L2/20-008

Proposal to synchronize the Core Specification

For consideration by Unicode Technical Committee

2020-01-06
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

This proposal adds to the response to Action item 161-A1 in that it aims at synchronizing the Core Specification with changes already effected in other parts of the Unicode Standard, notably UAX #14, or suggested in *Proposal to make material changes to UAX #14* and *Proposal to make focused changes to the Code Charts text,* submitted simultaneously.

By coincidence, this proposal is also part of Unicode 13.0 beta feedback.

1 Narrow No-Break Space

TUS 12.0.0, ch. 6, 6.2 General Punctuation, *Space Characters,* p. 265

Change from:

Narrow No-Break Space. U+202F NARROW NO-BREAK SPACE (NNBSP) is a narrow version of U+00A0 NO-BREAK SPACE. The NNBSP can be used to represent the narrow space occurring around punctuation characters in French typography, which is called an "espace fine insécable." It is used especially in Mongolian text, before certain grammatical suffixes, to provide a small gap that not only prevents word breaking and line breaking, but also triggers special shaping for those suffixes. See "Narrow No-Break Space" in *Section 13.5, Mongolian,* for more information.

Change to:

Narrow No-Break Space. U+202F NARROW NO-BREAK SPACE (NNBSP) is ~~a narrow~~ best thought of as a non-breaking version of ~~U+00A0 NO-BREAK SPACE~~ U+2009 THIN SPACE. The NNBSP can be used as a numeric group separator in numerous scripts and to represent the narrow space occurring ~~around~~ next to certain punctuation characters in French ~~typography~~ text, ~~which~~ where it is currently called an "*espace fine (insécable)*~~.~~" [literally "(no-break) thin space" (supposed to be always non-breaking)]. ~~It~~ The NNBSP is used especially in Mongolian text, before certain grammatical suffixes, to provide a small gap that not only prevents word breaking and line breaking, but also triggers special shaping for those suffixes. See "Narrow No-Break Space" in *Section 13.5, Mongolian,* for more information.

Note about colors:

Highlighting is yellow for new text, lime green for reused, and ~~purple & barred~~ for deleted. That color scheme aims at distinguishing moved, copy-pasted or case-converted strings, from those that are added from scratch. Using another color for deletions (plus line through) is for easier fast-reading.

Rationale:

Here, TUS needs to reflect a change that was already made to UAX #14 for Unicode 12.0.0 (March 2019). UAX #14 in its last intermediate revision 42 (also L2/19-030 from 2018-10-03) already got this untrue information removed ("~~is a narrow version of NO-BREAK SPACE~~"). Some missing essential information should in my opinion also be added today, such as a commented English translation of the French term and a number of caveats so as to stick with facts instead of standing on thin air.

Details:

Despite its misleading name, NARROW NO-BREAK SPACE is by no means "a narrow version of NO-BREAK SPACE", because if it were, it would be justifying, yet it is not. From the beginning on, when it was encoded for Mongolian, NNBSP was intended to be fixed-width, and it has never departed from that design. The fact is that Unicode proposed Mongolian authorities to use NO-BREAK SPACE instead, and only after a point was made for Mongolian using both NO-BREAK SPACE and the Mongolian space, Unicode agreed to encode a new space character, but in the General Punctuation block as it was explicitly assumed that such a space could be useful to other scripts alike. Please see L2/19-112 and L2/19-115.

NNBSP, THIN SPACE and FOUR-PER-EM SPACE have a great variety of ratios to each other across fonts. This is mainly due to its underspecification since its encoding for Unicode 3.0, leaving font designers to themselves until in version Unicode 6.0.0 the chart of General Punctuation started hinting that NNBSP may be a no-break THIN SPACE or a no-break MID SPACE as well as anything else around these ("a narrow form of a no-break space, typically the width of a thin space or a mid space"), while stating in UAX #14 that SPACE and NO-BREAK SPACE are the only justifying spaces in ordinary contexts, so NBSP has no narrow form.

In practice, NNBSP is the interoperable THIN SPACE tailored as non-breaking in DTP (wysiwyg and TeX) where it is used to space off certain punctuation marks in French text, and as a numeric group separator. Consistently, the most-used fonts Times New Roman and Arial give NNBSP exactly the width of THIN SPACE. A test page comparing nine fonts can be seen on page 12 of *Proposal to make focused changes to the Code Charts text.*

I think that the best that Unicode can do today is delivering useful recommendations, rather than loosely commenting on a situation that it left unfolding during a decade (1999–2010). Unicode as a standards body is expected to provide more accurate information, especially about what NNBSP *should be,* not to merely delete the statement "~~is a narrow version of NO-BREAK SPACE~~" without putting anything in its place, as was done in UAX #14. As a consequence, I suggest the phrase "is best thought of as", and the concept of "a non-breaking version of U+2009 THIN SPACE."

The primary use of NNBSP is as a group separator in numbers for locales grouping digits into triads using space, listed in L2/19-112. These locales actually implement a recommendation from the International System of Units (BIPM). The requirement is that the group separator space be always narrower than a digit, not only by default, but always in display. Therefore, that use case should be mentioned first.

Next comes the use of NNBSP in French [1][2], since in 11.0.0 this was raised (see subsection *Space Characters*, because many direct bookmarks were disabled since that version), after it was added below the Mongolian usage of NNBSP in TUS 7.0.0 (see paragraph *Narrow No-Break Space*):

1. I'd say "next to" rather than "around", since for any related mark it occurs always on the same side (always before except for opening angle quotation marks [there are two in French too, the double or standard,

but also the single or half one, used by scholars and sometimes—and in some locales—for nested quotes]), never on the opposite side like what is done around EN DASH with justifying spaces.
2. I'd add "certain" as not all punctuation marks are concerned, only the tall ones ("double" marks—even COLON in new school like in some high-ranking print media—and single guillemet).
3. The term "punctuation characters" is specific to Unicode for "punctuation marks" and must stay.
4. I'd replace "typography" with "text" exactly like below "in Mongolian text." It's about a feature not only for typesetters but for everyone's current use. Moreover, part of the graphic industry keeps using systems doing without NNBSP by tailoring SPACE to the appropriate width and line breaking behavior, regardless of any considerations about interoperability as long as the output is for PDF and printed matter. NNBSP by contrast is for everyone to write with, possibly to post on the internet.
5. Then "is called" is connected to "French text" using "where".
6. We need to add "currently" because there is also another French name, in the Code Charts' French translation: *ESPACE INSÉCABLE ÉTROITE*. Sometimes people in France are fooled into taking this for the French name of that space, but since it's an early and maintained translation of the Code Charts, even a recent update proved really unable to adapt the character's disguise to current usage and couldn't help prolonging the quid pro quo.
7. I'd suggest italicizing "*espace fine insécable*" as a non-English term (and applying this rule throughout, just like when TUS italicizes "*virama*").
8. Parentheses around "*insécable*" are in fact necessary, because "*la fine*" was ever thought of as non-breaking. Only in the Unicode era need we to distinguish between the non-breaking one and the breaking one. Sometimes "*espace fine insécable*" is called a pleonasm.
9. That calls for an explanation inside the English-translation brackets, consistently parenthesized.
10. Adding an English translation of the French term is good practice, and here it is extremely useful as it shows how NNBSP is called in the locale that in Latin script makes the most extensive use of NNBSP.

When the sentence about usage in other scripts grows forcibly longer, the sentence about Mongolian usage needs to start with the repeated subject.

2 Space Characters

TUS 12.0.0, ch. 6, 6.2 General Punctuation, *Space Characters*, p. 264

Change from:

The main difference among other space characters is their width. U+2000..U+2006 are standard quad widths used in typography. U+2007 FIGURE SPACE has a fixed width, known as tabular width, which is the same width as digits used in tables. U+2008 PUNCTUATION SPACE is a space defined to be the same width as a period. U+2009 THIN SPACE and U+200A HAIR SPACE are successively smaller-width spaces used for narrow word gaps and for justification of type. The fixed-width space characters (U+2000..U+200A) are derived from conventional (hot lead) typography. Algorithmic kerning and justification in computerized typography do not use these characters. However, where they are used (for example, in typesetting mathematical formulae), their width is generally font-specified, and they typically do not expand during justification. The exception is U+2009 THIN SPACE, which sometimes gets adjusted.

In addition to the various fixed-width space characters, there are a few script-specific space characters in the Unicode Standard. U+1680 OGHAM SPACE MARK is unusual in that it is generally rendered with a visible horizontal line, rather than being blank.

Change to:

The main differences among other space characters ~~is~~ are their fixed width and their line breaking behavior. U+2000 EN QUAD and U+2001 EM QUAD are wide spaces that are a (canonically equivalent) breaking alternative to U+2002..U+2003 where the other fixed-width spaces are non-breaking. U+2002..U+2006 are standard quad widths ~~used in~~ originating from traditional (hot lead) typography, coded in the order ½, 1, ⅓, ¼, and ⅙. U+2007 FIGURE SPACE ~~has a fixed width, known as tabular width, which is the same width as digits used in tables.~~ and U+2008 PUNCTUATION SPACE have the width of a digit and of ~~is a space defined to be the same width as~~ a period, respectively, and can be used to align figures in tables. FIGURE SPACE is non-breaking. U+2009 THIN SPACE is used in numbers and next to punctuation if it is non-breaking; else, U+202F NARROW NO-BREAK SPACE is used instead. ~~and~~ U+200A HAIR SPACE ~~are~~ is ~~successively smaller width~~ the thinnest space~~s for narrow word gaps and~~ for manual justification ~~of type. The fixed-width space characters (U+2000..U+200A) are derived from~~ in ~~conventional~~ traditional ~~(hot lead)~~ typography. Algorithmic kerning and justification in computerized ~~typography~~ typesetting do not use these characters. ~~However, where they are used (for example, in typesetting mathematical formulae), their~~ width is generally font-specified, and they typically do not expand during justification. The exception is Unlike the other fixed-width spaces, U+2009 THIN SPACE~~, which sometimes~~ may ~~gets adjusted~~ slightly expand during justification when used in mathematical formulae.

In addition ~~to the various fixed-width space characters~~, there are a few script-specific space characters in the Unicode Standard. U+1680 OGHAM SPACE MARK is unusual in that it ~~is generally rendered with a visible horizontal~~ continues the stemline, ~~rather than being~~ but it is blank in stemless fonts. See *Section 8.12, Ogham*, for more information.

Change to (cleared):

The main differences among other space characters are their fixed width and their line breaking behavior. U+2000 EN QUAD and U+2001 EM QUAD are wide spaces that are a (canonically equivalent) breaking alternative to U+2002..U+2003 where the other fixed-width spaces are non-breaking. U+2002..U+2006 are standard quad widths originating from traditional (hot lead) typography, coded in the order ½, 1, ⅓, ¼, and ⅙. U+2007 FIGURE SPACE and U+2008 PUNCTUATION SPACE have the width of a digit and of a period, respectively, and can be used to align figures in tables. FIGURE SPACE is non-breaking. U+2009 THIN SPACE is used in numbers and next to punctuation if it is non-breaking; else, U+202F NARROW NO-BREAK SPACE is used instead. U+200A HAIR SPACE is the thinnest space for manual justification in traditional typography. Algorithmic kerning and justification in computerized typesetting do not use these characters. Unlike the other fixed-width spaces, U+2009 THIN SPACE may slightly expand during justification when used in mathematical formulae.

In addition, there are a few script-specific space characters in the Unicode Standard. U+1680 OGHAM SPACE MARK is unusual in that it continues the stemline, but it is blank in stemless fonts. See *Section 8.12, Ogham*, for more information.

Rationale:

I found the actual wording confusing. Today it looks to me like Unicode was burying facts under redundancies. The reader expects to learn what these spaces are used for, and this is also the place to report irregularities so as to assist the user. Patrick Andries [2] reports (section 6.3.2, page 144; see p. 2 of L2/19-115) that THIN SPACE is “unfortunately” breaking. Clearly line breaking behavior really matters and should be mentioned, notably the fact that FIGURE SPACE is non-breaking. Discovering such things by themselves only makes people suspicious: Why does Unicode not disclose those facts in the first place? I suggest a thorough rewording by cutting down on redundancies and adding essential information.

Details:

The behavior in line breaking is another main difference between space characters, also in the range U+2000..U+200A since there is one no-break space (FIGURE SPACE). The Standard cannot skip that particular information, and it should also cover the entire range as of these spaces' line breaking behavior, so that the user knows what issues to expect.

The fact that the width of these spaces is fixed should be mentioned in the first place. Actually these two paragraphs mention "fixed(-|)width" three times, but first only about FIGURE SPACE. Only below is the entire range declared as fixed-width. That is not straightforward. In the suggested rewording I keep using "fixed" three times, but for the first one I don't wait until presenting U+2007. By adding "fixed" as an attribute before the first instance of "width" I suggest noting both the difference with SPACE mentioned in the preceding paragraph, and introducing width as a key parameter.

One disturbance in the actual text is that only the second half of the range is presented in detail, character after character, while the first half is obfuscated by only mentioning their overall design principle. I'd suggest shedding some light on the hidden part of that range, starting with the two canonically equivalent wide spaces. These are not in the XCCS. The documentation of the Xerox Coded Character Standard only mentions "quad" as an alias of "space" for these characters. Never could Unicode duplicate them "by mistake." An explanation is wanted as of why there is a set of canonically equivalent wide spaces at range start, and this is the single best place in TUS to provide it.

Still I'm very cautious about the suggested wording. It is crystal clear to me that Unicode cannot disclose the full history, nor will the eyewitnesses. Therefore I'd suggest simply sticking with the facts as we know them today, namely that the duplicate spaces are used in TeX to provide a breaking alternative to the two widest spaces of the range U+2002..U+200A, that is non-breaking there throughout: "U+2000 EN QUAD and U+2001 EM QUAD are wide spaces that are a (canonically equivalent) breaking alternative to U+2002..U+2003 where the other fixed-width spaces are non-breaking." Based on that hint, the user will be able to start making a better sense of the range than when stumbling right over those duplicates while checking the Code Charts, then wondering why TUS refrains from commenting on them.

I'd suggest being precise about what "typography" is referred to as soon as in the first instance. Rather than "used in typography" contradicting the information about computerized typography below, I'd put "originating from traditional (hot lead) typography," where "originating from" replaces "used in" both because these spaces are not used today the way they were by the time, and because EM and EN SPACE are not restricted to typography proper but can be used everywhere the system has basic Unicode support (examples in *Proposal suggesting formal edits to UAX #14,* item 3.80); where "traditional" replaces "conventional" because the change is in tradition, not in convention, and "conventional" is negatively connotated; and where "(hot lead)" is raised from below.

The range description would not meet reader expectations without mentioning the encoding order, since that is not straightforward: the half space comes first, then the full one, and only the other fractions are in decreasing order: "coded in the order ½, 1, ⅓, ¼, and ⅙."

FIGURE SPACE and PUNCTUATION SPACE should in my opinion be treated together since they are used together, and unlike the actual text of TUS I'd advocate being brief and to-the-point by deleting the lengthy phrases about tabular width, and rather adding a useful hint about the actual use case instead of letting the reader figure it out: "and can be used to align figures in tables."

In the wake, TUS should mention that "FIGURE SPACE is non-breaking." I don't suggest explaining why so. To find it out, I needed to check TUS 1.0 and TUS 2.0, quoted in the *Background* section below. TUS 13.0.0 is only expected to draw attention to what is noted in the compatibility decomposition mapping found in the Code Charts.

Since THIN SPACE is breaking, it cannot be used as a numeric group separator and to space off certain punctuation marks, but the fact is that this is its regular use case much more than "narrow word gaps," and it is still real where THIN SPACE is tailored as non-breaking. TUS cannot get around mentioning that fact, at least in my opinion: "U+2009 THIN SPACE is used in numbers and next to punctuation if it is non-breaking;" For the sake of standard Unicode implementations, I also insist on directing the reader to NARROW NO-BREAK SPACE within the description of THIN SPACE. The reader anyway finds out when reading on, so doing it here is simply a matter of good form: "else, U+202F NARROW NO-BREAK SPACE is used instead."

The rest of the proposed first paragraph is self-explaining based on the above, and results from a streamlining effort that makes it finish in: "Unlike the other fixed-width spaces, U+2009 THIN SPACE may slightly expand during justification when used in mathematical formulae." (Deletions not quoted.)

The second paragraph about script-specific spaces has focus on OGHAM SPACE MARK. I'd suggest adapting its content to stick closer with Michael Everson's feedback L2/07-392 not stating that the presence of the stemline is general in Ogham fonts. The first paragraph is quoted from page 9 of Irish Standard 424:1999.

> 1. The centre line is optional. In printing and in manuscript Ogham it is conventional to design with a centre line, but this is not necessary. In implementations without the centre line, the character OGHAM SPACE MARK should be given its conventional width, and simply left blank like SPACE.
>
> See http://www.evertype.com/standards/iso10646/pdf/is434.pdf
>
> The OGHAM SPACE MARK **may** have a visible glyph. But it also **may not** have a visible glyph. And when it does have a visible glyph, that glyph is not actually part of the letter, any more than the stemline in OGHAM LETTER FEARN is a part of the letter. OGHAM LETTER FEARN is three strokes to the right of the edge of the stone. That edge may be drawn in a font (in which case OGHAM SPACE MARK should have a visible glyph) or it may not be (in which case OGHAM SPACE MARK should not have a visible glyph). The stemline is more an indication of layout; it is not an integral part of the letter.

Actually TUS is stating that OGHAM SPACE MARK "is generally rendered with a visible horizontal line, rather than being blank." I'd suggest changing this to: "continues the stemline, but it is blank in stemless fonts."

Because of the word "stemline" and for better support like what TUS already does for Mongolian (see section 1 above) and many other scripts, I also suggest adding a cross-reference: "See *Section 8.12, Ogham,* for more information."

Background

The significance of FIGURE SPACE changed in Unicode 3.0 when TUS (unlike UAX #14) started highlighting its relationship with typesetting tables. The reader is welcome to figure out how the digit-wide glyph of this space integrates with digits so as to emulate the effect of a decimal tab stop. Another easy guess is how PUNCTUATION SPACE fills in the gap where other lines have a thousands separator (example on page 4 of L2/19-115).

FIGURE SPACE indeed got its name not because it should be used *in* figures, but because it is used *before* figures, for indentation and alignment purposes. It was so in typesetting tables the old-fashioned way, hence its alias “tabular space” (currently ESPACE TABULAIRE in French), which is in another way confusing, given that Unicode has actually two tabular spaces, PUNCTUATION SPACE being the other one. Both may still be used so in plain text; they were designed for proportional fonts, where they make actual sense.

Long before UAX #14 was set up—its first available Draft version dates from May 1998 and is de facto (draft) version Unicode 2.1.2—the Unicode 2.0 Core Specification already stated that FIGURE SPACE is intended to be used as a group separator in numbers, after Unicode 1.0 had provided it for that purpose. TUS 2.0 reads:

> ***Space Characters***
>
> ***Typographical Space Characters.*** Spaces generally have the semantics of being word-break characters. Other than that, the main difference is in the width of the characters. U+2000 → U+2006 are standard quad widths used in typography. U+2007 FIGURE SPACE is intended to be used as a thousands separator in cases where countries use space to separate groups of digits. Typically it has a fixed width the same size as a digit in a particular font. U+2007 FIGURE SPACE behaves like a numeric separator for the purposes of bidirectional layout. ([…]) U+2008 PUNCTUATION SPACE is a space defined to be the same width as a period. U+2009 THIN SPACE and U+200A HAIR SPACE are successively smaller-width spaces used for narrow word gaps and for justification of type. All of the fixed-width space characters are derived from conventional (hot lead) typography. Their functions are mostly replaced by algorithmic kerning and justification in computerized typography. Characters drawn with a dotted box are invisible in normal rendering.
>
> Zero-width space characters can be used in languages that have no visible word spacing in order […] […] U+200B ZERO WIDTH SPACE may be significant for searching or sorting operations.
>
> - ➥ It is important to note that not all space characters have word- or line-breaking properties.
>
> Space characters may also be found in other character blocks in the Unicode Standard. […]

So far on page 6-68 of *The Unicode Standard Version 2.0.* For reference, page 75 of *The Unicode Standard Version 1.0* was slightly different only by the absence of some information:

> *Typographical Space Characters.* Spaces all have […]. Other than […]. U+2000 → U+2006 […]. The figure space is provided for use in some languages as a thousands separator. The punctuation space is […]. The thin space and hair space are […]. All of […]. Their functions […].
>
> The zero-width space can […] […] for searching or sorting operations.

As soon as Mongolian entered for Unicode 3.0 in 1999, and the NARROW NO-BREAK SPACE with it, editing the Core Specification resulted in deleting the point about FIGURE SPACE as a group separator (p. 149–150):

> ***Space characters***
>
> The most commonly used space character is U+0020 SPACE. Also often used is its non-breaking counterpart, U+00A0 NO-BREAK SPACE. These two characters have the same width, but behave differently for line breaking. U+00A0 NO-BREAK SPACE behaves like a numeric separator for the purposes of bidirectional layout. ([…]) In ideographic text, U+3000 IDEOGRAPHIC SPACE is commonly used because its width matches that of the ideographs.
>
> The main difference among other space characters is their width. U+2000..U+2006 are standard quad widths used in typography. U+2007 FIGURE SPACE has a fixed width, known as tabular width, which is the same width as digits used in tables. U+2008 PUNCTUATION SPACE is a space defined to be the same width as a period. U+2009 THIN SPACE and U+200A HAIR SPACE are successively smaller-width spaces used for narrow word gaps and for justification of type. The fixed-width space characters (U+2000..U+200A) are derived from conventional (hot lead) typography. Algorithmic kerning and justification in computerized typography do not use these characters. However, where they are used, as, for example, in typesetting mathematical formulae, their width is generally font-specified, and they typically do not expand during justification. The exception is U+2009 THIN SPACE, which sometimes gets adjusted.
>
> Space characters with special behavior in word or line breaking are described in "Line and Word Breaking" in Section 13.2, Layout Controls.
>
> Space characters may also be found in other character blocks in the Unicode Standard.

This full quote shows that the locale preference for any space as a group separator is no longer supported. (The only instance of "thousands" in Chapter 6 of TUS 3.0 is about the decimal or thousands separator as a dot, and no instance of "group" matches.) NARROW NO-BREAK SPACE is not mentioned either outside Mongolian, despite it was encoded outside the Mongolian block, in General Punctuation, to cater for the needs of other scripts. Prior to being introduced in section 6.2 of TUS 6.1 (2012), NNBSP probably needed to be supported in fonts outside Mongolian.

Please see also *Proposal to make material changes to UAX #14.*

References

[1] Deborah Anderson, Ken Whistler, Roozbeh Pournader, Lisa Moore, Liang Hai, *Recommendations to UTC #160 July 2019 on Script Proposals,* p. 14, #20 [L2/19-286].

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

[3] Jakub Stachu, in "Non-breakable space justification in Word 2016", page 3, Microsoft Community, 2017.

[4] Jukka "Yucca" Korpela, "*Unicode line breaking rules: explanations and criticism",* *IT and communication,* 2000-10-11, 2005-08-28/29, 2008-05-21.

[5] Shriramana Sharma, "NBSP supposed to stretch, right?", Unicode Public Mailing List, December 2019.

Acknowledgments

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

Thanks to Microsoft for Word Online, OneDrive, VS Code and MSKLC.

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