standard's names list (which is not meant to be a complete reference for contextual rules and FVS usage).

- Fonts today actually loosely follow the Users' Convention and some de facto agreements, neither of which is properly represented in the Unicode Standard.
 - Also, the authors made contradictory claims that, A) fonts are forced to automatically handle all shaping cases, and B) text should be encoded statically with FVSes because of the Unicode Standard's Mongolian names list.
3. The authors' argument for disunifying KE–GE from QA–GA is weak.
- In order to argue why the two proposed characters should be disunified from the existing two and why they should not be encoded as a single character, controversial grammar theories and far-fetched grapheme analysis are presented.
 - Authors need a more significant attempt of analyzing the existing encoding model and should try to derive arguments from the model's internal logic.
 - There're major misunderstandings about reusing Ali Gali Ka (U+1889).
4. Apparently it would've been a sensible idea to disunify KE–GE from QA–GA if it were proposed when Mongolian was originally encoded.
- However, now after nearly two decades, it would be harmful to make major changes to text representation only to simplify the shaping logic.
 - As no architectural problems of the current KE–GE encoding have been revealed, the issues remain on the level of specification and implementation, which is the proper level to make changes.
 - Authors don't seem to have realized that the standardization of shaping specification is the required first step for resolving compatibility issues. Altering the encoding alone won't actually resolve the shaping incompatibility.
5. For considering newly proposed Mongolian characters in the future:
- Introducing additional ways of encoding is indeed not as harmful for Mongolian as for other scripts. Because the Mongolian encoding already heavily suffers from architectural confusable issues and incompatible implementations (which require different ways of encoding), and implementations already have to support multiple encodings forever for

backward compatibility.

- Note new ways of encoding (although theoretically superior) are effectively just additional ways of encoding, because the old ways continue to exist and often continue to be the major representation. We've had a lot of failures before, trying to improve an unideal (but actually workable) encoding but ended up with introducing duplications forever: Tamil śrī, Malayalam chillus, Devanagari eyelash Ra, Bangla khanda Ta...
 - It would take years for platforms to consistently support the new characters, while the situation of existing characters can be improved (and need to be improved) with the current versions of platforms. Shaping engines themselves today generally are okay for rendering Mongolian and the major issues lie in the lack of shaping specification and the bad quality of font implementation.
 - It's only worth considering to significantly alter encoding when the change can resolve some architectural issues that are not resolvable in implementation and thus the change can significantly improve the encoding. Eg, the graphetic model proposed earlier would significantly improve the encoding.
6. The authors have pointed out that word processors like LibreOffice might be using inappropriate strings for previewing Mongolian fonts.