L2/19-116

Proposal to clarify the purpose of U+202F NARROW NO-BREAK SPACE

For consideration by Mongolian Working Group
For consideration by Unicode Technical Committee
For consideration by CLDR Technical Committee

2019-04-03

Marcel Schneider (charupdate@orange.fr)

"We must always say what we see.
Above all we must always
— which is more difficult —
see what we see."

Alain Finkielkraut quoting Charles Péguy

Proposal History

This proposal is submitted following instructions that Asmus Freytag gave on the Unicode Public Mailing List on January 19, 2019.

In order to prevent this paper from becoming an omnibus proposal, and keep it focused on U+202F NARRROW NO-BREAK SPACE, two other proposals have been submitted beforehand so that they can be cited below for further reference, instead of discussing here two issues with loose relationship but decisive incidence on the topic discussed in the following sections:

1. *Proposal to focus break prevention design on end-user input* (about WORD JOINER and ZWNBSP);
2. *Proposal to ensure usability of fixed-width spaces* (about the range U+2000..U+200A).

That was done after catching up an old but still unresolved issue related to bidi-mirroring — *Proposal to spread a warning about legibility of six bidi-mirrored symbols* —, because it had precedence over those issues related to whitespace characters, that while affecting the digital representation of virtually all languages, do not make them downright illegible.

Due to collision of deadlines (deadline clash) and especially with respect to the Mongolian Working Group Meeting at Ulaanbaatar on April 3–5, 2019, this proposal's essentials are summarized below for timely submission, before some related but differently scoped proposal(s) shall add for completion. Please see the upcoming *Proposal to define a space character for group separator*.

Problem

On the background of defective space character encoding (please refer to the preceding *Proposal to ensure usability of fixed-width spaces*), the *MONGOLIAN SPACE converted to NARROW NO-BREAK SPACE (NNBSP) and moved from the upcoming Mongolian block (where it was proposed as *U+1800) to the General Punctuation block (where it is U+202F), is actually a savior for virtually all other scripts. Thanks to Unicode,

the original *MONGOLIAN SPACE became the dreamt-of catch-up filling in the hole left by then-missing *no-break thin space,* or the lacuna ripped in the Standard by not making U+2009 THIN SPACE non-breaking.

On the other hand, while NNBSP started being supported by fonts and implementations for the sake of other scripts, it fails to meet all requirements for rendering when used in Mongol script. Part of the problem comes from its status of a de-facto compound of a space and a format control when used to write Mongolian — it turned out being too inflexible when text columns grow narrower while Mongolian enclitics may be long and may also concatenate with each other — and another part is due to soft-hinted glyph shaping based on an open set of enclitics. If NNBSP were proper to Mongol script, its properties could be freely fine-tuned, but as its encoding appeared as the only way of granting all scripts the long-expected no-break thin space, the *MONGOLIAN SPACE designed as a Mongolian Suffix Connector (MSC) is actually hijacked by virtually all of the world's writing systems, and particularly by French.

The semantic overload of NNBSP is thus another consequence of wrongly encoding U+2009 THIN SPACE as a breaking space (line break class BA) instead of giving it the correct Line_Break property value GL. If the Unicode whitespace model were functional, moving *NONGOLIAN SPACE outside Mongolian block would have been pointless, and Mongol script would have a connecting character of its own, whose property values could be freely adapted for optimization of the Mongolian encoding model.

Background

The Public Review Issue #308 was about changing the General_Category value of NNBSP from whitespace (Zs, separator space) to that of a connecting punctuation mark (Pc). The attempt was finally dropped, but the fact that it was on the table for a while illustrates the trail of interference that the merger of the *Mongolian space* functionality with the *no-break thin space* functionality brought into existence. It was lenified on the Background page stating:

> The only other widely noted use for U+202F NNBSP is for representation of the thin non-breaking space (*espace fine insécable*) regularly seen next to certain punctuation marks in French style typography.

Actually, the no-break thin space is so widely noted that it has its own TeX short form command backslash comma, and it is required in so common use cases as German "z.<no-break thin space>B." (English "e.g."), and of course as a group separator in all locales using space in that context (please refer to the upcoming *Proposal to define a space character for group separator*). Hence, restricting the use of NNBSP to Mongolian (Mongol script) and French (Latin script) may be understood as a compromise between two competing encoding goals: Per the Unicode Standard it was encoded for Mongolian and is primarily used in Mongol script, but can also be used to represent the *no-break thin space* (French: *espace fine insécable*) occurring next to a number of punctuation marks in French.

When during the extensive data collection for CLDR version 34 (summer 2018), the French locale data was updated accordingly, it came up that this is also the only appropriate space character for grouping digits in large numbers for all locales using the SI-conformant *thin space* for that purpose. As a consequence, the group separator moved from U+00A0 to U+202F in French — because asking for that fix was the responsibility of French vetters — while simultaneously the CLDR Technical Committee was urged to get all eligible locales implement the update of the group separator from NBSP to NNBSP, thanks to growing availability in fonts of

that character awaited since a long time. The corresponding CLDR ticket #11423 from September 17, 2018 starts:

> To be cost-effective, locales using space as numbers group separator should migrate at once from the wrong U+00A0 to the correct U+202F. I didn't aim at making French stand out, but at correcting an error in CLDR. […]

And it goes on in comment #1 from the same day (boldened text is original: the grammar mistake is corrected in square brackets):

> After having painstakingly [caught] up support of some narrow fixed-width no-break space (U+202F). the industry is now ready to migrate from U+00A0 to U+202F. Doing it in a single rush is way more cost-effective than migrating one locale this time, another locale next time, a handful locales the time after, possibly splitting them up in sublocales with different migration schedules. I really believed that now Unicode proves ready to adopt the real group separator in French, all relevant locales would be consistently pushed for correcting that value in release 34. The v34 alpha overview makes clear they are not.
>
> http://cldr.unicode.org/index/downloads/cldr-34#TOC-Migration
>
> I aimed at correcting an error in CLDR, not at making French stand out. Having many locales and sublocales stick with the wrong value makes no sense any more.
>
> https://www.unicode.org/cldr/charts/34/by_type/numbers.symbols.html#a1ef41eaeb6982d
>
> The only effect is implementers skipping migration for fr-FR while waiting for the others to catch up, then doing it for all at once.
>
> There seems to be a misunderstanding: The **locale setting** is whether to use period, comma, space, apostrophe, U+066C ARABIC THOUSANDS SEPARATOR, or another graphic.
> Whether "space" is NO-BREAK SPACE or NARROW NO-BREAK SPACE is **not a locale setting,** but it's all about Unicode **design** and Unicode **implementation**.
> I really thought that that was clear and that there's no need to heavily insist on the ST "French" forum. When referring to the "French thousands separator" I only meant that unlike comma- or period-using locales, the French locale uses space and that the group separator space should be the correct one. That did **not** mean that French should use **another** space than the other locales using space.

However, the use of NNBSP in other scripts than Mongol script unveils as a mere workaround in an attempt to cope with the Unicode encoding error affecting U+2009 THIN SPACE, so far as this space is breaking instead of non-breaking. An error that it has in common with other space characters, and that should be fixed prior to being able to propose any further fixes for non-Mongolian scripts such as Latin.

Proposed actions

1. After correcting the Line_Break property value of whitespaces as proposed in the preceding *Proposal to ensure usability of fixed-width spaces*, state that THIN SPACE is the preferred representation of the non-breaking thin space used in virtually all scripts, and cease directing French users to represent the

French *espace fine insécable* by using NNBSP instead of THIN SPACE as soon as all implementations are upgraded within the time-lapse between two Unicode major versions.

2. Change the Script_Extensions property value of NNBSP from {Latn, Mong} to {Mong}, or its Script property value from "Common" to "Mong".
3. In the Code Charts, add an informative alias to U+202F NARROW NO-BREAK SPACE: = Mongolian space

References

Microsoft Typography: *Whitespace*
https://docs.microsoft.com/en-us/typography/develop/character-design-standards/whitespace

Kenneth Whistler, UTR 14 and U+202F NARROW NO-BREAK SPACE, 2007-08-08, L2/07-209R
https://www.unicode.org/L2/L2007/07209-whistler-uax14.txt

Liang Hai, *COMMENTS ON L2/17-036 (MONGOLIAN SUFFIX CONNECTOR),* 2017-01-25, L2/17-052
https://www.unicode.org/L2/L2017/17052-mongolian-cmt.pdf

Badral Sanlig, Munkh-Uchral Enkhtur, *Solution for NNBSP issues,* 2018-09-10, L2/18-293
https://www.unicode.org/L2/L2018/18293-nnbsp-solution.pdf

Liang Hai, *Comments on L2/18-293 and L2/18-294,* 2018-10-14, L2/18-316
https://www.unicode.org/L2/L2018/18316-mongolian-doc-cmt.pdf

Debbie Anderson, *Mongolian Ad Hoc meeting summary,* 2018-09-22 / 2018-10-30, L2/18-314
https://www.unicode.org/L2/L2018/18314-mongolian-ad-hoc.pdf

Deborah Anderson, *Mongolian Ad Hoc Report,* 2016-09-29, L2/16-297
https://www.unicode.org/L2/L2016/16297-n4769-mongolian-ad-hoc.pdf

Jirimutu, Siqin, Bao Haishan, Burigudu, Menghejiya, *The Proposal for deprecation of MSC/NNBSP Mongolian Suffix Form Controlling Behavior,* MWG/2-N1
https://www.unicode.org/~lisa/mongoliandocs/mwg2-1Mongolian-Jirimutu-Proposal.pdf

Mrunix.de, *Gesperrte Leerzeichen* [forum thread], 2007-05-03
https://www.mrunix.de/forums/showthread.php?50696-Gesperrte-Leerzeichen

Acknowledgments

Thanks to everyone who directly or indirectly helped put this paper together.

Thanks to Google for Google Search.

Thanks to Microsoft for Word Online and OneDrive.