L2/19-317

**Proposal to update some statements about space characters in *Unicode Standard Annex #14: Unicode Line Breaking Algorithm***

For consideration by Unicode Technical Committee

2019-09-27

Marcel Schneider (charupdate@orange.fr)

*We should always say what we see.*
*Above all we should always*
*—which is most difficult—*
*see what we see.*

Charles Péguy

# Introduction

The specification of the *Unicode Line Breaking Algorithm* (*Unicode Standard Annex #14*) is to reflect state-of-the-art knowledge in spacing numbers and other parts of speech, so as to become citable in the first place when documenting end-user interfaces, mainly with respect to NARROW NO-BREAK SPACE in Latin script. The rationale and background are provided at the end of this paper.

The version of UAX #14 cited for reference is current latest version Unicode 12.0.0 (2019-02-15), which is revision 43.

Deleted text is highlighted in ~~red and barred~~. Proposed new wording is highlighted in yellow. Where existing text is reordered while elements are kept unchanged, these are highlighted in lime green.

# Proposed Actions

## 1) GL: Non-breaking (“Glue”) (XB, XA)

**Change from:**

Non-breaking characters prohibit breaks on either side, but that prohibition can be overridden by SP or ZW. In particular, when NO-BREAK SPACE follows SPACE, there is a break opportunity after the SPACE and the NO-BREAK SPACE will go as visible space onto the next line. See also WJ. The following are examples of characters of line break class GL:

| | |
|---|---|
| 00A0 | NO-BREAK SPACE (NBSP) |
| 202F | NARROW NO-BREAK SPACE (NNBSP) |
| 180E | MONGOLIAN VOWEL SEPARATOR (MVS) |

NO-BREAK SPACE is the preferred character to use where two words are to be visually separated but kept on the same line, as in the case of a title and a name “Dr.<NBSP>Joseph Becker”. When SPACE follows NO-BREAK SPACE, there is no break, because there never is a break in front of SPACE.

NARROW NO-BREAK SPACE has exactly the same line breaking behavior as NO-BREAK SPACE, but with a narrow display width. The MONGOLIAN VOWEL SEPARATOR acts like a NARROW NO-BREAK SPACE in its line breaking behavior. Both of these characters are regularly used in Mongolian text, where they participate in special shaping behavior, as described in *Section 13.5, Mongolian* of [Unicode].

When NARROW NO-BREAK SPACE occurs in French text, it should be interpreted as an “espace fine insécable”.

| | |
|---|---|
| 034F | COMBINING GRAPHEME JOINER |

This character has no visible glyph and its presence indicates that adjoining characters are to be treated as a graphemic unit, therefore preventing line breaks between them. The use of *grapheme joiner* affects other processes, such as sorting, therefore, U+2060 WORD JOINER should be used if the intent is to merely prevent a line break.

| | |
|---|---|
| 2007 | FIGURE SPACE |

This is the preferred space to use in numbers. It has the same width as a digit and keeps the number together for the purpose of line breaking.

**Change to:**

Non-breaking characters prohibit breaks on either side, but that prohibition can be overridden by SP or ZW. In particular, when NO-BREAK SPACE follows SPACE, there is a break opportunity after the SPACE, and the NO-BREAK SPACE will go as visible space onto the next line. See also WJ. The following are examples of characters of line break class GL:

| | |
|---|---|
| 00A0 | NO-BREAK SPACE (NBSP) |
| 180E | MONGOLIAN VOWEL SEPARATOR (MVS) |
| 202F | NARROW NO-BREAK SPACE (NNBSP) |

NO-BREAK SPACE has exactly the same behavior as SPACE in horizontal justification, but without providing any line break opportunity. It is ~~the preferred character to~~ used where two words are to be visually separated but kept on the same line, as in the case of a title and a name: “Dr.<NBSP>Joseph Becker”. When NO-BREAK SPACE precedes SPACE ~~follows~~, there is no break, because there never is a break in front of SPACE.

NARROW NO-BREAK SPACE has ~~exactly~~ only the same line breaking behavior as NO-BREAK SPACE, while it does not change in width when horizontal justification is enabled. This is the preferred space to use where a non-breaking THIN SPACE is required. Examples include grouping digits, visually separating contiguous quotation marks as in English, and setting off tall punctuation marks in French text, where it is called “espace fine insécable”.

The MONGOLIAN VOWEL SEPARATOR acts like a NARROW NO-BREAK SPACE in its line breaking behavior. Historically, both of these characters are regularly used in ~~Mongolian text~~ Mongol script, where they participate in special shaping behavior, as described in *Section 13.5, Mongolian* of [Unicode].

| | |
|---|---|
| 034F | COMBINING GRAPHEME JOINER (CGJ) |

This character has no visible glyph, and its presence indicates that adjoining characters are to be treated as a graphemic unit, therefore preventing line breaks between them. The use of *grapheme joiner* affects other processes, such as sorting; therefore, if the intent is to merely prevent a line break, U+2060 WORD JOINER should be used instead.

| | |
|---|---|
| 2007 | FIGURE SPACE (FSP) |

This ~~is the preferred~~ space ~~to~~ is used ~~in~~ to indent numbers as a way of vertically aligning decimal separators. It has the same width as a digit and ~~keeps the number together for the purpose of line breaking~~ is thus too wide as a group separator. See NNBSP.

## 2) IS: Infix Numeric Separator (XB)

**Change from:**

*Note:* FIGURE SPACE, not being a punctuation mark, has been given the line break class GL.

**Change to:**

*Note:* FIGURE SPACE~~,~~ and PUNCTUATION SPACE are used in front of numbers as a way of vertically aligning decimal separators. Not being ~~a punctuation mark, has~~ intended for use as infix numeric separators, they have been given ~~the~~ other line break classes ~~GL~~.

## 3) BA: Break After (A)

**Change from:**

| | |
|---|---|
| 1680 | OGHAM SPACE MARK |
| 2000 | EN QUAD |
| 2001 | EM QUAD |
| 2002 | EN SPACE |
| 2003 | EM SPACE |
| 2004 | THREE-PER-EM SPACE |
| 2005 | FOUR-PER-EM SPACE |
| 2006 | SIX-PER-EM SPACE |
| 2008 | PUNCTUATION SPACE |
| 2009 | THIN SPACE |
| 200A | HAIR SPACE |
| 205F | MEDIUM MATHEMATICAL SPACE |
| 3000 | IDEOGRAPHIC SPACE |

All of these space characters have a specific width, but otherwise behave as breaking spaces. In setting a justified line, none of these spaces normally changes in width, except for THIN SPACE when used in mathematical notation. See also the **SP** property**.**

**Change to:**

| | |
|---|---|
| 1680 | OGHAM SPACE MARK |
| 2000 | EN QUAD (NQSP) |
| 2001 | EM QUAD (MQSP) |
| 2002 | EN SPACE (ENSP) |
| 2003 | EM SPACE (EMSP) |
| 2004 | THREE-PER-EM SPACE (3/MSP) |
| 2005 | FOUR-PER-EM SPACE (4/MSP) |
| 2006 | SIX-PER-EM SPACE (6/MSP) |
| 2008 | PUNCTUATION SPACE (PSP) |
| 2009 | THIN SPACE (THSP) |
| 200A | HAIR SPACE (HSP) |
| 205F | MEDIUM MATHEMATICAL SPACE (MMSP) |
| 3000 | IDEOGRAPHIC SPACE (IDSP) |

~~All of~~ These ~~space~~ characters ~~have a specific width, but otherwise~~ behave as breaking spaces, but some of them such as THIN SPACE, EN SPACE, EM SPACE are currently tailored to meet user expectations by non-breaking behavior in some environments. [PARAGRAPH BREAK]

~~All of these space characters have a specific width,~~ and ~~In setting a~~ when lines are justified ~~line~~, none of ~~these spaces~~ them normally changes in width, except for THIN SPACE when used in mathematical notation. See also the **SP** property.

## Rationale

In the wake of Mongolian Working Group Meeting #3 — see L2/19-134, *Summary of Mongolian property updates in Unicode since MWG2*, and L2/19-141, *Summary of proposals made during Mongolian Working Group Meeting 3 (MWG3)*, from Roozbeh Pournader — and further item #20 on page 14 in L2/19-286 *Recommendations to UTC #160 July 2019 on Script Proposals*, from Deborah Anderson, et al., Script Ad Hoc group, and the related Action item #103 of UTC meeting #160, updating the Unicode Standard and Standard Annexes is necessary in order to reflect the new status of NNBSP, given that end-user documentation — such as documentation of keyboard layouts for the French locale — is expected to point to these resources.

In particular, about ***not*** using FIGURE SPACE in numbers but a non-breaking THIN SPACE, thus NNBSP in practice, please refer to L2/19-112 *Proposal to define a space character as a group separator*. We are aware that locale-specific data is curated in CLDR. However, UAX #14 throughout all accessible versions (earliest is revision 4) is making a statement about which space character "is the preferred space to use in numbers", which means, as a group separator. Obviously, UAX #14 is the correct level for sorting out which space is preferred in any locale using a space rather than a punctuation mark for the purpose of grouping digits, since

there is actually as little choice as where the group separator is FULL STOP, especially because NBSP as a group separator is an inappropriate pre-Unicode fallback, that cannot be used for that purpose in standard plain text unless horizontal justification is turned off, which means, it is a non-standard representation of the group separator space; and FIGURE SPACE is even less usable.

More detailed rationales are provided on a per-section basis as follows:

## 1) GL: Non-breaking ("Glue") (XB, XA)

In the first quoted paragraph, a comma is missing.

In the first quoted table, lines are not sorted on code points in ascending order. Hence rows 2 and 3 should be permuted.

In the first paragraph after the first quoted table, key information about justifying behavior of NO-BREAK SPACE is missing. It has been added in a sentence at paragraph start on the pattern found at the beginning of the next paragraph.

In the conditional clause starting the last sentence of the same paragraph, the character names SPACE and NO-BREAK SPACE should be arranged in logical order, starting with NO-BREAK SPACE as the subject of the clause because it is the topic of the paragraph. The verb is then changed accordingly.

The second paragraph after that table starts talking about NARROW NO-BREAK SPACE before switching to MONGOLIAN VOWEL SEPARATOR, and then to both, while the next brief paragraph is again about NARROW NO-BREAK SPACE only. This is for historical reasons, since the third paragraph has been added later (in 2007). The time has come to rearrange these two paragraphs, grouping information about NNBSP in one paragraph, while giving an extra paragraph to characters used in Mongol script.

In the first sentence about NARROW NO-BREAK SPACE, the word "exactly" is unnecessary, and since the clause it is used in, is followed by the clause "but with a narrow display width", it is even misleading, as it suggests that the only difference between the two spaces is in display width. Hence it is replaced with "only", and missing information about the behavior of NNBSP in horizontal justification is added, as well as a small number of sample use cases ending with French in order to connect to the already existing locale information, now raised by one paragraph.

The role of NARROW NO-BREAK SPACE and MONGOLIAN VOWEL SEPARATOR in Mongolian text is about to change, probably (see L2/19-169, *Proposal to establish NARROW NO-BREAK SPACE as a definitely usable avatar of THIN SPACE*) in a way that MVS is freed from its legacy usage and takes over the format controlling part of NNBSP. So far it is already clear that Mongol script [1] is turned away from using NNBSP as a format control. To account for the shift, "Historically," is prepended to the sentence.

In the second quoted table, the acronym (for instance, CGJ) is missing. It is added according to the Code Chart of the block *Combining Diacritical Marks*, following the example of the first quoted table that yields the acronyms of all three characters.

The paragraph about the COMBINING GRAPHEME JOINER is lacking two punctuation marks, and the last two clauses should be swapped to stick with logical order; "instead" is appended in the process.

In the third quoted table, the acronym (for instance, FSP) is missing. It is added according to the Code Chart of the block *General Punctuation*, following the example of the first quoted table that yields the acronyms of all three characters.

Lastly, the paragraph about FIGURE SPACE needs to have an untrue statement corrected, given that FIGURE SPACE is undocumented in its alleged role as a group separator, while its true role in number indentation is halfway indirectly hinted in TUS. See L2/19-112 *Proposal to define a space character as a group separator*.

### 2) IS: Infix Numeric Separator (XB)

The Note is updated in accordance with the change exposed in the preceding paragraph.

### 3) BA: Break After (A)

In the last quoted table, the acronyms are missing, except for the OGHAM SPACE MARK, that does not have any acronym in the Code Chart of the block *Ogham*. They are added according to the Code Chart of the block *General Punctuation* and the Code Chart of the block *CJK Symbols and Punctuation*, following the example of the first quoted table that yields the acronyms of all three characters.

The logic of having the clause “but otherwise behave as breaking spaces” in its actual place is non-obvious, since before and after, this paragraph is about display width. Also, given the line break behavior of theses spaces is non-obvious either and raises much concern, as exposed in L2/19-115, *Proposal to ensure usability of fixed-width spaces*, some additional information about line break behavior is mandatory here. That leads to split this paragraph, dedicating the first of the resulting paragraphs to line break behavior. The second is then about advance width, with minimal changes to the existing text.

Although tailoring is dealt with in *Section 8: Customization*, it is key information here with respect to the misuse of THIN SPACE on the internet by users either fooled into thinking that user agent rendering engines would keep its line break behavior as tailored in DTP and LaTeX, or inappropriately exporting text from these environments, not noticing that the non-breaking behavior (of THIN SPACE) is merely due to tailoring and won’t subsist in a standard environment. When enumerating examples, EN SPACE and EM SPACE are added to THIN SPACE in an attempt to draw attention to the fact that tailoring is agnostic of canonical equivalence here, as EN QUAD and EM QUAD are kept breaking in LaTeX.

## Background

The use of NNBSP is not (and was never intended to be) limited to publishing. As NNBSP is a part of the interoperable representation of numerous languages including Armenian, English, French, Georgian, German and Tuareg using Tifinagh script, it is expected to be available out-of-the-box. One example is any French keyboard layout generating a text stream without or with manual or automated punctuation spacing as exposed in CLDR ticket #10904. In particular, the lack of *NO-BREAK THIN SPACE, or rather, the wrong line break class attributed to THIN SPACE in Unicode was a main disturbance impacting keyboard layout development.

Patrick Andries [2] reported the trouble with using non-breaking thin space, reflected also in the cited document L2/19-286.

# References

[1] Badral Sanlig, Munkh-Uchral Enkhtur, *Solution for NNBSP issues,* 2018-09-10, L2/18-293, page 1, note 1: "We use the term 'Mongol script' to indicate the traditional Mongolian script that is used in Mongolia and outside the country among other Mongolians. While the term 'Mongolian script' implies to scripts that is used only in the country Mongolia."

[2] ANDRIES Patrick, *Unicode 5.0 en pratique : codage des caractères et internationalisation des logiciels et des documents*, Dunod, Paris, 2008 [Read on Google Books]

# Acknowledgments

Special thanks to Deborah Anderson, Ken Whistler, Roozbeh Pournader, Lisa Moore, and Liang Hai from the Script Ad Hoc group, to the Mongolian Working Group, and to the Unicode Technical Committee, for helping fix the non-breaking thin space issue.

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

Thanks to Microsoft for Word Online and OneDrive.

Thanks to Google for Google Search, Google Books, Google Translate and Gmail.