L2/20-010

Proposal to make focused changes to the Code Charts text

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 as it is complementary to *Proposal to extend support for abbreviations,* itself added to the response.

Beyond that initial scope, additional suggestions related to the same topic of normative abbreviations (aliases) and informative aliases in the Code Charts are aggregated, and some supplemental edits are also suggested in the process, so as to avoid overly breaking up simultaneous proposals into small units.

Materially this proposal is about editing the file NamesList.txt in the Unicode Character Database.

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

1 SPACE U+0020

Change from:

```
0020    SPACE
        * sometimes considered a control code
        * other space characters: 2000-200A
        x (no-break space - 00A0)
        x (zero width space - 200B)
        x (word joiner - 2060)
        x (ideographic space - 3000)
        x (zero width no-break space - FEFF)
```

Change to:

```
0020    SPACE
        * sometimes considered a control code
        * currently abbreviated as SP
        * other space characters: 2000-200A
        x (no-break space - 00A0)
        x (zero width space - 200B)
        x (narrow no-break space - 202F)
        x (word joiner - 2060)
        x (ideographic space - 3000)
        x (zero width no-break space - FEFF)
```

Rationale:

Some annotations yielding abbreviations are missing from the Code Charts, while those abbreviations are already in the UCD and are widely used. SPACE, although not called “SP” in TUS, is so in UAX #14, and “SP” is in NameAliases.txt. By contrast, in the Code Charts, “SP” is only the mnemonic in the dashed box (represented as a pair of brackets in the Code Charts column of the table below).

cp	Name	NamesAlias.txt	Core Spec	Code Charts	UAX14 lists	UAX14 else
0020	SPACE	SP	SPACE	[SP]	(SP)	SP

The Core Specification not using “SP” is consistent with the Unicode 1.0 Code Charts where the dashed box of SPACE is labeled “SPACE”, but since Unicode 2.0, “SP” is fairly common. Therefore I’d suggest adding that annotation to the entry of SPACE, where “currently” is used instead of “commonly” with respect to TUS that does not use the abbreviation.

In the Code Charts, the particle “as” after “abbreviated” is used the first three times out of 24. This proposal consistently suggests the more streamlined form without a particle (also common in other locales) for all code points above U+034F.

Adding a pointer to U+202F NARROW NO-BREAK SPACE is suggested with respect to the importance of that character in writing numerous languages in various scripts (listed in L2/19-112), at least when it comes to writing numbers in digits. A dedicated cross-reference is the more necessary as U+202F is outside the range pointed in an annotation (although it is cross-referenced there from THIN SPACE), and even more as five space-related characters are already cross-referenced from SPACE, among which NO-BREAK SPACE and ZERO WIDTH NO-BREAK SPACE.

2 LEFT-TO-RIGHT ISOLATE..POP DIRECTIONAL ISOLATE U+2066..2069

Change from:

```
@                Format characters
2066     LEFT-TO-RIGHT ISOLATE
2067     RIGHT-TO-LEFT ISOLATE
2068     FIRST STRONG ISOLATE
2069     POP DIRECTIONAL ISOLATE
```

Change to:

```
@                Format characters
2066     LEFT-TO-RIGHT ISOLATE
         * commonly abbreviated LRI
2067     RIGHT-TO-LEFT ISOLATE
         * commonly abbreviated RLI
2068     FIRST STRONG ISOLATE
         * commonly abbreviated FSI
2069     POP DIRECTIONAL ISOLATE
         * commonly abbreviated PDI
```

Rationale:

This is an exact synchronization with what is found in the first bidi-control range U+200E..U+200F and in the second one at U+202A..U+202E:

```
202E     RIGHT-TO-LEFT OVERRIDE
         * commonly abbreviated RLO
```

These four characters are the only bidi controls that do not have their abbreviation brought to the Code Charts reader in an annotation, as symbolized in this table where the brackets in the Code Charts column stand for "only in the glyph inside the dashed box":

cp	Name	NameAliases.txt	Core Spec	Code Charts
2066	LEFT-TO-RIGHT ISOLATE	LRI	LRI	[LRI]
2067	RIGHT-TO-LEFT ISOLATE	RLI	RLI	[RLI]
2068	FIRST STRONG ISOLATE	FSI	FSI	[FSI]
2069	POP DIRECTIONAL ISOLATE	PDI	PDI	[PDI]

Adding these annotations in the Code Charts is the more mandatory as the Core Specification is currently using these abbreviations too, as shown in the table above.

2 OGHAM SPACE MARK U+1680

Change from:

```
@              Space
1680     OGHAM SPACE MARK
         * glyph is blank in "stemless" style fonts
         x (space - 0020)
```

Change to:

```
 @             Space
1680     OGHAM SPACE MARK
         * glyph is blank in "stemless" style fonts
         * abbreviated OGSP
         x (space - 0020)
```

Rationale:

The Ogham space is among the spaces lacking a standard abbreviation. The simultaneously submitted *Proposal to extend support for abbreviations* encompasses those spaces. OGHAM SPACE MARK is particular in that its glyph in the Code Charts is lacking also a mnemonic despite the stemline is optional. Synchronizing it is part of Proposal to synchronize seven glyphs in the Code Charts. The mnemonic "OGSP" is found in L2/08-142.

Alongside completing NameAliases.txt, abbreviations newly added to NameAliases.txt should also be introduced in the Code Charts by adding annotations in NamesList.txt. For instance, neither "commonly" nor "currently" are suggested to be employed, unless common or current usage is established.

Although in the Unicode Standard, regularity is only a remote ideal, consistently with the diversity of the world's writing systems, I'm on the side of those thinking that spaces are some of the good places to keep promoting consistency.

3 EN QUAD..HAIR SPACE U+2000..U+200A

Change from:

```
2000    EN QUAD
        : 2002 en space
2001    EM QUAD
        = mutton quad
        : 2003 em space
2002    EN SPACE
        = nut
        * half an em
        # 0020 space
2003    EM SPACE
        = mutton
        * nominally, a space equal to the type size in points
        * may scale by the condensation factor of a font
        # 0020 space
2004    THREE-PER-EM SPACE
        = thick space
        # 0020 space
2005    FOUR-PER-EM SPACE
        = mid space
        # 0020 space
2006    SIX-PER-EM SPACE
        * in computer typography sometimes equated to thin space
        # 0020 space
2007    FIGURE SPACE
        * space equal to tabular width of a font
        * this is equivalent to the digit width of fonts with fixed-width
        digits
        # <noBreak> 0020
2008    PUNCTUATION SPACE
        * space equal to narrow punctuation of a font
        # 0020 space
2009    THIN SPACE
        * a fifth of an em (or sometimes a sixth)
        x (narrow no-break space - 202F)
        # 0020 space
200A    HAIR SPACE
        * thinner than a thin space
        * in traditional typography, the thinnest space available
        # 0020 space
```

Change to:

```
2000    EN QUAD
        * currently abbreviated NQSP
        : 2002 en space
2001    EM QUAD
        = mutton quad
        * currently abbreviated MQSP
        : 2003 em space
2002    EN SPACE
        = nut
        * half an em
        * currently abbreviated ENSP
        # 0020 space
2003    EM SPACE
        = mutton
        * nominally, a space equal to the type size in points
        * may scale by the condensation factor of a font
        * currently abbreviated EMSP
        # 0020 space
2004    THREE-PER-EM SPACE
        = thick space
        * abbreviated THPMSP
        * also abbreviated 3/MSP
        # 0020 space
2005    FOUR-PER-EM SPACE
        = mid space
        * abbreviated FPMSP
        * also abbreviated 4/MSP
        # 0020 space
2006    SIX-PER-EM SPACE
        * in computer typography sometimes equated to thin space
        * abbreviated SPMSP
        * also abbreviated 6/MSP
        # 0020 space
2007    FIGURE SPACE
        * space equal to tabular width of a font
        * this is equivalent to the digit width of fonts with fixed-width digits
        * currently abbreviated FSP
        # <noBreak> 0020
2008    PUNCTUATION SPACE
        * space equal to narrow punctuation of a font
        * currently abbreviated PSP
        # 0020 space
2009    THIN SPACE
        * a fifth of an em (or sometimes a sixth)
        * currently abbreviated THSP
        x (narrow no-break space - 202F)
        # 0020 space
200A    HAIR SPACE
        * thinner than a thin space
        * in traditional typography, the thinnest space available
```

```
* currently abbreviated HSP
# 0020 space
```

Rationale:

All these spaces are lacking a standard abbreviation. The simultaneously submitted *Proposal to extend support for abbreviations* encompasses those spaces.

Alongside completing NameAliases.txt, abbreviations newly added to NameAliases.txt should also be introduced in the Code Charts by adding annotations in NamesList.txt. For instance, “currently” is suggested to be employed as far as the added abbreviations are in current use in the glyphs.

Those three spaces whose current glyph mnemonics are not fit for standardization (as they start with a digit and contain a slash) have conformant abbreviations added. Even when these are used in the glyphs too, as suggested in *Proposal to synchronize seven glyphs in the Code Charts,* the legacy abbreviations will still need to be mentioned for reference, although not in the first place.

4 IDEOGRAPHIC SPACE U+3000

Change from:

```
3000     IDEOGRAPHIC SPACE
         x (space - 0020)
         # <wide> 0020
```

Change to:

```
3000     IDEOGRAPHIC SPACE
         * currently abbreviated IDSP
         x (space - 0020)
         # <wide> 0020
```

Rationale:

This is part of the space abbreviations and is related to sections 2 and 3.

5 CHARACTER TABULATION..LINE FEED U+0009..U+000A

Change from:

```
0009     <control>
         = CHARACTER TABULATION
         = horizontal tabulation (HT), tab
         000A  <control>
         = LINE FEED (LF)
         = new line (NL), end of line (EOL)
```

Change to:

```
0009     <control>
         = CHARACTER TABULATION
         = horizontal tabulation (HT)¶ [LINE BREAK]
         = tab (TAB)
         000A  <control>
         = LINE FEED (LF)
         = new line (NL)¶ [LINE BREAK]
         = end of line (EOL)
```

Rationale:

The tab character needs an urgent fix since nothing makes clear that “TAB” is a standard abbreviation. It isn’t even mentioned as such, only because the lowercase informative alias has same spelling. Hence I’d suggest putting “tab” on a new line and appending the normative alias “TAB” in usual parentheses. I think it becomes sufficiently clear from the context that uppercasing the abbreviation isn’t meant to stop users from pronouncing it as an acronym.

Collapsing two fully-fledged informative aliases with their standard abbreviations on a single line twice seems to me a disturbance, unless the Code Charts layout actually requires sparing two lines below this chart, which is not the case since the printed list spans over 3¼ pages. The actual wording dates back to Unicode 3.2.0 (2002), and in the next Code Charts Unicode 4.0.0 the block 1 list already took 3⅓ pages; there was no issue.

The fact is that these two characters are standing out amongst all the surrounding C0 controls by the number of aliases crowding their entries, but that’s in the nature of the matter and deserves full spotlight.

6 HYPHEN..HORIZONTAL BAR U+2010..U+2015

Change from:

```
@                Dashes
2010     HYPHEN
         x (hyphen-minus - 002D)
         x (soft hyphen - 00AD)
2011     NON-BREAKING HYPHEN
         x (hyphen-minus - 002D)
         x (soft hyphen - 00AD)
         # <noBreak> 2010
2012     FIGURE DASH
2013     EN DASH
2014     EM DASH
         * may be used in pairs to offset parenthetical text
         x (two-em dash - 2E3A)
         x (katakana-hiragana prolonged sound mark - 30FC)
2015     HORIZONTAL BAR
         = quotation dash
         * long dash introducing quoted text
```

Change to:

```
@                Dashes and hyphens
2010     HYPHEN
         * same glyph as hyphen-minus in most fonts
         * one quarter of an em dash
         * abbreviated HY
         x (hyphen-minus - 002D)
         x (soft hyphen - 00AD)
2011     NON-BREAKING HYPHEN
         * same glyph as hyphen
         * abbreviated NBHY
         x (hyphen-minus - 002D)
         x (soft hyphen - 00AD)
         # <noBreak> 2010
2012     FIGURE DASH
         * dash equal to tabular width of a font
2013     EN DASH
         * half an em dash
         * may be used in pairs to offset parenthetical text
2014     EM DASH
         * may be used in pairs to offset parenthetical text
         x (two-em dash - 2E3A)
         x (katakana-hiragana prolonged sound mark - 30FC)
2015     HORIZONTAL BAR
         = quotation dash
         * long dash introducing quoted text
         * three quarters of an em dash in some fonts
```

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:

“Dashes” does not accurately sum up this range starting with two hyphens. “Dashes and hyphens” is a very common name for a group of punctuation marks so as to pay tribute to the special status of the hyphen, although *typographically* it is considered a quarter-em *dash* (French: “*tiret quart de cadratin*”).

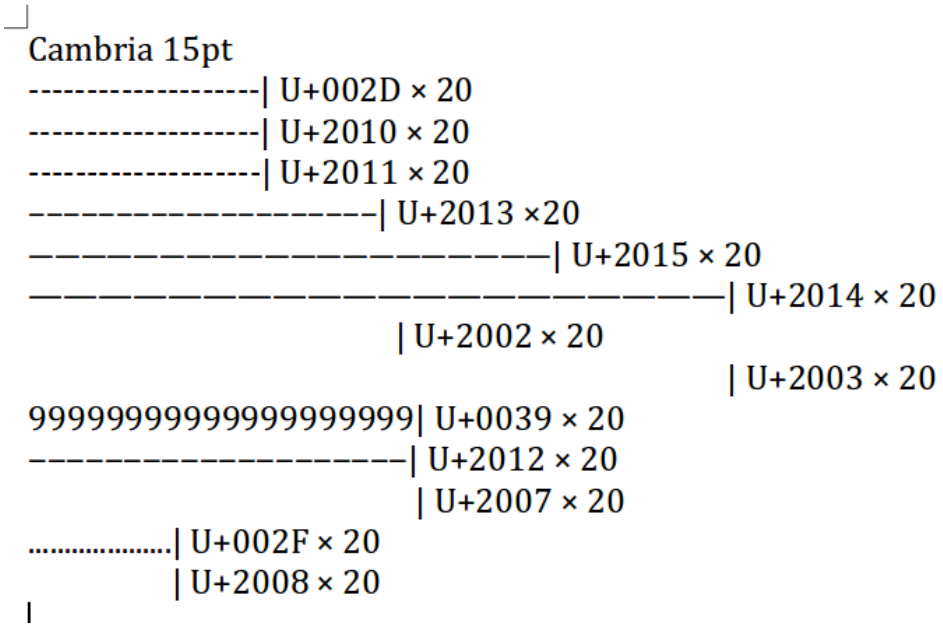
The two lines proposed for deletion are duplicates of cross-references three lines above each one, where they make actual sense given that the SOFT HYPHEN is breaking, not non-breaking. I’d suggest rather providing additional information instead, the rationales of which are grouped hereafter:

1. **Lengths:** As already mentioned, the HYPHEN is currently considered one quarter the length of an EM DASH. The scale goes on with EN DASH and then—perhaps surprisingly—with HORIZONTAL BAR as shown in the test below. Although U+2015 is often given the length of U+2014, some fonts make it an intermediate dash between U+2013 and U+2014. Some other fonts make a difference between U+2014 and U+2015 by giving the latter side bearings that the former does not have in many fonts, no more than U+2013. While I don’t see the use of side bearings around these dashes, I do think that a *THEE-QUARTER

EM DASH is fairly useful, especially when EM DASH is too long, and EN DASH too short for a given design. Adding length-related annotations is meant to incentivize thoughts about how to make the most of these characters instead of fussing about an alleged encoding design inconsistency.

2. **Tabular:** One aspect of length is tabular width as given to U+2007 FIGURE SPACE and consistently to U+2012 FIGURE DASH. While it's surely up to the Core Specification to show how to use the three tabular characters for typesetting tables without tab stops, the Code Charts should at least give the FIGURE DASH the same support as it gives the FIGURE SPACE, where an annotation reads: "space equal to tabular width of a font", with another one explaining what that means. I've copy-pasted and adapted that annotation to FIGURE DASH. The same holds true for EN DASH, with one of the suggested annotations copied from EN SPACE.
3. **Glyphs:** The issue with some fonts having dissimilar glyphs for U+002D HYPHEN-MINUS on one hand, and for U+2010 HYPHEN on the other hand, is a UX disturbance and is considered a design mistake, at least as far as U+2212 MINUS SIGN is to be used under the same conditions than U+00D7 MULTIPLICATION SIGN and U+00F7 DIVISION SIGN. Therefore HYPHEN-MINUS is expected to look like HYPHEN in proportional fonts. Adding some hints right in the Code Charts widely used by font designers appears in this light a good idea. I've tried to suggest a couple of straightforward annotations. Some extra information would certainly be useful but is probably not indispensable.
4. **Usage:** Some locales are using EN DASH for the same purpose than others are using EM DASH as of bracketing parentheticals. Hence the same annotation as for EM DASH goes surely also to EN DASH.
5. **Abbreviations:** *Proposal to extend support for abbreviations* suggests standardizing "NBHY" used in UAX #14 from Unicode 3.0 to Unicode 10.0.0 (over 18 years long). Additionally, adding "HY" is recommended for symmetry and with respect to the issues surrounding U+2010 HYPHEN, mostly confusable with the keyboard hyphen (see also point 3 above). The Unicode Standard should in my opinion be supportive of facilitating string description by providing a few handy descriptors.

Test page for length ratios, provided to support points 1, 2 and 3 of the rationale above



7 LINE SEPARATOR..NARROW NO-BREAK SPACE U+2028..U+202F

Change from:

```
@               Format characters
2028    LINE SEPARATOR
        * may be used to represent this semantic unambiguously
2029    PARAGRAPH SEPARATOR
        * may be used to represent this semantic unambiguously
202A    LEFT-TO-RIGHT EMBEDDING
        * commonly abbreviated LRE
202B    RIGHT-TO-LEFT EMBEDDING
        * commonly abbreviated RLE
202C    POP DIRECTIONAL FORMATTING
        * commonly abbreviated PDF
202D    LEFT-TO-RIGHT OVERRIDE
        * commonly abbreviated LRO
202E    RIGHT-TO-LEFT OVERRIDE
        * commonly abbreviated RLO
202F    NARROW NO-BREAK SPACE
        * commonly abbreviated NNBSP
        * a narrow form of a no-break space, typically the width of a thin
        space or a mid space
        x (no-break space - 00A0)
        x (four-per-em space - 2005)
        x (thin space - 2009)
        # <noBreak> 0020
@               General punctuation
```

Change to:

```
@               ~~Format characters~~ Separators
2028    LINE SEPARATOR
        * may be used to represent this semantic unambiguously
        * commonly abbreviated LS
2029    PARAGRAPH SEPARATOR
        * may be used to represent this semantic unambiguously
        * commonly abbreviated PS
@               Format characters
202A    LEFT-TO-RIGHT EMBEDDING
        * commonly abbreviated LRE
202B    RIGHT-TO-LEFT EMBEDDING
        * commonly abbreviated RLE
202C    POP DIRECTIONAL FORMATTING
        * commonly abbreviated PDF
202D    LEFT-TO-RIGHT OVERRIDE
        * commonly abbreviated LRO
202E    RIGHT-TO-LEFT OVERRIDE
        * commonly abbreviated RLO
@               Space
202F    NARROW NO-BREAK SPACE
        * commonly abbreviated NNBSP
```

```
            * ~~a narrow form of a no-break space,~~ typically the width of ~~a~~ thin
            space ~~or a mid space~~
            * in Mongolian also a format character
            x (no-break space - 00A0)
            x (four-per-em space - 2005)
            x (thin space - 2009)
            # <noBreak> 0020
@                   General punctuation
```

Rationale:

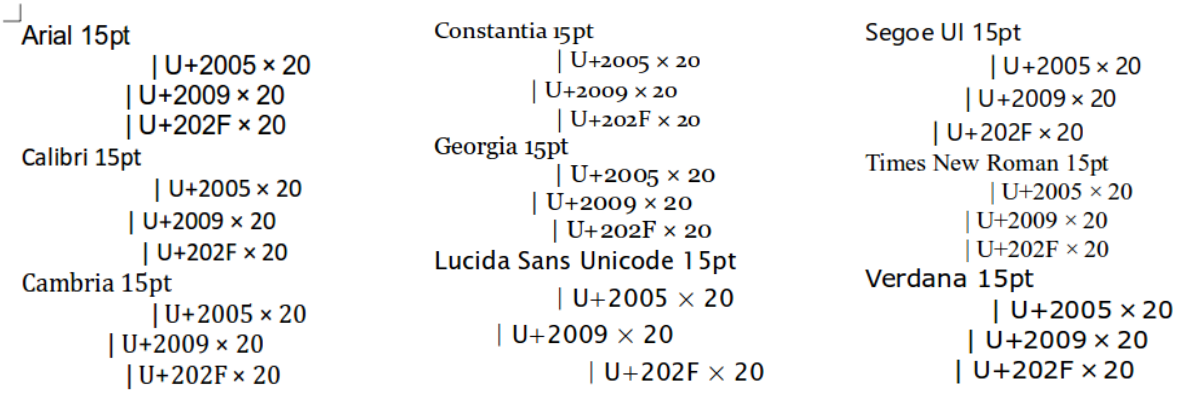
Since LINE SEPARATOR and PARAGRAPH SEPARATOR are of General Category = Line Separator (gc=Zl) and Paragraph Separator (gc=Zp) respectively, not Format (gc=Cf), they need another subheading. The most straightforward is the name of the major class Z.

The abbreviations of LINE SEPARATOR and PARAGRAPH SEPARATOR are added below the existing annotations that are referring to the character names ("this semantic"). The rest of the range is cited in extenso so as to show the ubiquity of abbreviation annotations in this range. The Core Specification makes use of these abbreviations, and *Proposal to extend support for abbreviations* makes a case for standardizing them among other unsupported yet expected abbreviations.

Since the subheading of (the rest of) this range is "Format characters", NARROW NO-BREAK SPACE needs an extra subheading "Space" to account for its General Category = Space Separator (gc=Zs).

The phrase "is a narrow version of no-break space" has been deleted from UAX #14 for Unicode 12.0.0; see revision 42 (also L2/19-030) for the deletion markup. This proposal suggests synching the Code Charts, while *Proposal to synchronize the Core Specification* does the same for TUS. The rationale is that since NNBSP is not justifying, the space that it is a narrow version of is necessarily a fixed-width space. The only non-breaking one is FIGURE SPACE. It is easier however to take a breaking space of the same width, typically THIN SPACE, and to say that NNBSP is "a non-breaking form of a thin space".

The Code Charts are implying that NNBSP can be the width of a mid space (FOUR-PER-EM SPACE) from Unicode 6.0.0 (2010) on, after a decade of undirected implementation (since NNBSP was encoded for Unicode 3.0.0, 1999). Normally it is expected to be the width of THIN SPACE, given that this is the space used in DTP to space off certain punctuation marks, and NNBSP is supposed to take its place for interoperability, following the fairly explicit design goal of Unicode for NNBSP. Consistently, the most used proportional fonts Arial and Times New Roman give NNBSP the width of THIN SPACE. However, since there is a stunning diversity of ratios across fonts between FOUR-PER-EM SPACE, THIN SPACE and NARROW NO-BREAK SPACE for no apparent reason, the only statement that is both safe *and useful* is that NNBSP is "typically the width of thin space" (lowercase for NamesList syntax sake), where "typically" means "in most-used fonts," namely Times New Roman for serif, and Arial for sans-serif.



Hopefully a streamlined annotation about NNBSP's nature and width can incentivize usable font design.

To the suggestions I've added an annotation about the use of NNBSP in Mongolian as a format character. This information is in my opinion better located in an annotation to the character, than in a subheading the character is added under despite that subheading catches scarcely more than half of the characters in the range it is placed over (4 out of 7).

8 Provisional

The simultaneously submitted *Proposal to extend support for abbreviations* contains additional suggestions resulting from a comprehensive data review of all invisible characters. Some characters have been spotted in Arabic, Brahmi, Kaithi and Tifinagh that could eventually be given a useful abbreviation. By contrast, unlike U+061C ARABIC LETTER MARK, short ALM, none of the nine following characters has a known abbreviation yet. Hence these are provisional while feedback from communities is needed:

```
0600  ARABIC NUMBER SIGN
* abbreviated ANS
0602  ARABIC FOOTNOTE MARKER
* abbreviated AFM
0605  ARABIC NUMBER MARK ABOVE
* abbreviated ANMA
06DD  ARABIC END OF AYAH
* abbreviated AEOA
08E2  ARABIC DISPUTED END OF AYAH
* abbreviated ADEOA
2D7F  TIFINAGH CONSONANT JOINER
* abbreviated TCJ
1107F BRAHMI NUMBER JOINER
* abbreviated BNJ
110BD KAITHI NUMBER SIGN
* abbreviated KNS
110CD KAITHI NUMBER SIGN ABOVE
* abbreviated KNSA
```

9 Moot

The abbreviations for these four mathematical format characters are moot, as the suggested initialisms are not renownedly backed by mathematicians.

Some names were even hard to abbreviate due to interference with initialisms IP and IT:

```
2061  FUNCTION APPLICATION

* abbreviated FA

2062  INVISIBLE TIMES

* abbreviated IMS for "invisible multiplication sign"

2063  INVISIBLE SEPARATOR

* abbreviated IS

2064  INVISIBLE PLUS

* abbreviated IPS for "invisible plus sign"
```

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.