

Proposal to define a space character as a group separator

For consideration by Unicode Technical Committee

For consideration by CLDR Technical Committee

For consideration by Mongolian Working Group

2019-04-18

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](#). It relies on the main issues regarding locale preferences and feasibility that Mark Davis listed in [CLDR ticket #11423, comment 13](#) from December 10, 2018.

In order to prevent this paper from becoming an omnibus proposal, three related proposals have been submitted beforehand:

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).
3. *Proposal to clarify the purpose of U+202F NARROW NO-BREAK SPACE*.

That was done after requesting that, if convenient for diplomatic reasons, they be kept in the queue until the series related to NNBSB completes.

This paper is less developed than it would have been without the impact of the news of the Notre-Dame disaster in Paris. However, as it is the last item missing from the NNBSB-related series and must meet the deadline of UTC meeting #159, it is a short form.

Problem

The issue of the *non-breaking thin space* in Unicode is composed of two questions:

1. Whether the Unicode Standard should support it, and if so, at which existing or new code point;
2. Whether the Unicode Common Locale Data Repository (CLDR) should feature it, notably in the role of a *group separator* for the 55 locales and 6 sub-locales using a space for that purpose.

Part of question (1) is easy to answer, and answering the rest is too early for now. One doesn't need to look farther than into the United States Government Publishing Office's Style Manual, that prescribes a thin space notably as a group separator and to separate double and single quotation marks. (For details please refer to the Background section.) The Unicode encoding principles and practice around accurate, interoperable and

streamlined representation necessitate that this be achieved by a single non-tailored space character rather than by formatting, tailoring, or composing a non-breaking space thanks to WORD JOINER, although that may be done on the web or in word processing or DTP applications. The space + WORD JOINER solution is too less straightforward, its usability is heavily challenged due to the late duplication of the break preventer, and consequently fonts supporting it are outnumbered by those already supporting NARROW NO-BREAK SPACE (NNBSP). Please refer to the preceding proposals for more information. It now depends on the decisions made for the *Mongol script* whether NNBSP will be fully available as a replacement of what THIN SPACE should be, as intended by Unicode [6] (this is subject to clarification in the Background section below). If it won't, the line break property value of THIN SPACE should be corrected as requested in the *Proposal to ensure usability of fixed-width spaces*, because encoding a new *NO-BREAK THIN SPACE is likely a non-starter.

Question (2) is definitely easy to answer with *yes*. Both NO-BREAK SPACE (NBSP; legacy, CLDR) and FIGURE SPACE (Unicode) potentially or permanently cause a whitespace to appear that is *optically* wider than a digit (see Background below). There is a strong assumption that a locale preferring such a representation of numbers has not been found yet. Standards bodies — namely the International System of Units (SI) and its American English implementations [1][3] — and sources reflecting best practice such as the popular style guide from the French State Printing Office [4] as well as German Wikipedia [7] unanimously advocate a *thin space* when recommending to use space to form triads of digits in numbers. An “acceptable fallback” does basically not exist, given NBSP is justifying by default, and any text is potentially subject to justified layout. Only a number of more or less suboptimal workarounds are customarily implemented, which are either application-specific, or limited to certain environments, or bound to peculiar layout preferences, and are thus not interoperable.

The residual problem is due to the excessive delay in implementing the obvious solution. Part of this delay is a consequence of encoding mistakes, a related lack of recommendations in the *Unicode Standard*, and misleading or downright wrong statements in *Unicode Standard Annex #14* documenting the *Unicode Line Break Algorithm* (see below and the preceding *Proposal to ensure usability of fixed-width spaces*). Another part is the responsibility of top level management and system administrators failing to upgrade critical infrastructure for better cybersecurity. [10] Sluggish demand for up-to-date fonts may be understood as a side-effect of that neglect. The Unicode Community should in no way caution such strategic misconduct.

Background

CLDR uniformly represented the group separator space as NBSP, until for version 34 the French main locale made the move to NNBSP, that is already in general use with big punctuation in that locale (in all cases according to the new school). Since it started in the ordinary vetting process, it was too late for cross-locale synchronization, despite this was suggested as soon as CLDR TC had accepted the principle and v34 alpha was out. [9] After the release of version 34, issues presumably related to font support and backwards compatibility were raised on the cited CLDR ticket #11423 and fueled part of a discussion on the Unicode Public Mailing List in January 2019. [11]

The baseline of the **opponents' position** is that both for the purpose of displaying user interfaces — even in a context of internationalization- and localization-aware products — and in expected end-user input, NBSP should be used as a group separator preferredly to any other non-breaking space (notably NNBSP), because the advantage of being compatible with the bulk of current fonts, i.e. also with fonts never used in, nor intended for, UI display, and of being compatible with pre-Unicode dependencies still in use in spite of the

threat to cyber-security posed by outdated systems and hardware, [10] far outweighs all the advantages that are normally expected from accurate software localization and from enabling all end-users to write their language in its correct digital representation.

That position is based on a distinction made between so-called “good” or “fine” typography — better called *correct* typography — on one hand, and how the common of mortals is supposed to get their writing represented when typing on a computer on the other hand. One reason why that distinction is an abuse is that there is a continuum between correct typography and draft style. Another reason is that a community’s, staff’s, team’s or individual’s positioning in that continuum usually depends on extraneous constraints or personal commitment, so that making general assumptions about the matter is problematic, especially when made in the sense of the non-Mongolian opponents to the general use of NNBS.

Moreover, **Unicode’s intent when encoding NNBS** makes those assumptions pointless. When in 1998 the Unicode Technical Committee accepted to encode a *new* whitespace character while encoding the *Mongol script*, instead of the proposed *MONGOLIAN SPACE in the new *Mongolian* block, they decided to move that space into the *General Punctuation* block, and to change its name to NARROW NO-BREAK SPACE. The rationale as stated in [L2/98-268R](#) [6] was as follows:

1. Mongolian Space

The UTC accepts the Chinese and Mongolian requirements for encoding a separate Mongolian space. There is a reasonable case for the common usage of a non-breaking space of Mongolian-specific layout width that can be used for Mongolian-specific (common) purposes and which could **meaningfully contrast with a regular NBSP** used in the same text.

The UTC suggests that this character be named NARROW NO BREAK SPACE and that it be encoded in the General Punctuation block at U+202F. The concept of the Mongolian space (**a non-breaking space, narrower than a normal non-breaking space, and contrasting with it in usage**) could be of use in other scripts as well; therefore it is better to make this a general use **punctuation character**, rather than limiting it to the Mongolian script.

The language is clear and unambiguously points the need of what may be commonly called a *no-break thin space* for general use, i.e. in virtually any script. As already explained in the *Proposal to ensure usability of fixed-width spaces* and in the *Proposal to clarify the purpose of U+202F NARROW NO-BREAK SPACE*, Unicode used this occasion to correct the biased assignment of line break property values to typographic spaces. Originally — and straightforwardly — U+2009 THIN SPACE would be non-breaking (GL). That was presumably well understood in 1998, so that the cautious wording is somewhat surprising.

The rest of this Background section is a collection of **evidence for thin space as a group separator**. Let’s note already that according to NIST *Guide for the Use of the International System of Units (SI)*, “The practice of using a space to group digits is not usually followed in certain specialized applications, such as engineering drawings and financial statements.”

First, we’ll demonstrate why NBSP is inappropriate as a group separator except as a fallback in some well-delimited cases, namely when justification is turned off and font size is small. Although those circumstances are forcibly common, they basically represent an edge case that is intrinsically unfit to serve as a basis for character encoding design.

The following screenshots show a number whose digits are separated into groups by U+00A0 NO-BREAK SPACE (NBSP), or by U+202F NARROW NO-BREAK SPACE (NNBSP), between a preceding word and part of the following word.

1.1. With **NBSP**, without line justification:



1.2. With **NNBSP**, without line justification:



2.1. With **NBSP**, justified in different contexts with increasing interword spacing:



2.2. With **NNBSP**, justified in the same contexts as above:



For completeness, here is the already shown (in *Proposal to ensure usability of fixed-width spaces*, p. 3) variant with U+2007 **FIGURE SPACE**, recommended in UAX #14 stating about that space in all available versions: “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.”



Obviously, neither NO-BREAK SPACE nor FIGURE SPACE do actually “keep together” the numbers that we “use [them] in,” except to keep the groups on the same line. To devise FIGURE SPACE as the only non-breaking space in that range — as opposed to picking THIN SPACE if really it should be only a single one — was likely to make a bad joke. However, given to set up the Unicode Standard was not about joking, but about getting character encoding right, we have to seriously investigate, because providing our end-users with a meaningful explanation that respects logical reasoning is our duty.

Hence these are the **requirements for the group separator space**, as it has to be:

- non-breaking, i.e. of line break class GL (glue), not BA (break after);
- fixed-width (also called “fixed”), not justifying;
- thin (or “narrow”), not too wide.

That is what standards bodies such as BIPM (SI), ISO, NIST, style manuals such as US-GPO, Chicago, Imprimerie nationale (France), as well as power users on Wikipedia [7][9] are recommending. Let’s have a brief review including a small number of examples. This is enough for the Unicode Standard to include an appropriate character, whereas for CLDR all involved locales should be prompted for feedback, based on the comprehensive *list of locales using space as a group separator*, as it stands below.

The **United States Government Publishing Office *Style Manual*** [1] does not specify the width of the space to be used as a group separator, as rule 12.9 (p. 276) simply prompts:

Use spaces to separate groups of three digits in a decimal fraction.

The width of those spaces may be extrapolated from other instances specifying “thin space” to set off section and paragraph signs (rule 10.6, p. 263) and “to separate double and single quotation marks” (rule 8.137, p. 218).

The **National Institute of Standards and Technology (U.S. Department of Commerce) *Guide for the Use of the International System of Units (SI)*** [3] dedicates section 10.5.3 (p. 37) to *Grouping digits* (see also the note about specialized applications quoted above, that comes after the text quoted here):

Because the comma is widely used as the decimal marker outside the United States, it should not be used to separate digits into groups of three. Instead, digits should be separated into groups of three, counting from the decimal marker towards the left and right, by the use of a thin, fixed space. However, this practice is not usually followed for numbers having only four digits on either side of the decimal marker except when uniformity in a table is desired.

Examples: 76 483 522 but not: 76,483,522

43 279.168 29 but not: 43,279.168 29

8012 or 8 012 but not: 8,012

0.491 722 3 is highly preferred to: 0.4917223

0.5947 or 0.594 7 but not: 0.59 47

8012.5947 or 8 012.594 7 but not: 8 012.5947 or 8012.594 7

Consistently, the leading *Check List for Reviewing Manuscripts* in the same publication reminds as item 11 on page vi:

The digits of numerical values having more than four digits on either side of the decimal marker are separated into groups of three using a thin, fixed space counting from both the left and right of the decimal marker. For example, 15 739.012 53 is highly preferred to 15739.01253. Commas are not used to separate digits into groups of three. (See Sec. 10.5.3.)

Let’s look at the source. The bilingual brochure about ***The International System of Units (SI)*** from the **International Bureau of Weights and Measures (BIPM)** has section 5.3.4 (p. 133) about *Formatting numbers, and the decimal separator*. There we read:

Following the 9th CGPM (1948, Resolution 7) and the 22nd CGPM (2003, Resolution 10), for numbers with many digits the digits may be divided into groups of three by a thin space, in order to facilitate reading. Neither dots nor commas are inserted in the spaces between groups of three. However, when there are only four digits before or after the decimal marker, it is customary not to use a space to isolate a single digit. The practice of grouping digits in this way is a matter of choice; it is not always followed in certain specialized applications such as engineering drawings, financial statements, and scripts to be read by a computer.

Resolution 7 of the 9th CGPM (1948), quoted in Appendix 1 (p. 146) does not specify the width of the group separator space:

In numbers, the comma (French practice) or the dot (British practice) is used only to separate the integral part of numbers from the decimal part. Numbers may be divided in groups of three in order to facilitate reading; neither dots nor commas are ever inserted in **the spaces** between groups.

Resolution 10 of the 22nd CGPM (2003, p. 169 sq) only considers the confusion between dot and comma, declares that the decimal separator may be either, and reaffirms the principle about dividing numbers in groups of three, without elaborating upon the properties of the space character.

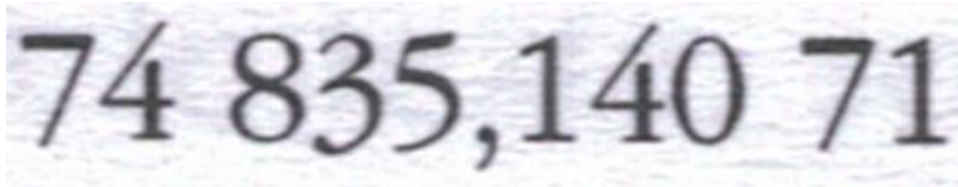
The editor country's State Printing Office **Imprimerie nationale** is clear about that, however. In the entry about *Numbers in Arabic digits* of its popular and widely followed style guide [4], a highlighted note on page 124 recommends [translation followed by French text]:

The numerals made up of digits expressing a quantity are written in groups of three digits (thousands) separated by **a non-breaking and non-justifying space**, both for the integer part and for the decimal part. These groups are formed starting from the comma towards the left for the integer part, and towards the right for the decimal part:

74 835,140 71

“Les nombres en chiffres exprimant une quantité s'écrivent par tranches de trois chiffres (tranches de mille) séparées par **une espace insécable et non dilatable**, tant pour la partie entière que pour la partie décimale. Ces groupes sont constitués en allant vers la gauche pour la partie entière, vers la droite pour la partie décimale, à partir de la virgule :”

An enlarged scan of the example makes clear that the space used is **not** FIGURE SPACE. It isn't NO-BREAK SPACE neither, since that is justifying. So the only remaining option as per latest Unicode version 12.0.0 is U+202F NARROW NO-BREAK SPACE:



Here is a list of all **locales sorted by scripts actually in CLDR preferring space** as group separator:

Language	Script	CLang	CScript
Armenian	Armenian	hy	Armn
Belarusian	Cyrillic	be	Cyrl
Bulgarian	Cyrillic	bg	Cyrl
Church Slavic	Cyrillic	cu	Cyrl
Kyrgyz	Cyrillic	ky	Cyrl
Ossetic	Cyrillic	os	Cyrl
Russian	Cyrillic	ru	Cyrl
Sakha	Cyrillic	sah	Cyrl
Tajik	Cyrillic	tg	Cyrl
Tatar	Cyrillic	tt	Cyrl
Ukrainian	Cyrillic	uk	Cyrl
Georgian	Georgian	ka	Geor
Afrikaans	Latin	af	Latn

Aghem	Latin	agq	Latn
Albanian	Latin	sq	Latn
Bafia	Latin	ksf	Latn
Basa (Cameroon)	Latin	bas	Latn
Breton	Latin	br	Latn
Central Atlas Tamazight	Latin	tzm	Latn
Colognian	Latin	ksh	Latn
Czech	Latin	cs	Latn
Duala	Latin	dua	Latn
English (Finland)	Latin	en_FI	Latn
English (South Africa)	Latin	en_ZA	Latn
English (Sweden)	Latin	en_SE	Latn
Esperanto	Latin	eo	Latn
Estonian	Latin	et	Latn
Ewondo	Latin	ewo	Latn
Finnish	Latin	fi	Latn
French	Latin	fr	Latn
Fulah	Latin	ff	Latn
German (Austria)	Latin	de_AT	Latn
Hungarian	Latin	hu	Latn
Inari Sami	Latin	smn	Latn
Jola-Fonyi	Latin	dyo	Latn
Kabuverdianu	Latin	kea	Latn
Kabyle	Latin	kab	Latn
Kako	Latin	kk	Latn
Koyra Chiini	Latin	khq	Latn
Koyraboro Senni	Latin	ses	Latn
Kwasio	Latin	nmg	Latn
Latvian	Latin	lv	Latn
Lithuanian	Latin	lt	Latn
Morisyen	Latin	mfe	Latn
Northern Sami	Latin	se	Latn
Norwegian Bokmål	Latin	nb	Latn
Norwegian Nynorsk	Latin	nn	Latn
Polish	Latin	pl	Latn
Portuguese (Portugal), European Portuguese	Latin	pt_PT	Latn
Prussian	Latin	prg	Latn
Slovak	Latin	sk	Latn
Spanish (Costa Rica)	Latin	es_CR	Latn
Swedish	Latin	sv	Latn
Tachelhit (Latin)	Latin	shi_Latn	Latn
Tasawaq	Latin	twq	Latn
Turkmen	Latin	tk	Latn
Uzbek	Latin	uz	Latn
Uzbek (Cyrillic)	Latin	uz_Cyrl	Latn
Xhosa	Latin	xh	Latn
Yangben	Latin	yav	Latn

Zarma	Latin	dje	Latn
Standard Moroccan Tamazight	Tifinagh	zgh	Tfng
Tachelhit	Tifinagh	shi	Tfng

Not only Latin script, but four other scripts are involved: Armenian, Cyrillic, Georgian and Tifinagh. As a consequence of the migration of the group separator space from the wrong U+00A0 to the best-fit U+202F, the `Script_Extensions` property value of U+202F should be set to {Armn, Cyrl, Geor, Latn, Mong, Tfng} instead of {Latn, Mong}.

Proposed actions

1. In **UAX #14**, correct the text about FIGURE SPACE to make clear that it is not used as a group separator, but to indent numbers in columns and tables for horizontal alignment (typically on decimal separator). Synch UAX #14 with the **Core Specification, § 6.2, Space Characters**, that already states: “U+2007 FIGURE SPACE has a fixed width, known as *tabular width*, which is the same width as digits used in tables.” Mention in this context in both instances that PUNCTUATION SPACE is in synergy with FIGURE SPACE (as it indents figures by the width of a group separator when this is COMMA, or FULL STOP).
2. Wait until a decision is made about the fate of NARROW NO-BREAK SPACE in Mongol script. As that decision is expected this year (2019), the issues around using NNBS in other scripts, and the related issue about the line break property value of THIN SPACE, should be settled soon.

References

- [1] United States Government Publishing Office, *Style Manual: An Official Guide to the Form and Style of Federal Government Publishing*. Washington, DC: U.S. Government Publishing Office, 2016. [[Read online](#)]
- [2] International Bureau of Weights and Measures (BIPM), *The International System of Units (SI)*, Paris: International Committee for Weights and Measures (CIPM; General Conference on Weights and Measures, CGPM), 8th edition, 2006. [[Read online](#)]
- [3] Ambler Thompson and Barry N. Taylor. *Guide for the Use of the International System of Units (SI)*. 2008 Edition, 2nd printing. NIST Special Publication 811. Gaithersburg, MD 20899: National Institute of Standards and Technology, U.S. Department of Commerce, 2008. [[Read online](#)]
- [4] Collectif Imprimerie nationale (France). *Lexique des règles typographiques en usage à l’Imprimerie nationale*. Printing March 2017. Paris: Imprimerie nationale, 2008.
- [5] Patrick Andries, *Unicode 5.0 en pratique : codage des caractères et internationalisation des logiciels et des documents*, Dunod, Paris, 2008. [[View on Google Books](#)]
- [6] Ken Whistler and the Unicode Technical Committee, *Analysis and UTC position regarding Mongolian Encoding issues* (Expert Contribution and UTC Position Paper), July 30, 1998, [L2/98-268R](#), section “UTC Positions on Mongolian Encoding”, 1. *Mongolian Space*, ¶ 2.

- [7] *Wikipedia* (Germanophone), “Schmales Leerzeichen” (Thin Space), first version from September 29, 2006 (see [article history](#)); latest version at the time of submission from [January 10, 2019](#). [Guarantee: This article reflects the unbiased Germanophone community’s views, so far as the submitter of this paper never contributed.]
- [8] *Wikipedia* (Anglophone), “Thin Space”, first version from August 4, 2010 (see [article history](#)); latest version at the time of submission from [March 26, 2019](#). [Guarantee: This article reflects the unbiased Anglophone community’s views, so far as the submitter of this paper never contributed.]
- [9] *CLDR ticket #11423 (accepted): Group separator migration from U+00A0 to U+202F* (bug report), September 17, 2018. [[View ticket](#)]
- [10] Jim Edwards, “Someone Is Trying to Take Entire Countries Offline and Cybersecurity Experts Say ‘It’s a Matter of Time Because It’s Really Easy.’ ” *Business Insider France*, December 22, 2018. [[Read online](#)]
- [11] Richard Wordingham via Unicode, “NNBSP (was: A last missing link for interoperable representation)”, Unicode Public Mailing List, [January 16, 2019](#).

Acknowledgments

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 and Google Books.