L2/20-088

Proposal to adjust space characters in UAX #14

For consideration by the Unicode Technical Committee

2020-03-27
Marcel Schneider (charupdate@orange.fr)

We should always tell what we see.
Above all we should always
—which is more difficult—
see what we see.

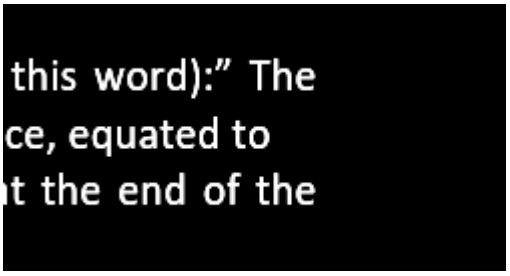
Charles Péguy, 1910

This proposal is part of the solution suggested in *Proposal to fix Hebrew in UAX #14 by ruling out LB21a..b,* submitted simultaneously. It is also a part of the set designed to replace the *Proposal to re-engineer spaces and punctuation in UAX #14* announced in L2/20-005 *Proposal to make material changes to UAX #14,* Due to the actual pandemic, this is the last part to be worked out and submitted so far. A missing proposal about spaces and punctuation in UAX #14 and other proposals are postponed.

1 Problems

1.1 Most breaking spaces do not hang into the margin

UAX #14 treats breaking spaces other than SPACE as visible word separator characters (subsection 5.7), causing them to leave "a visual 'hole' or indentation in the flush edge." While optional for hanging punctuation, hanging into the margin is mandatory for breaking spaces. Here, UAX #14 introduced a design flaw causing unexpected behavior at line end, and pretends that fixing this is outside its scope, while for SPACE it is inside. It makes a special case for U+1680, while OGHAM SPACE MARK is only one example of the bug. Through assigning breaking spaces the line breaking class BA instead of SP, Unicode made them definitely unusable in flowing text, since except IDEOGRAPHIC SPACE—see L2/10-309 *Comments on UAX#14* from Eric Muller—breaking spaces are expected to hang into the margin, as they do today in TeX—and most probably did so since the 1980ies—but not in Unicode implementations like Word Online, used to write this up:



Unicode-conformant but wrong behavior of breaking EM SPACE at line end, with or without following SPACE, in Word Online.

Assigning BA to spaces was a mistake that has no presentable explanation, yet OGHAM SPACE MARK is the only BA space with a weak recommendation to palliate the resulting unexpected behavior. Not only should that recommendation be extended to em and en spaces. The problem must be addressed properly. UAX #14 admits that BA causes inappropriate behavior, so it needs to recognize that restricting SP to SPACE was wrong. SPACE is not the only space causing indirect word breaks. The exact same holds true for all breaking spaces used in text, except IDEOGRAPHIC SPACE, that has different semantics.

Calling something “a TeX convention that does not follow general typography” is to put things upside doxn: Donald Knuth carefully implemented good, traditional, “general” typography; therefore, TeX is a template of what Unicode is expected to perform in plain text. That was well understood at Unicode prior to implementing an overthrow designed to limit the Standard’s usability, as explained below under 3.7.

1.2 Too many spaces are breaking instead of non-breaking

The narrower fixed-width spaces are overwhelmingly expected to be non-breaking throughout. Assigning these spaces the line breaking class BA instead of GL is extremely disruptive to pre-Unicode practice, that keeps widely overlapping with Unicode practice via both application-specific tailoring and unawareness of end-users assuming regular behavior, where “regular” means “like users expect it based on what they know from pre-Unicode practice and what they see in accurately tailored environments.”

Bug fixes are often dismissed to tailoring even inside a given locale, while tailoring is thought of as a means to meet expectations on a per-locale basis. Heavily relying on application-specific tailoring inside a given locale should be considered abusive, as opposed to locale-specific tailoring that is typically considered a Unicode-conformant feature. The only exception is mathematics, that may rely on very specific tailorings implemented in dedicated TeX packages or specific applications, and are outside the scope of this proposal.

Despite U+202F NARROW NO-BREAK SPACE has been encoded in the block General Punctuation in 1999, in 2010 Hebrew communities still complained about not having enough non-breaking fixed-width spaces, we assume in *Proposal to fix Hebrew in UAX #14 by ruling out LB21a..b.* For Hebrew indeed, that problem has then been solved badly since Unicode 6.1.0 (2012) by adding rule LB21a, a neither sustainable nor Unicode-conformant “quick-and-dirty” solution harmful to Hebrew itself.

Action 3 of L2/19-115 *Proposal to ensure usability of fixed-width spaces* was about “[c]hang[ing] the Line_Break property value of U+2002..U+2006 and U+2008..U+200A from BA to GL,” but at meeting #159, the UTC advised (159-A14) that this “would he too disruptive to existing data.” However, that proposal was based on the assumption that the canonical equivalence of U+2000 to U+2002 and of U+2001 to U+2003 should be cancelled after adding a clause to the decomposition stability policy “stipulating that decomposition mappings **defined in contradiction with character identity** may be canceled.” **(Bold added.)** UTC determined that this is not possible. That substantially compromised the proposal through not fulfilling the precondition. UTC’s advice about disruption to existing data should be interpreted in this light, which allows to be specific: Existing data would be disrupted only as far as U+2002 and U+2003—and their canonically decomposables U+2000 and U+2001—are involved, accumulated evidence tends to prove.

2 Solution

The only viable solution consists in changing the line breaking class assignments as shown in the following table.

1. The upper five space characters U+1680 and U+2000..U+2003 are expected to cause indirect line breaks so as to hang into the margin. Therefore, they need to be assigned line breaking class SP.
2. The lower six spaces U+2004..U+200A \ U+2007 are expected to not provide a line break opportunity, so they need to be GL. This change is not disruptive to existing data, except to the virtually nonexistent data not following general typography or opened in applications not tailored to meet the requirements of general typography. and not yet converted to a stable format such as PDF.

CP	Name	Old line breaking class	New line breaking class
1680	OGHAM SPACE MARK	BA	SP
2000	EN QUAD	BA	SP
2001	EM QUAD	BA	SP
2002	EN SPACE	BA	SP
2003	EM SPACE	BA	SP
2004	THREE-PER-EM SPACE	BA	GL
2005	FOUR-PER-EM SPACE	BA	GL
2006	SIX-PER-EM SPACE	BA	GL
2008	PUNCTUATION SPACE	BA	GL
2009	THIN SPACE	BA	GL
200A	HAIR SPACE	BA	GL

Some space characters are not included in this solution, because they already behave appropriately or are outside the scope of this proposal:

1. U+205F MEDIUM MATHEMATICAL SPACE (class BA) is outside the scope of this proposal, given that mathematical formulae are typically handled by TeX-based or specialized rendering engines and therefore may not need any fully functional plain text representation, even less using MMSP.
2. U+3000 IDEOGRAPHIC SPACE has been accurately assigned the line breaking property value BA, because “it is treated as any other CJK ideograph, and by default remains in the measured part of lines when at a line extremity (i.e. it does not disappear or hang in the margin like a U+0020)”, Eric Muller reported in L2/10-309 *Comments on UAX#14.*

3 Rationale

This section is basically about accumulating evidence of practice and subsequent expectations related to fixed-width spaces, and tends to prove that there are virtually zero existing documents that would be disrupted when upgrading from Unicode 13.0.0 to Unicode 14.0.0 while the line breaking property values of fixed-width spaces are changed for Unicode 14.0.0 as suggested above.

3.1 Hebrew script

Rule LB21a of the Unicode Line Breaking Algorithm showcases how much the Hebrew user communities are interested in non-breaking fixed-width spaces, to such an extent that a secret lobbying effort seems to have been directed at Apple and picked up there by Unicode. Since Hebrew user communities are okay that U+2000..U+200A are all non-breaking (not only U+2007), while Hebrew has nothing particular with respect to these spaces that Cyrillic, Greek, Latin and so on would not have, rule LB21a is strong evidence that these spaces can change line breaking class, except perhaps the two (four) largest ones U+2000..U+2003, since these were duplicated, as pointed out further down this list.

Please see *Proposal to fix Hebrew in UAX #14 by ruling out LB21a..b.*

3.2 French style guide from Louis Guéry of the Centre of Journalists Perfection (CFPJ)

Yet this interest is not proper to Hebrew. For example, Latin user communities likewise are interested in a full range of non-breaking fixed-width spaces. As a consequence, pre-Unicode computerized typesetting engines currently provided the various fixed-width spaces as non-breaking spaces. Consistently, one of the most

renowned and widely taught and followed French style guides, the *Dictionnaire des règles typographiques* (Lexicon of Typographical Rules) [1] from Louis R. M. Guéry (1919–2016) [2], journalist, newspaper editor in chief, founder and head (1969–1984) of the Centre de formation et de perfectionnement des journalistes CFPJ (Centre of Journalist Perfection), states on page 90:

> ◆ les espaces à valeur fixe qui sont toujours insécables (voir ce mot) :
> – le cadratin, espace dont la largeur est celle de la force de corps du caractère : 12 points dans un corps 12, 7 dans un corps 7, etc. ;
> – le demi-cadratin ;
> – le quart de cadratin, parfois appelé espace fine, ou tout simplement fine :
> C'est par une fine que l'on sépare les tranches de trois chiffres.

The first line translates to: "the fixed-width spaces that are always non-breaking (see this word):" The second-level bullet list enumerates the em space, the en space, and the four-per-em space, equated to the thin space, that is used as a group separator in numbers, the example sentence at the end of the snippet states.

That is strong evidence that all fixed-width spaces were assumed to be non-breaking in pre-Unicode usage.

3.3 French National Printing Office (Imprimerie Nationale) style guide

Consistently, the style guide of the French National Printing Office [3], on page 148 about punctuation spacing, refers to the thin space without being specific about its line breaking behavior, obviously assuming that it is non-breaking:

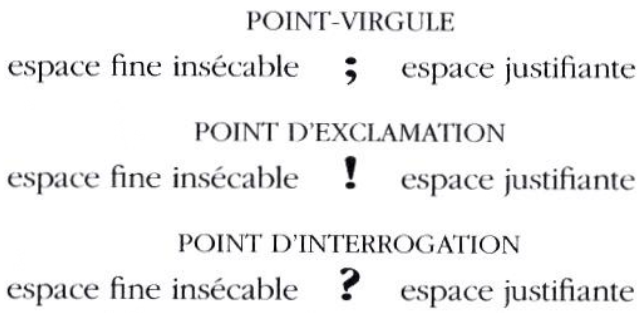
> – une espace fine doit être placée devant le *point-virgule*, le *point d'exclamation* et le *point d'interrogation* (qui ne seront jamais collés au mot qui précède) ;

Translation: "– a thin space must be placed before the semicolon, the exclamation mark and the question mark (that shall never be glued to the word that precedes);"

There is to say that this source represents the transition from French old school typesetting, acting like quoted above but using a justifying no-break space with colon and quotation marks (the latter already typeset with thin space while still ruling otherwise), to French new school typesetting that uses thin space throughout.

The French term "espace fine insécable" quoted in TUS is used in the overview table on the next page (page 149) of the same source [3]. That demonstrates how in French, the full name of the non-breaking thin space was used mainly when there was plenty of space, while the non-breaking behavior was so obvious that authors sometimes felt no need to mention it in flowing text.

(The scanned snippet is again limited to the intersection of old school and new school.)



3.4 The International System of Units

Unicode could not ignore that FIGURE SPACE is unfit to separate groups of digits in SI-conformant numbers, since the International Bureau of Weights and Measures (BIPM) decided in 1948, when establishing the International System of Units (SI), that digits shall be grouped using space (see resolution #7 of the 9th General Conference on Weights and Measures (CGPM), referenced on page 133 (139 of PDF) of [4], subsection 5.3.4, and quoted in full text on page 146), and in 2003 at the 22nd CGPM, resolution #10, still refrained from specifying the width of the digit-grouping space. Only the brochure, at least in its 8th edition (2006), is more specific **(bold added):**

> 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.

But the narrow width of the group separator was an obvious assumption in 1948 and in 2003, just like its non-breaking behavior remains still implicit.

The SI rule is mirrored on page 37 of [5]. Please see a more comprehensive overview on pages 4 through 6 of L2/19-112 *Proposal to define a space character as a group separator*.

Obviously, at least Adobe knew—while helping set up Unicode—that without a non-breaking thin space, a usable plain text representation of many languages is impossible.

3.5 Donald E. Knuth's TeX

Long before Unicode was set up, "thin space" was handled in TeX as a non-breaking six-per-em space through the macro "\kern .16667em", as explained already on pages 5 and 10 (pages 16 and 21/494 of PDF) of the TeXbook (1983) [6]: "[…] you certainly won't want to type ' ', because that space is much too large—it's just as large as the space between words—and TEX might even start a new line at such a space when making up a paragraph! The solution is to type '\thinspace'' […]."

As per page 352 (363 of PDF) of *The TeXbook* [6], TeX had in 1983—and still has—both breaking and non-breaking fixed-width spaces, Jonathan Coxhead reported on Unicode Public on July 10, 2000. The following table sums up the spaces natively provided by TeX:

TeX breaking space macros	TeX non-breaking space macros	Width equivalent
\qquad	–	2 × EM SPACE
\quad	–	EM SPACE
\enskip	\enspace	EN SPACE
–	\thinspace	THIN SPACE

This table allows to easily check that as of fixed-width spaces, the word “quad” is preferably used for breaking spaces, while the word “space” is preferably used for non-breaking spaces. The duplicate presence of the en space in two macros, one breaking and the other non-breaking, is also striking. These observations are enough to explain why Unicode could come up with a fully-fledged whitespace array featuring both breaking and non-breaking spaces as of em and en spaces, but only non-breaking spaces as of narrower spaces. So that scheme prefigures as a template what Unicode designed a few years later, because TeX was so wildly successful that it established itself as a de-facto writing and publishing standard on the marketplace, that Unicode was committed to follow under its compatibility principle.

Another important point about these breaking spaces in TeX **hanging into the margin** should also be noted here, because Unicode missed that point by limiting the line breaking class SP to SPACE instead of extending it to include all breaking spaces, among which U+1680 OGHAM SPACE MARK to begin with.

3.6 The Xerox Character Code Standard

TeX was able to generate a non-breaking thin space by calling a macro for ⅛ em advance width at a time (1977–1983) when the simultaneously upcoming (1980) Xerox Character Code Standard had neither a thin space nor a six-per-em space, as it probably unified it with the punctuation space 0xEE 0x24. A possible rationale relies on the equal advance width of numbers grouped by comma or period and those grouped by punctuation space. This is based on character set 0xEE as shown in a *table on Wikipedia*.

3.7 Unicode 1.0

So Unicode was going to encode the full range of fixed-width spaces, plus a breaking em space and a breaking en space, because it was well understood that while Unicode was committed to stick with legacy practice and to be fully backwards compatible, it was also important to look forward and see that the two largest spaces would be used also in contexts where they should provide a break opportunity and, of course, hang into the margin.

For getting there, Unicode had to make the most of the XCCS. Therefore, as Ken Whistler witnessed on Unicode Public, July 7, 2000: “The EM SPACE and EN SPACE were added to fill out the set of fixed-width spaces, which themselves were mostly derived from XCCS 1980.” We can interpret this statement by considering that Unicode needed to complete the set of fixed-width non-breaking spaces after using up the two quads to encode breaking large spaces, and made a clever use of some aliases provided in XCCS and reported by Ken Whistler on Unicode Public the same day, as the discussion went on:

Code(s)	Name	Also called	Width
357 | 55	em quad	em space	1.0 em fixed width
357 | 54	en quad	en space	½ em fixed width

Prior to releasing version 1.0 of the Standard, Unicode hustled this well-formed encoding scheme, presumably triggering one of the “too many crises” that saw “authors dropping like flies”, obviously because nobody wanted to endorse the biased paragraph published on page 75 of TUS 1.0 about “*Typographical Space Characters*” alleging that “Spaces all have the semantics of being word-break characters” prior to contradicting itself by pretending three sentences further that “The figure space is provided for use in some languages as a thousands separator”, while providing no hint about what the punctuation space, mentioned next, should be used for. (As we know today, FIGURE SPACE and PUNCTUATION SPACE help typeset tables in the absence of decimal tab stops.)

So that crisis would have been the pulldown of the typographical whitespace range, turning it into a zone of limited usability as we know it. Consistently, Patrick Andries later complained in § 6.3.2, page 144 of [7] (see longer quote on page 2 of L2/19-115 *Proposal to ensure usability of fixed-width spaces*):

> […] unfortunately, THIN SPACE is breaking.

References

[1] Louis Guéry, *Dictionnaire des règles typographiques,* CFPJ-Éditions, 1994; 5th edition, ediSens, Paris, 2019, ISBN 978-78235113-357-6.

[2] Geneviève Dermenjian, Bruno Duriez, "Notice GUÉRY Louis, René, Marie", *Le Maitron,* 2010-07-08, last updated 2017-02-17 [online].

[3] Collectif Imprimerie nationale (France). *Lexique des règles typographiques en usage à l'Imprimerie nationale.* Printing March 2017, Paris, Imprimerie nationale, 2008, ISBN 978-2-7433-0482-9.

[4] 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 [online].

[5] 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 [online].

[6] Donald E. Knuth, *The TeXbook,* Addison–Wesley Publishing Company, Reading, Mass., 1983 [online].

[7] Patrick Andries, *Unicode 5.0 en pratique : codage des caractères et internationalisation des logiciels et des documents*, Dunod, Paris, 2008 [online].