L2/20-007

Proposal to extend support for abbreviations

For consideration by Unicode Technical Committee

2020-01-13 (revised; first submitted 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 encompasses part of the response to Action item 161-A1 as it doesn't fully integrate into either *Proposal suggesting formal edits to UAX #14* or *Proposal to make material changes to UAX #14*. It needs indeed to care about the underlying data in the first place.

Please see also *Proposal to adapt TUS to extended support for abbreviations, Proposal to make focused changes to the Code Charts text, Proposal to synchronize seven glyphs in the Code Charts,* and *Proposal to synchronize two glyphs in the Core Specification.*

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

For more information about the ins and outs, please see the *Rationale* and *Background* sections below. For detailed data reviews, please see *annexes A* and *B,* and for an overview, to *annex C.*

1 Suggested Abbreviations

1.1 Top priority

cp	Name	Suggested abbreviation
070F	SYRIAC ABBREVIATION MARK	SAM
1680	OGHAM SPACE MARK	OGSP
2000	EN QUAD	NQSP
2001	EM QUAD	MQSP
2002	EN SPACE	ENSP
2003	EM SPACE	EMSP
2004	THREE-PER-EM SPACE	THPMSP
2005	FOUR-PER-EM SPACE	FPMSP
2006	SIX-PER-EM SPACE	SPMSP
2007	FIGURE SPACE	FSP
2008	PUNCTUATION SPACE	PSP
2009	THIN SPACE	THSP
200A	HAIR SPACE	HSP
2010	HYPHEN	HY
2011	NON-BREAKING HYPHEN	NBHY
2028	LINE SEPARATOR	LS
2029	PARAGRAPH SEPARATOR	PS
3000	IDEOGRAPHIC SPACE	IDSP

Please see the rationales in sections *2 Rationale* and *3 Background*, and in annex *A Grouped data review*, where they have been relegated in order to streamline the proposal by raising these tables.

In particular, for SAM please see subsection *2.2.2 SYRIAC ABBREVIATION MARK U+070F.*

For spaces, in particular OGHAM SPACE MARK, please see subsection *2.2.3 Spaces* and annex *A.6 Space Characters,* and please see also *Proposal to synchronize seven glyphs in the Code Charts.*

For HYPHEN and NON-BREAKING HYPHEN, please see subsection *2.3 Inclusion of hyphens.*

LINE SEPARATOR and PARAGRAPH SEPARATOR are mentioned in *2.2.4 Line and paragraph separators.*

See also section *4 Data for UCD,* suggesting some edits to a dozen lines in the file header along with providing lines for the file body as usual.

1.2 Provisional

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

Please see the rationales in the following sections of annex A where they have been relegated in order to streamline the proposal by raising these tables:

- *A.2 Prepended Concatenation Marks* (Arabic, Kaithi);
- *A.3 Number Joiner* (Brahmi);
- *A.5 Other Combining Joiners* (Tifinagh).

See also section *4 Data for UCD.*

1.3 Moot

cp	Name	Suggested abbreviation
2061	FUNCTION APPLICATION	FA
2062	INVISIBLE TIMES	IMS
2063	INVISIBLE SEPARATOR	IS
2064	INVISIBLE PLUS	IPS

Please see the rationale in annex *A.11 Mathematical format characters*, where it has been relegated in order to streamline the proposal by raising these tables. — See also section *4 Data for UCD.*

2 Rationale

This section is about why L2/19-317 ***Proposal to update some statements about space characters in* Unicode Standard Annex #14: Unicode Line Breaking Algorithm** suggested adding more abbreviations, and whether doing so would be worth making material or editorial changes to several parts of the Unicode Standard.

2.1 Main reasons

The main reasons why in my opinion the Unicode Standard needs to support more abbreviations are:

1. Abbreviations like CGJ or SAM are skipped in UAX #14 when introducing example characters, although they are widely used also in the Core Specification *The Unicode Standard* (TUS), and despite for some other characters, UAX #14 does provide abbreviations. SAM is even missing from NameAliases.txt. The reason why seems non-obvious. Adding both is a minor edit in UAX #14, but for SAM it involves a material change to the Unicode Character Database (UCD).
2. Abbreviations like ANS for ARABIC NUMBER SIGN are missing from the Standard, even while others like ALM for ARABIC LETTER MARK are present in NameAliases.txt. Completing the set is considered a matter of equity and consistency between Arabic and Syriac, and between bidi controls and prepended concatenation marks, though in practice it is limited to fully English names. The preference about not using abbreviations for invisible stackers is also followed.
3. Characters like THREE PER EM SPACE seem to lack any conformant abbreviation. Adding abbreviations like THPMSP for 3/MSP is suggested proactively for users' convenience, and for consistency and completeness. That contradicts the scope of existing and widely or commonly used abbreviations stated in TUS and NameAliases.txt, but I think it is worth the change.

Although this proposal pays special attention to UAX #14, it widens its focus in an attempt to synchronize all related parts of the Unicode Standard, so that UAX #14 won't stand out after copy-editing.

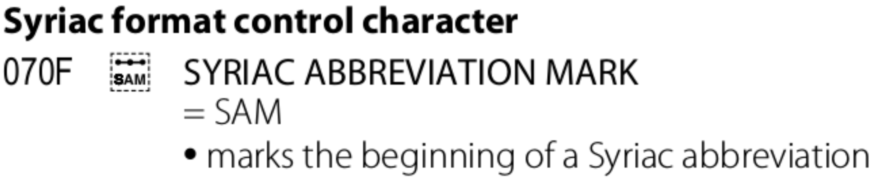
2.2 Evidence for a need to change

2.2.1 Frequency of use

Abbreviations prove useful in the Standard and in other ICT documentation when describing strings, the more as characters like THIN SPACE is massively relied upon in publishing (both wysiwyg and TeX) thanks to their tailorable line breaking property values. Other examples of frequently used spaces include EM SPACE and EN SPACE, but also FIGURE SPACE even if not as a group separator. It is used along with PUNCTUATION SPACE in positions such as next to the start of line. Yet none of these spaces is present in NameAliases.txt.

2.2.2 SYRIAC ABBREVIATION MARK U+070F

The SYRIAC ABBREVIATION MARK initialism "SAM" is found both in TUS, inside the dashed box of its reference glyph in the Code Charts, and even as an informative alias there: "= SAM".



Syriac format control character

070F SAM SYRIAC ABBREVIATION MARK
= SAM
• marks the beginning of a Syriac abbreviation

2.2.3 Spaces

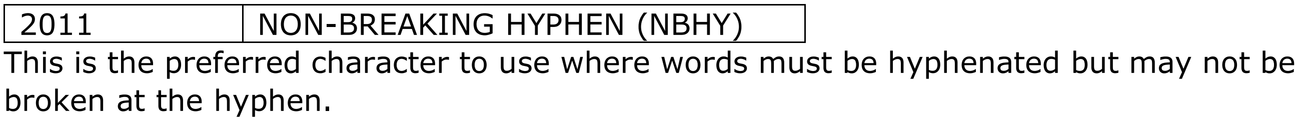
In UAX #14, among the blank or invisible characters listed as examples without having their parenthesized abbreviation appended to their name, two (CGJ, MMSP) are already in NameAliases.txt. But nine (NQSP, MQSP, ENSP, EMSP, FSP, PSP, THSP, HSP, IDSP) are only in their Code Charts dashed boxes. Three others are there while unfit for standardization (3/MSP, 4/MSP, 6/MSP) and need to be redesigned, and OGSP for OGHAM SPACE MARK, to be revived (at least in my opinion).

2.2.4 Line and paragraph separators

Among those abbreviations not yet in NameAliases.txt, “LS” and “PS” are best found in the Core Specification rather than in their reference glyphs, where they read “LSEP” and “PSEP”; please see also *Proposal to synchronize seven glyphs in the Code Charts*.

2.3 Inclusion of hyphens

“NBHY” for NON-BREAKING HYPHEN is a handy abbreviation helping streamline string documentation, as that character is widely used. Asmus Freytag introduced the abbreviation into UAX #14 alongside the character in Revision 6.0 for Unicode 3.0.0. Like what is done for other abbreviations, “NBHY” was parenthesized and appended to the character name:



2011	NON-BREAKING HYPHEN (NBHY)

This is the preferred character to use where words must be hyphenated but may not be broken at the hyphen.

From Version Unicode 5.0.1 on (see draft Revision 20), “NBHY” was used in two instances of the string descriptor “<SHY, NBHY>”.

It wasn’t until PRI #376 about updating UAX #14 for Unicode 11.0.0, that “NBHY” came under scrutiny; yet Charlotte Buff suggested no more than that “some difference in formatting could be introduced as to not imply that NBHY is a stable identifier.” Nevertheless, Andy Heninger was prompted in 155-A26 to remove the abbreviation “NBHY” from the table and change it to the full name in the string descriptor.

The other way of fixing the abbreviation of NON-BREAKING HYPHEN was by adding NBHY to NameAliases.txt. I’d suggest rather doing that, and accordingly I’ve extended the scope of this proposal. I’m aware, though, that the string “NBHY” does occur on the internet in unrelated contexts.

For balance and completeness, the abbreviation “HY” for HYPHEN is suggested alongside, not only because the two code points are contiguous, and because <HY, NBHY> form a pair much like <SP, NBSP>, but also because U+2010 HYPHEN is confusable with HYPHEN-MINUS in Basic Latin and should never have been encoded, as its main effect was to motivate very few font designers to give U+002D a glyph that makes the hyphen on the keyboard hardly usable in practice when typing plain English in such a font, except in technical documentation where there may be no issue.

3 Background

3.1 Initial suggestions

Proposal L2/19-317 started suggesting that the most possible abbreviations be provided for blank or invisible characters when listed as examples of line breaking class members, but it did so only in the four tables quoted in the parts proposed for changes. The first of these tables has already all parenthesized abbreviations, the others don't. Abbreviations were added on the fly as they were found. That resulted in relying also on the sample glyphs in the Code Charts without paying attention to the limits of the Unicode namespace. It resulted also in OGHAM SPACE MARK standing out by lack of an abbreviation.

In that proposal, abbreviations are mistakenly called "acronyms", while "SP", used in isolation and as part of other abbreviations, is not an acronym (but an initialism). The only rationale provided was consistency with the first quoted table, while not seeking consistency with NameAliases.txt. Example:

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

Neither is "FSP" a legal abbreviation per the Unicode Standard's [UCD]/NameAliases.txt (latest version 12.1.0), nor is it used anywhere in the Core Specification (*The Unicode Standard*, version 12.0.0), nor in the Code Charts' annotations (per version 12.1.0 of NamesList.txt).

3.2 Limitations

The concept of using abbreviations as mnemonic identifiers, successfully implemented for bidirectional layout controls and for Mongolian, was initially applied also to Brahmic scripts, when the Kharoshthi *virama* was labeled "KV" in version 4.1.0. But this lasted only until the next version. Where the use of acronyms would inflate the number of script-specific items representing a functionally similar invisible *virama*, Unicode switched to smart graphics instead, like the combining plus sign below a dotted circle in a dashed box for *viramas.* This design already proved successful for U+17D2 KHMER SIGN COENG.

The same concept was applied to Tifinagh when a consonant joiner was added in version 6.0.0 and abbreviated "TFNCJ". But calling U+2D7F TIFINAGH CONSONANT JOINER "TFNCJ" was as ephemeral a choice as calling U+10A3F KHAROSHTHI VIRAMA "KV". I'm guessing that "TFNCJ" was considered overlong, as "TFN" is nearly the entire script code (Tfng), in contrast to Kharoshthi and especially to Mongolian, where initialisms are considered sustainable ("MVS", not "MONVS" after the code Mong). Hence, while this proposal refrains from suggesting abbreviations for invisible viramas, it tries "TCJ".

I think there is a balance between ensuring that every blank or special-use character has an abbreviation in the Standard, and meeting locale preferences and existing practice favoring non-Latin labels or non-alphabetic visuals. As a rule of thumb, in my opinion, when it's up to support handy abbreviations for general-use spaces and format characters, no other line should be drawn than deprecation.

4 Data for UCD

4.1 Body data

This section is providing the final data for addition to NameAliases.txt.

Proposals currently may contain such data to streamline processes in the case of acceptance. Some proposals even come in plain text for better interoperability. While I've thought at submitting also plain text files. I'm actually refraining from doing so, in order to not overly breaking up already numerous simultaneously submitted proposals in this rush.

- Top priority:

```
070F;SAM;abbreviation

1680;OGSP;abbreviation

2000;NQSP;abbreviation

2001;MQSP;abbreviation

2002;ENSP;abbreviation

2003;EMSP;abbreviation

2004;THPMSP;abbreviation

2005;FPMSP;abbreviation

2006;SPMSP;abbreviation

2007;FSP;abbreviation

2008;PSP;abbreviation

2009;THSP;abbreviation

200A;HSP;abbreviation

2010;HY;abbreviation

2011;NBHY;abbreviation

2028;LS;abbreviation

2029;PS;abbreviation

3000;IDSP;abbreviation
```

- Provisional:

```
0600;ANS;abbreviation

0602;AFM;abbreviation

0605;ANMA;abbreviation

06DD;AEOA;abbreviation

08E2;ADEOA;abbreviation

2D7F;TCJ;abbreviation

1107F;BNJ;abbreviation

110BD;KNS;abbreviation

110CD;KNSA;abbreviation
```

- Moot:

```
2061;FA;abbreviation

2062;IMS;abbreviation

2063;IS;abbreviation

2064;IPS;abbreviation
```

4.2 File header

Since some abbreviations are likely to not yet be in use, the file header needs to be adapted to account for the change in scope. After the edit, abbreviations are not qualified by frequency any more.

Nor would the header keep enumerating involved character classes, as the actual enumeration is already incomplete. Notably in the actual set description, CGJ (gc=Mn) is not caught.

Fixing this in the Core Specification is among the purposes of *Proposal to adapt TUS to extended support for abbreviations.*

Change from:

```
# 5. abbreviation

#       Commonly occurring abbreviations (or acronyms) for control codes,

#       format characters, spaces, and variation selectors

#
```

```
# The formal name aliases are part of the Unicode character namespace, which

# includes the character names and the names of named character sequences.

# The inclusion of ISO 6429 names and other commonly occurring names and

# abbreviations for control codes and format characters as formal name aliases

# is to help avoid name collisions between Unicode character names and the

# labels which commonly appear in text and/or in implementations such as regex, for

# control codes (which for historical reasons have no Unicode character name)
# or for format characters.
```

Change to:

```
# 5. abbreviation
#      ~~Commonly occurring a~~Abbreviations (may be acronyms or initialisms) for various
#      blank or special-use characters such as spaces or variation selectors
#
# The formal name aliases are part of the Unicode character namespace, which
# includes the character names and the names of named character sequences.
# The inclusion of ISO 6429 names and other ~~commonly occurring~~ useful names and
# abbreviations for control codes and format characters as formal name aliases
# is to help avoid name collisions between Unicode character names and the
# labels which may ~~commonly~~ appear in text and/or in implementations such as regex, for
# control codes (which for historical reasons have no Unicode character name),
# ~~or~~ for format characters or for other invisible characters.
```

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.

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.

Annex A: Grouped data review

This section is grouping items by type to propose abbreviations or not. These are then used in the *Proposed Actions* section.

A.1 Bidi Controls

This table contains all characters whose Bidi_Control property value is Yes. These are cited for reference and because the last four do not have Code Charts annotations mentioning their abbreviation, as do the other ones. This is properly NamesList.txt feedback, but for completeness it is included in this proposal.

cp	Name	NameAliases.txt	Core Specification	Charts
061C	ARABIC LETTER MARK	ALM	ALM	ALM
200E	LEFT-TO-RIGHT MARK	LRM	LRM	LRM
200F	RIGHT-TO-LEFT MARK	RLM	RLM	RLM
202A	LEFT-TO-RIGHT EMBEDDING	LRE	LRE	LRE
202B	RIGHT-TO-LEFT EMBEDDING	RLE	RLE	RLE
202C	POP DIRECTIONAL FORMATTING	PDF	PDF	PDF
202D	LEFT-TO-RIGHT OVERRIDE	LRO	LRO	LRO
202E	RIGHT-TO-LEFT OVERRIDE	RLO	RLO	RLO
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]

A.2 Prepended Concatenation Marks

This table lists all characters with Prepended_Concatenation_Mark=Yes. None, not even SAM, is in NameAliases.txt. Abbreviations are suggested for some of these, more precisely for those not having their local name in their Unicode name, following the example of SYRIAC ABBREVIATION MARK.

cp	Name	Core	Charts	UAX14 lists	14 else	sugg
0600	ARABIC NUMBER SIGN	<none>	[<Arabic>]	<collapsed>	<none>	ANS
0601	ARABIC SIGN SANAH	<none>	[<Arabic>]	<collapsed>	<none>	<none>
0602	ARABIC FOOTNOTE MARKER	<none>	[<Arabic>]	<collapsed>	<none>	AFM
0603	ARABIC SIGN SAFHA	<none>	[<Arabic>]	<collapsed>	<none>	<none>
0604	ARABIC SIGN SAMVAT	<none>	[<Arabic>]	<collapsed>	<none>	<none>
0605	ARABIC NUMBER MARK ABOVE	<none>	[<Arabic>]	#N/A	#N/A	ANMA
06DD	ARABIC END OF AYAH	<none>	[<graphic>]	<none>	<none>	AEOA
070F	SYRIAC ABBREVIATION MARK	SAM	SAM	<none>	<none>	SAM
08E2	ARABIC DISPUTED END OF AYAH	<none>	[<Arabic>]	#N/A	#N/A	ADEOA
110BD	KAITHI NUMBER SIGN	<none>	[<Kaithi>]	<none>	#N/A	KNS
110CD	KAITHI NUMBER SIGN ABOVE	<none>	[<Kaithi>]	#N/A	#N/A	KNSA

A.3 Number Joiner

This is the full list of Indic_Syllabic_Category=Number_Joiner. The nonspacing combining mark BRAHMI NUMBER JOINER has already a Latin capital letter mnemonic in its dashed-box glyph in the Code Charts. Hence standardizing that initialism seems appropriate within the scope of this proposal.

cp	Name	NameAliases.txt	Core	Charts	sugg
1107F	BRAHMI NUMBER JOINER	#N/A	<none>	[BNJ]	BNJ

A.4 Invisible Stackers

This is the full list of Indic_Syllabic_Category=Invisible_Stacker. None of these has an abbreviation in NameAliases.txt, and the established usage is to call them generically (without mentioning the script name) and to represent them in the Code Charts with a generic glyph resulting from enclosing the reference glyph of COMBINING PLUS SIGN BELOW — with its dotted circle — in a dashed box. Hence no abbreviation is suggested for any of these, but two easy-to-fix glyph issues are to be reported, one (already mentioned) in the Code Charts with SUNDANESE SIGN VIRAMA, and one in the Core Specification with KHAROSHTHI VIRAMA.

cp	Name	Core Specification	Code Charts
1039	MYANMAR SIGN VIRAMA	[◌̟], *virama*	[◌̟]
17D2	KHMER SIGN COENG	[◌̟]	[◌̟]
1A60	TAI THAM SIGN SAKOT	<none>	[◌̟]
1BAB	SUNDANESE SIGN VIRAMA	<none>	[̟]
AAF6	MEETEI MAYEK VIRAMA	<none>	[◌̟]
10A3F	KHAROSHTHI VIRAMA	[KV]	[◌̟]
11133	CHAKMA VIRAMA	*virama*	[◌̟]
11A47	ZANABAZAR SQUARE SUBJOINER	[◌̟], subjoiner	[◌̟]
11A99	SOYOMBO SUBJOINER	subjoiner	[◌̟]
11D45	MASARAM GONDI VIRAMA	[◌̟], *virama*	[◌̟]
11D97	GUNJALA GONDI VIRAMA	[◌̟], *virama*	[◌̟]

A.5 Other Combining Joiners

The COMBINING GRAPHEME JOINER and the TIFINAGH CONSONANT JOINER have been included as invisible characters, of General_Category=Nonspacing_Mark. Both have an abbreviation, but CGJ is in wider use than TFNCJ. The reference glyph of TIFINAGH CONSONANT JOINER has a similar issue like KHAROSHTHI VIRAMA: Its Latin capital letter mnemonic has been replaced in the Code Charts but persists in the Core Specification. I'd suggest standardizing a Latin capital initialism shorter than the legacy abbreviation using almost (but not totally) the full script code Tfng.

cp	Name	NaAl	Core	Charts	14 lists	14 else	sugg
034F	COMBINING GRAPHEME JOINER	CGJ	CGJ	CGJ	<none>	CGJ	
2D7F	TIFINAGH CONSONANT JOINER	#N/A	[TFNCJ]	[<◌......>]	#N/A	#N/A	TCJ

A.6 Space Characters

This is a subset of White_Space=Yes, including only characters with General_Category=Space_Separator in order to shorten the list by grouping the other characters into the next subsection.

Most of these (13 out of 17) are lacking an official abbreviation, as they are missing from NameAliases.txt. For these, a unique abbreviation is suggested, most (12) of which are already found (9) in dashed-box glyphs or are derived (3) from the latter for compliance with Unicode character name constraints. Using non-standard Latin abbreviations in the Unicode Standard is problematic. For an abbreviation to be standard, it needs to stand in NameAliases.txt and thus to comply to the Unicode character namespace constraints. These are prohibiting slashes and leading digits as found in the glyphs of THREE-, FOUR- and SIX-PER-EM SPACE. Converting these to conformant abbreviations is easy when remembering that a word-leading "th" enters the abbreviation as a "TH".

For OGHAM SPACE MARK, "OGSP" would similarly be found in its dashed-box glyph, additionally to the stemline, according to a wide consensus (UTC #113, Action Item A15) to change this glyph to the fourth one depicted in L2/08-142. If L2/08-318, section 9.14, fell short of completing that change, making a new attempt now is appropriate because omitting the four Latin capitals in the reference glyph of OGHAM SPACE is contradicting the claim—even made in the Code Charts—about the non-mandatory status of the stemline. Since Ogham can be written without the stemline, and stemless Ogham fonts do exist on the marketplace, a bare stemline can by no means be a distinctive sign of the OGHAM SPACE MARK reference glyph. Think of the stemline as missing, and you will end up today with an empty dashed box as a reference glyph of OGHAM SPACE MARK. Initially there wasn't even a dashed box. When that flaw was on the table, everything was ready to get "OG" and "SP" into the glyph alongside. Now that all spaces are hopefully going to get a standard abbreviation, OGHAM SPACE MARK included (that isn't actually a "mark," so OGSP is fine), there is a good occasion for making Ogham script cease standing out, by completing its space glyph to make it a fully-fledged Code Charts space character reference glyph.

cp	Name	NaAl	Core	Charts	UAX14 lists	UAX14 else	sugg
0020	SPACE	SP	SPACE	[SP]	(SP)	SP	
00A0	NO-BREAK SPACE	NBSP	NBSP	NBSP	(NBSP)	NBSP	
1680	OGHAM SPACE MARK	#N/A	<none>	[—] /<blank>	<none>	<full>	OGSP
2000	EN QUAD	#N/A	<none>	[NQSP]	<none>	<none>	NQSP
2001	EM QUAD	#N/A	<none>	[MQSP]	<none>	<none>	MQSP
2002	EN SPACE	#N/A	<none>	[ENSP]	<none>	<none>	ENSP
2003	EM SPACE	#N/A	<none>	[EMSP]	<none>	<none>	EMSP
2004	THREE-PER-EM SPACE	#N/A	<none>	[3/MSP]	<none>	<none>	THPMSP
2005	FOUR-PER-EM SPACE	#N/A	<none>	[4/MSP]	<none>	<none>	FPMSP
2006	SIX-PER-EM SPACE	#N/A	<none>	[6/MSP]	<none>	<none>	SPMSP
2007	FIGURE SPACE	#N/A	<none>	[FSP]	<none>	<full>	FSP
2008	PUNCTUATION SPACE	#N/A	<none>	[PSP]	<none>	<none>	PSP
2009	THIN SPACE	#N/A	<none>	[THSP]	<none>	<full>	THSP
200A	HAIR SPACE	#N/A	<none>	[HSP]	<none>	<none>	HSP
202F	NARROW NO-BREAK SPACE	NNBSP	NNBSP	NNBSP	(NNBSP)	<full>	

205F	MEDIUM MATHEMATICAL SPACE	MMSP	<none>	MMSP	<none>	<none>	
3000	IDEOGRAPHIC SPACE	#N/A	<none>	[IDSP]	<none>	<none>	IDSP

A.7 Line and paragraph separators

The abbreviations “LS” and “PS” are used in the Core Specification and should become standard.

cp	gc	Unicode name or ISO 6429 name	NaAl	Core	Charts	UAX14 lists	UAX14 else	sugg
0009	Cc	CHARACTER TABULATION	HT, TAB	HT	HT	TAB	tab	
000A	Cc	LINE FEED	LF	LF	LF	(LF)	LF	
000B	Cc	LINE TABULATION	VT	VT	VT	(VT)	VT	
000C	Cc	FORM FEED	FF	FF	FF	(FF)	FF	
000D	Cc	CARRIAGE RETURN	CR	CR	CR	(CR)	CR	
0085	Cc	NEXT LINE	NEL	NEL	NEL	(NEL)	NEL, NL	
2028	Zl	LINE SEPARATOR	#N/A	LS	[LSEP]	<none>	<full>	LS
2029	Zp	PARAGRAPH SEPARATOR	#N/A	PS	[PSEP]	<none>	<full>	PS

The abbreviations “LS” and “PS” are consistent with these control characters and their Control Picture glyphs in the Code Charts, while the mnemonics in the reference glyphs of LS and PS, actually LSEP and PSEP, need to be synched as suggested in section 2 of *Proposal to synchronize seven glyphs in the Code Charts.*

Templates backing the two-letter abbreviations of LINE SEPARATOR and PARAGRAPH SEPARATOR:

cp	ISO 6429 name	Alias	Abbreviation
001C	INFORMATION SEPARATOR FOUR	FILE SEPARATOR	FS
001D	INFORMATION SEPARATOR THREE	GROUP SEPARATOR	GS
001E	INFORMATION SEPARATOR TWO	RECORD SEPARATOR	RS
001F	INFORMATION SEPARATOR ONE	UNIT SEPARATOR	US

A.8 Line break format characters

Among the rest of format characters, these are designed to generate or inhibit break opportunities, and nothing else. Their abbreviations are well supported in all considered documents. They are cited here for completeness and to broaden the set of well supported characters for demonstration purposes within this proposal.

cp	Name	NaAl	Core	Charts	UAX14 lists	UAX14 else
00AD	SOFT HYPHEN	SHY	SHY	SHY	(SHY)	SHY
200B	ZERO WIDTH SPACE	ZWSP	ZWSP	ZWSP	(ZWSP)	ZWSP
2060	WORD JOINER	WJ	<none>	WJ	(WJ)	WJ, <full>
FEFF	ZERO WIDTH NO-BREAK SPACE	BOM, ZWNBSP	BOM, ZWNBSP	ZWNBSP, BOM	(ZWNBSP)	ZWNBSP

A.9 Shaping format characters

These are designed to control the letter shaping behavior of rendering engines. Their abbreviations are well supported in all considered documents. They are cited here for completeness and to broaden the set of well supported characters for demonstration purposes within this proposal.

cp	Unicode name	NaAl	Core	Charts	UAX14 lists	UAX14 else
180E	MONGOLIAN VOWEL SEPARATOR	MVS	MVS	MVS	(MVS)	<full>
200C	ZERO WIDTH NON-JOINER	ZWNJ	ZWNJ	ZWNJ	#N/A	#N/A
200D	ZERO WIDTH JOINER	ZWJ	ZWJ	ZWJ	(ZWJ)	ZWJ

A.10 Hyphens

The NON-BREAKING HYPHEN is widely used and had an abbreviation in UAX #14 from Unicode 3.0.0 through 10.0.0, but failing to be standardized, it disappeared in Unicode 11.0.0. This proposal makes a case for “NBHY” being worth standardizing.

Alongside, “HY” for HYPHEN is suggested as an inevitable counterpart, as advocated in the section *Inclusion of hyphens* above.

cp	Name	NaAl	Core	Charts	UAX14 lists	UAX14 else	sugg
2010	HYPHEN	#N/A	<none>	[-]	<none>	hyphen	HY
2011	NON-BREAKING HYPHEN	#N/A	<none>	[NB-]	<none>	*non-breaking hyphen*	NBHY

A.11 Mathematical format characters

Since these characters help disambiguate formulae, abbreviations could be handy. Mathematicians alone are able to assess the usefulness of “<FA>” vs “<f()>” and so on, but these characters are considered in scope because they fall within the considered sets and are in UAX #14 (although in a collapsed range). The provisionally suggested abbreviations might be used in UAX #14 when expanding that range.

To design these abbreviations, the names MULTIPLICATION SIGN and PLUS SIGN from *Basic Latin* and *Latin-1 Supplement*, are used instead of TIMES and PLUS to disambiguate the resulting initialisms from “IP” (INVISIBLE PLUS) and “IT” (INVISIBLE TIMES). “IS” in turn is consistent with the “LS” and “PS” set.

cp	Unicode name	NaAl	Core	Charts	UAX14 lists	UAX14	sugg
2061	FUNCTION APPLICATION	#N/A	[f()]	[f()]	<collapsed>	#N/A	FA
2062	INVISIBLE TIMES	#N/A	<none>	[×]	<collapsed>	#N/A	IMS
2063	INVISIBLE SEPARATOR	#N/A	<none>	[,]	<collapsed>	#N/A	IS
2064	INVISIBLE PLUS	#N/A	<none>	[+]	<collapsed>	#N/A	IPS

Annex B: Scalar data review

Prior to narrowing down the scope of this proposal as reflected in *annex A*, this data review has been broadened until encompassing the union of characters fulfilling one of the following conditions:

- White_Space=Yes
- General_Category=Format
- Indic_Syllabic_Category=Invisible_Stacker
- Indic_Syllabic_Category=Number_Joiner
- Name=COMBINING_GRAPHEME_JOINER
- Name=HYPHEN
- Name=NON-BREAKING_HYPHEN
- Name=TIFINAGH_CONSONANT_JOINER
- Name=MUSICAL_SYMBOL_NULL_NOTEHEAD

A full list of the considered characters is attached below in *annex C Data overview.*

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

The following tables are in continuous ascending order of code points. A green background means no issues spotted.

As an example of what in my understanding is good support, here is the list of the first few characters. Except for three of them, I see no issues with these. They have an abbreviation listed in NameAliases.txt. The abbreviation occurs in the Core Specification *The Unicode Standard.* (That is not the case of "SP", consistently with its 1.0 dashed-box glyph.) The Code Charts are providing the abbreviation in an annotation to the character and are using it inside the dashed box. Where the mnemonic there is not backed by an annotation, the string or graphic is bracketed. Lastly in the tables, the abbreviation's appearance in UAX #14 in lists of discussed examples, and how the character is referred to elsewhere in UAX #14.

For U+0009 CHARACTER TABULATION, in UAX #14, "TAB" (one of the two existing abbreviations) stands instead of the ISO 6429 name — despite for U+000B that name is provided followed by the parenthesized abbreviation of its Unicode 1.0 name — and the rest of the document is lowercasing the abbreviation.

As of "CGJ", UAX #14 is skipping it when introducing COMBINING GRAPHEME JOINER, while still using it elsewhere in the document.

The string "<none>" means that no abbreviation is used in the Core Specification, in the Code Charts, or in UAX #14 in the tables listing example characters, respectively. The last column (UAX14 elsewhere) differs in that it shows "<none>" if that character is not mentioned in the rest of the document.

cp	gc	White Space	Name (Unicode or ISO 6429)	NameAliases.txt	Core Spec	Code Charts	UAX14 lists	UAX14 elsewhere
0009	Cc	Yes	CHARACTER TABULATION	HT, TAB	HT	HT	TAB	tab
000A	Cc	Yes	LINE FEED	LF	LF	LF	(LF)	LF
000B	Cc	Yes	LINE TABULATION	VT	VT	VT	(VT)	VT
000C	Cc	Yes	FORM FEED	FF	FF	FF	(FF)	FF

000D	Cc	Yes	CARRIAGE RETURN	CR	CR	CR	(CR)	CR
0020	Zs	Yes	SPACE	SP	SPACE	[SP]	(SP)	SP
0085	Cc	Yes	NEXT LINE	NEL	NEL	NEL	(NEL)	NEL, NL
00A0	Zs	Yes	NO-BREAK SPACE	NBSP	NBSP	NBSP	(NBSP)	NBSP
00AD	Cf	No	SOFT HYPHEN	SHY	SHY	SHY	(SHY)	SHY
034F	Mn	No	COMBINING GRAPHEME JOINER	CGJ	CGJ	CGJ	<none>	CGJ

Next comes a set of Arabic or Syriac marks whose abbreviations are unsupported except for ARABIC LETTER MARK. The SYRIAC ABBREVIATION MARK has its abbreviation SAM given In the Code Chart and used in TUS, but this character did not make it into NameAliases.txt. Hence it should be added there, and its abbreviation appended in UAX #14. I guess that some of the others also need an abbreviation in Latin letters for equity with ALM.

The column header "NaAl" refers to the NameAliases.txt file in the UCD, version 12.1.0. The content of the NaAl column has initially been generated programmatically with a spreadsheet formula checking the subset of entries labeled "abbreviation" in NameAliases.txt.

cp	**gc**	**Name**	**NaAl**	**Core**	**Charts**	**UAX14 lists**	**14 else**
0600	Cf	ARABIC NUMBER SIGN	#N/A	<none>	[<Arabic>]	<collapsed>	<none>
0601	Cf	ARABIC SIGN SANAH	#N/A	<none>	[<Arabic>]	<collapsed>	<none>
0602	Cf	ARABIC FOOTNOTE MARKER	#N/A	<none>	[<Arabic>]	<collapsed>	<none>
0603	Cf	ARABIC SIGN SAFHA	#N/A	<none>	[<Arabic>]	<collapsed>	<none>
0604	Cf	ARABIC SIGN SAMVAT	#N/A	<none>	[<Arabic>]	<collapsed>	<none>
0605	Cf	ARABIC NUMBER MARK ABOVE	#N/A	<none>	[<Arabic>]	#N/A	#N/A
061C	Cf	ARABIC LETTER MARK	ALM	ALM	ALM	#N/A	#N/A
06DD	Cf	ARABIC END OF AYAH	#N/A	<none>	[<graphic>]	<none>	<none>
070F	Cf	SYRIAC ABBREVIATION MARK	#N/A	SAM	SAM	<none>	<none>
08E2	Cf	ARABIC DISPUTED END OF AYAH	#N/A	<none>	[<Arabic>]	#N/A	#N/A

Among the next characters, most have the property Indic_Syllabic_Category=Invisible_Stacker (labeled "InvSt" in this table). Users may not desire abbreviations for these, because of the already hinted reasons. By contrast, the initialism MVS is supported, as it stands for a format character. SUNDANESE SIGN VIRAMA seems to have a problem in its reference glyph, where the dotted circle is missing in the Code Charts, hence the NBSP instead of U+25CC in the table below. — In the "UAX #14 else" column, "<full>" means that the character is cited by its full Unicode name.

For the purpose of correcting UAX #14, U+1680 OGHAM SPACE MARK has special focus, since it stands out in the list of examples of space characters belonging to the line break class BA by not having any hint of an abbreviation, despite it has the White_Space property, despite it shows a line only in stemline fonts while being blank in stemless fonts, and despite UTC and the US NB advocated an abbreviation, namely "OGSP", in 2007 (113-A15) and 2008 (L2/08-142), when the reference glyph of U+1680 (lacking the dashed box by then) had come under scrutiny (L2/07-340, L2/07-392). In the wake, a discussion of OGHAM SPACE MARK was added as section 5.7 of UAX #14 for version Unicode 5.1 (2007).

cp	gc	InSC	Name	NaAl	Core	Charts	UAX14 lists	UAX14 else
1039	Mn	InvSt	MYANMAR SIGN VIRAMA	#N/A	[◌္], *virama*	[◌္]	#N/A	#N/A
1680	Zs		OGHAM SPACE MARK	#N/A	<none>	[ ],<blank>	<none>	<full>
17D2	Mn	InvSt	KHMER SIGN COENG	#N/A	[◌្]	[◌្]	#N/A	#N/A
180E	Cf		MONGOLIAN VOWEL SEPARATOR	MVS	MVS	MVS	(MVS)	<full>
1A60	Mn	InvSt	TAI THAM SIGN SAKOT	#N/A	<none>	[◌᩠]	#N/A	#N/A
1BAB	Mn	InvSt	SUNDANESE SIGN VIRAMA	#N/A	<none>	[᮫]	#N/A	#N/A

Now comes a whole range of unsupported space characters as of their abbreviations. One issue is with the leading digit in the Code Charts' mnemonic-in-dashed-box glyph of some of them, and another one is with the slash therein, both making a string unfit for NameAliases.txt. Yet that should not be a reason to deprive them of an abbreviation, as that digit can be replaced with the initial of its name, and that slash, with the capital letter P on the pattern of the (usually lowercase) "ppm" and "ppb" initialisms.

cp	Name	NaAl	Core	Charts	UAX14 lists	UAX14 else
2000	EN QUAD	#N/A	<none>	[NQSP]	<none>	<none>
2001	EM QUAD	#N/A	<none>	[MQSP]	<none>	<none>
2002	EN SPACE	#N/A	<none>	[ENSP]	<none>	<none>
2003	EM SPACE	#N/A	<none>	[EMSP]	<none>	<none>
2004	THREE-PER-EM SPACE	#N/A	<none>	[3/MSP]	<none>	<none>
2005	FOUR-PER-EM SPACE	#N/A	<none>	[4/MSP]	<none>	<none>
2006	SIX-PER-EM SPACE	#N/A	<none>	[6/MSP]	<none>	<none>
2007	FIGURE SPACE	#N/A	<none>	[FSP]	<none>	<full>
2008	PUNCTUATION SPACE	#N/A	<none>	[PSP]	<none>	<none>
2009	THIN SPACE	#N/A	<none>	[THSP]	<none>	<full>
200A	HAIR SPACE	#N/A	<none>	[HSP]	<none>	<none>

Next is a long list of various invisibles plus HYPHEN and NON-BREAKING HYPHEN. The first five are not in scope for this proposal, since their abbreviations are well supported. By contrast, LS for LINE SEPARATOR and PS for PARAGRAPH SEPARATOR are scarcely supported and need improvement, since although these characters are scarcely used in practice, they are in the Standard and thus require full support. The next six characters until NNBSP are again well supported. Not so the rest: UAX #14 does not give the abbreviation MMSP despite this is official; the initialism WJ is not used in TUS; the invisible mathematical symbols are lacking each one a Latin initialism despite those can be easily designed; and the Code Charts are lacking annotations indicating the abbreviations (yet in UCD) of the last four bidi controls.

The six bidirectional layout controls after this list are deprecated and are thus only in *annex C.*

cp	gc	WSpace	Name	NaAl	Core	Charts	UAX14 lists	UAX14 elsewhere
200B	Cf	No	ZERO WIDTH SPACE	ZWSP	ZWSP	ZWSP	(ZWSP)	ZWSP
200C	Cf	No	ZERO WIDTH NON-JOINER	ZWNJ	ZWNJ	ZWNJ	#N/A	#N/A
200D	Cf	No	ZERO WIDTH JOINER	ZWJ	ZWJ	ZWJ	(ZWJ)	ZWJ
200E	Cf	No	LEFT-TO-RIGHT MARK	LRM	LRM	LRM	#N/A	#N/A

200F	Cf	No	RIGHT-TO-LEFT MARK	RLM	RLM	RLM	#N/A	#N/A
2010	Pd	No	HYPHEN	#N/A	<none>	[-]	<none>	hyphen
2011	Pd	No	NON-BREAKING HYPHEN	#N/A	<none>	[NB-]	<none>	*non-breaking hyphen*
2028	Zl	Yes	LINE SEPARATOR	#N/A	LS	[LSEP]	<none>	<full>
2029	Zp	Yes	PARAGRAPH SEPARATOR	#N/A	PS	[PSEP]	<none>	<full>
202A	Cf	No	LEFT-TO-RIGHT EMBEDDING	LRE	LRE	LRE	#N/A	#N/A
202B	Cf	No	RIGHT-TO-LEFT EMBEDDING	RLE	RLE	RLE	#N/A	#N/A
202C	Cf	No	POP DIRECTIONAL FORMATTING	PDF	PDF	PDF	#N/A	#N/A
202D	Cf	No	LEFT-TO-RIGHT OVERRIDE	LRO	LRO	LRO	#N/A	#N/A
202E	Cf	No	RIGHT-TO-LEFT OVERRIDE	RLO	RLO	RLO	#N/A	#N/A
202F	Zs	Yes	NARROW NO-BREAK SPACE	NNBSP	NNBSP	NNBSP	(NNBSP)	NNBSP
205F	Zs	Yes	MEDIUM MATHEMATICAL SPACE	MMSP	<none>	MMSP	<none>	<none>
2060	Cf	No	WORD JOINER	WJ	<none>	WJ	(WJ)	WJ
2061	Cf	No	FUNCTION APPLICATION	#N/A	[f()]	[f()]	#N/A	#N/A
2062	Cf	No	INVISIBLE TIMES	#N/A	<none>	[×]	#N/A	#N/A
2063	Cf	No	INVISIBLE SEPARATOR	#N/A	<none>	[,]	#N/A	#N/A
2064	Cf	No	INVISIBLE PLUS	#N/A	<none>	[+]	#N/A	#N/A
2066	Cf	No	LEFT-TO-RIGHT ISOLATE	LRI	LRI	[LRI]	#N/A	#N/A
2067	Cf	No	RIGHT-TO-LEFT ISOLATE	RLI	RLI	[RLI]	#N/A	#N/A
2068	Cf	No	FIRST STRONG ISOLATE	FSI	FSI	[FSI]	#N/A	#N/A
2069	Cf	No	POP DIRECTIONAL ISOLATE	PDI	PDI	[PDI]	#N/A	#N/A

The last raw list snippet here starts with U+2D7F TIFINAGH CONSONANT JOINER mentioned above, and goes on with U+3000 IDEOGRAPHIC SPACE, cited above. The five *viramas* and two subjoiners are in the benefit of invisible *virama*-specific display. ZWNBSP/BOM shows no support problem, but U+1107F BRAHMI NUMBER JOINER raises concern by not having its initialism in NameAliases.txt, nor in the Code Charts except inside the dashed box. The two Kaithi number signs, having Kaithi labels, should for equity also have Latin abbreviations in the Standard. “InvSt” stands for “Invisible_Stacker”.

cp	gc	WSpace	InSC	Name	NaAl	Core	Charts	UAX14 lists	UAX14 else
2D7F	Mn	No		TIFINAGH CONSONANT JOINER	#N/A	[TFNCJ]	[<◌......>]	#N/A	#N/A
3000	Zs	Yes		IDEOGRAPHIC SPACE	#N/A	<none>	[IDSP]	<none>	<none>

AAF6	Mn	No	InvSt	MEETEI MAYEK VIRAMA	#N/A	<none>	[◌̟]	#N/A	#N/A
FEFF	Cf	No		ZERO WIDTH NO-BREAK SPACE	BOM, ZWNBSP	BOM, ZWNBSP	ZWNBSP, BOM	(ZWNBSP)	ZWNBSP
10A3F	Mn	No	InvSt	KHAROSHTHI VIRAMA	#N/A	[KV]	[◌̟]	#N/A	#N/A
1107F	Mn	No		BRAHMI NUMBER JOINER	#N/A	<none>	[BNJ]	#N/A	#N/A
110BD	Cf	No		KAITHI NUMBER SIGN	#N/A	<none>	[<Kaithi>]	#N/A	#N/A
110CD	Cf	No		KAITHI NUMBER SIGN ABOVE	#N/A	<none>	[<Kaithi>]	#N/A	#N/A
11133	Mn	No	InvSt	CHAKMA VIRAMA	#N/A	*virama*	[◌̟]	#N/A	#N/A
11A47	Mn	No	InvSt	ZANABAZAR SQUARE SUBJOINER	#N/A	[◌̟], subjoiner	[◌̟]	#N/A	#N/A
11A99	Mn	No	InvSt	SOYOMBO SUBJOINER	#N/A	subjoiner	[◌̟]	#N/A	#N/A
11D45	Mn	No	InvSt	MASARAM GONDI VIRAMA	#N/A	[◌̟], *virama*	[◌̟]	#N/A	#N/A
11D97	Mn	No	InvSt	GUNJALA GONDI VIRAMA	#N/A	[◌̟], *virama*	[◌̟]	#N/A	#N/A

Other format controls are used in Egyptian Hieroglyphs, in shorthand and in musical notation. Defining ASCII uppercase abbreviations for all of these is possible, and for each one of them a suggestion is made in *annex C.* They are not listed here because there is no issue with equity. Further, some Hieroglyph controls are since ever represented with ASCII symbols like asterisk or colon. For shorthand controls, lengthy Latin abbreviations may be considered pointless. The highly technical musical notation controls have most of their names spelled out even in dashed box glyphs. Proposing abbreviations for these characters as suggested in *annex C would require a separate proposal on a per-script basis. Hence these ranges are out of scope for this proposal.*

Annex C: Data overview

The full list of considered characters and suggested abbreviations is attached below.

Column headers

1. cp = Code Point
2. gc = General Category
3. Boolean = binary character properties with Boolean value (optional column)
4. Unicode name or ISO 6429 name (the latter for control codes)
5. NaAl = NameAliases.txt (the character's abbreviation therein, or #N/A)
6. Core = Core Specification *The Unicode Standard* 12.0.0 (abbreviation used, or <none>)
7. Charts = Code Charts (abbreviation in annotation, or bracketed if only in glyph, or other graphic)
8. UAX #14 lists = abbreviation in example list, or <none>, or #N/A if the character is not cited
9. UAX #14 else = how the character is referred to elsewhere in the document: abbreviation, or <full> if only the full, capitalized Unicode name is used, or <none> if it is mentioned only in an example list.
10. sugg = suggested abbreviation
11. uniq = uniqueness check for the suggested abbreviation

cp	gc	Boolean	InSC	Unicode name or ISO 6429 name	NaAl	Core	Charts	UAX14 lists	UAX14 els	sugg	uniq
0009	Cc	White_Space		CHARACTER TABULATION	HT, TAB	HT	HT	TAB	tab		
000A	Cc	White_Space		LINE FEED	LF	LF	LF	(LF)	LF		
000B	Cc	White_Space		LINE TABULATION	VT	VT	VT	(VT)	VT		
000C	Cc	White_Space		FORM FEED	FF	FF	FF	(FF)	FF		
000D	Cc	White_Space		CARRIAGE RETURN	CR	CR	CR	(CR)	CR		
0020	Zs	White_Space		SPACE	SP	SPACE	[SP]	(SP)	SP		
0085	Cc	White_Space		NEXT LINE	NEL	NEL	NEL	(NEL)	NEL, NL		
00A0	Zs	White_Space		NO-BREAK SPACE	NBSP	NBSP	NBSP	(NBSP)	NBSP		
00AD	Cf			SOFT HYPHEN	SHY	SHY	SHY	(SHY)	SHY		
034F	Mn			COMBINING GRAPHEME JOINER	CGJ	CGJ	CGJ	<none>	CGJ		
0600	Cf	Prepended_Concatenation_		ARABIC NUMBER SIGN	#N/A	<none>	[<Arabic>]	<collapsed>	<none>	ANS	OK
0601	Cf	Prepended_Concatenation_		ARABIC SIGN SANAH	#N/A	<none>	[<Arabic>]	<collapsed>	<none>	<none>	NOTOK
0602	Cf	Prepended_Concatenation_		ARABIC FOOTNOTE MARKER	#N/A	<none>	[<Arabic>]	<collapsed>	<none>	AFM	OK
0603	Cf	Prepended_Concatenation_		ARABIC SIGN SAFHA	#N/A	<none>	[<Arabic>]	<collapsed>	<none>	<none>	NOTOK
0604	Cf	Prepended_Concatenation_		ARABIC SIGN SAMVAT	#N/A	<none>	[<Arabic>]	<collapsed>	<none>	<none>	NOTOK
0605	Cf	Prepended_Concatenation		ARABIC NUMBER MARK ABOVE	#N/A	<none>	[<Arabic>]	#N/A	#N/A	ANMA	OK
061C	Cf	Bidi_Control		ARABIC LETTER MARK	ALM	ALM	ALM	#N/A	#N/A		
06DD	Cf	Prepended_Concatenation_		ARABIC END OF AYAH	#N/A	<none>	[<graphic>]	<none>	<none>	AEOA	OK
070F	Cf	Prepended_Concatenation_		SYRIAC ABBREVIATION MARK	#N/A	SAM	SAM	<none>	<none>	SAM	OK
08E2	Cf	Prepended_Concatenation_		ARABIC DISPUTED END OF AYAH	#N/A	<none>	[<Arabic>]	#N/A	#N/A	ADEOA	OK
1039	Mn		Invisible_St	MYANMAR SIGN VIRAMA	#N/A	[◌], virama	[◌]	#N/A	#N/A	<none>	NOTOK
1680	Zs	White_Space		OGHAM SPACE MARK	#N/A	<none>	[—], <blank>	<none>	<full>	OGSP	OK
17D2	Mn		Invisible_St	KHMER SIGN COENG	#N/A	[◌]	[◌]	#N/A	#N/A	<none>	NOTOK
180E	Cf			MONGOLIAN VOWEL SEPARATOR	MVS	MVS	MVS	(MVS)	<full>		
1A60	Mn		Invisible_St	TAI THAM SIGN SAKOT	#N/A	<none>	[◌]	#N/A	#N/A	<none>	NOTOK
1BAB	Mn		Invisible_St	SUNDANESE SIGN VIRAMA	#N/A	<none>	[᮪]	#N/A	#N/A	<none>	NOTOK
2000	Zs	White_Space		EN QUAD	#N/A	<none>	[NQSP]	<none>	<none>	NQSP	OK
2001	Zs	White_Space		EM QUAD	#N/A	<none>	[MQSP]	<none>	<none>	MQSP	OK
2002	Zs	White_Space		EN SPACE	#N/A	<none>	[ENSP]	<none>	<none>	ENSP	OK
2003	Zs	White_Space		EM SPACE	#N/A	<none>	[EMSP]	<none>	<none>	EMSP	OK
2004	Zs	White_Space		THREE-PER-EM SPACE	#N/A	<none>	[3/MSP]	<none>	<none>	THPMSP	OK
2005	Zs	White_Space		FOUR-PER-EM SPACE	#N/A	<none>	[4/MSP]	<none>	<none>	FPMSP	OK
2006	Zs	White_Space		SIX-PER-EM SPACE	#N/A	<none>	[6/MSP]	<none>	<none>	SPMSP	OK
2007	Zs	White_Space		FIGURE SPACE	#N/A	<none>	[FSP]	<none>	<full>	FSP	OK
2008	Zs	White_Space		PUNCTUATION SPACE	#N/A	<none>	[PSP]	<none>	<none>	PSP	OK
2009	Zs	White_Space		THIN SPACE	#N/A	<none>	[THSP]	<none>	<full>	THSP	OK
200A	Zs	White_Space		HAIR SPACE	#N/A	<none>	[HSP]	<none>	<none>	HSP	OK
200B	Cf			ZERO WIDTH SPACE	ZWSP	ZWSP	ZWSP	(ZWSP)	ZWSP		
200C	Cf			ZERO WIDTH NON-JOINER	ZWNJ	ZWNJ	ZWNJ	#N/A	#N/A		
200D	Cf			ZERO WIDTH JOINER	ZWJ	ZWJ	ZWJ	(ZWJ)	ZWJ		
200E	Cf	Bidi_Control		LEFT-TO-RIGHT MARK	LRM	LRM	LRM	#N/A	#N/A		
200F	Cf	Bidi_Control		RIGHT-TO-LEFT MARK	RLM	RLM	RLM	#N/A	#N/A		
2010	Pd	Hyphen		HYPHEN	#N/A	<none>	[-]	<none>	hyphen	HY	OK
2011	Pd	Hyphen		NON-BREAKING HYPHEN	#N/A	<none>	[NB-]	<none>	*non-breaki*	NBHY	OK
2028	Zl	White_Space		LINE SEPARATOR	#N/A	LS	[LSEP]	<none>	<full>	LS	OK
2029	Zp	White_Space		PARAGRAPH SEPARATOR	#N/A	PS	[PSEP]	<none>	<full>	PS	OK
202A	Cf	Bidi_Control		LEFT-TO-RIGHT EMBEDDING	LRE	LRE	LRE	#N/A	#N/A		
202B	Cf	Bidi_Control		RIGHT-TO-LEFT EMBEDDING	RLE	RLE	RLE	#N/A	#N/A		
202C	Cf	Bidi_Control		POP DIRECTIONAL FORMATTING	PDF	PDF	PDF	#N/A	#N/A		
202D	Cf	Bidi_Control		LEFT-TO-RIGHT OVERRIDE	LRO	LRO	LRO	#N/A	#N/A		
202E	Cf	Bidi_Control		RIGHT-TO-LEFT OVERRIDE	RLO	RLO	RLO	#N/A	#N/A		
202F	Zs	White_Space		NARROW NO-BREAK SPACE	NNBSP	NNBSP	NNBSP	(NNBSP)	<full>		
205F	Zs	White_Space		MEDIUM MATHEMATICAL SPACE	MMSP	<none>	MMSP	<none>	<none>		
2060	Cf			WORD JOINER	WJ	<none>	WJ	(WJ)	WJ, <full>		
2061	Cf	Other_Math		FUNCTION APPLICATION	#N/A	[f()]	[f()]	<collapsed>	#N/A	FA	OK
2062	Cf	Other_Math		INVISIBLE TIMES	#N/A	<none>	[×]	<collapsed>	#N/A	IMS	OK
2063	Cf	Other_Math		INVISIBLE SEPARATOR	#N/A	<none>	[,]	<collapsed>	#N/A	IS	OK
2064	Cf	Other_Math		INVISIBLE PLUS	#N/A	<none>	[+]	<collapsed>	#N/A	IPS	OK
2066	Cf	Bidi_Control		LEFT-TO-RIGHT ISOLATE	LRI	LRI	[LRI]	#N/A	#N/A		
2067	Cf	Bidi_Control		RIGHT-TO-LEFT ISOLATE	RLI	RLI	[RLI]	#N/A	#N/A		
2068	Cf	Bidi_Control		FIRST STRONG ISOLATE	FSI	FSI	[FSI]	#N/A	#N/A		
2069	Cf	Bidi_Control		POP DIRECTIONAL ISOLATE	PDI	PDI	[PDI]	#N/A	#N/A		
206A	Cf			INHIBIT SYMMETRIC SWAPPING	#N/A	<deprecated>	[ISS]	#N/A	#N/A		
206B	Cf			ACTIVATE SYMMETRIC SWAPPING	#N/A	<deprecated>	[ASS]	#N/A	#N/A		
206C	Cf			INHIBIT ARABIC FORM SHAPING	#N/A	<deprecated>	[IAFS]	#N/A	#N/A		
206D	Cf			ACTIVATE ARABIC FORM SHAPING	#N/A	<deprecated>	[AAFS]	#N/A	#N/A		
206E	Cf			NATIONAL DIGIT SHAPES	#N/A	<deprecated>	[NADS]	#N/A	#N/A		
206F	Cf			NOMINAL DIGIT SHAPES	#N/A	<deprecated>	[NODS]	#N/A	#N/A		
2D7F	Mn			TIFINAGH CONSONANT JOINER	#N/A	[TFNCJ]	[<◌......>]	#N/A	#N/A	TCJ	OK
3000	Zs	White_Space		IDEOGRAPHIC SPACE	#N/A	<none>	[IDSP]	<none>	<none>	IDSP	OK
AAF6	Mn		Invisible_St	MEETEI MAYEK VIRAMA	#N/A	<none>	[◌]	#N/A	#N/A	<none>	NOTOK
FEFF	Cf			ZERO WIDTH NO-BREAK SPACE	BOM, ZWNBSP	BOM, ZWNBSP	ZWNBSP, BOM	(ZWNBSP)	ZWNBSP		
10A3F	Mn		Invisible_St	KHAROSHTHI VIRAMA	#N/A	[KV]	[◌]	#N/A	#N/A	<none>	NOTOK
1107F	Mn		Number_Jo	BRAHMI NUMBER JOINER	#N/A	<none>	[BNJ]	#N/A	#N/A	BNJ	OK
110BD	Cf	Prepended_Concatenation_		KAITHI NUMBER SIGN	#N/A	<none>	[<Kaithi>]	<none>	#N/A	KNS	OK
110CD	Cf	Prepended_Concatenation_		KAITHI NUMBER SIGN ABOVE	#N/A	<none>	[<Kaithi>]	#N/A	#N/A	KNSA	OK
11133	Mn		Invisible_St	CHAKMA VIRAMA	#N/A	*virama*	[◌]	#N/A	#N/A	<none>	NOTOK
11A47	Mn		Invisible_St	ZANABAZAR SQUARE SUBJOINER	#N/A	[◌], subjoiner	[◌]	#N/A	#N/A	<none>	NOTOK
11A99	Mn		Invisible_St	SOYOMBO SUBJOINER	#N/A	subjoiner	[◌]	#N/A	#N/A	<none>	NOTOK
11D45	Mn		Invisible_St	MASARAM GONDI VIRAMA	#N/A	[◌], virama	[◌]	#N/A	#N/A	<none>	NOTOK
11D97	Mn		Invisible_St	GUNJALA GONDI VIRAMA	#N/A	[◌], virama	[◌]	#N/A	#N/A	<none>	NOTOK
13430	Cf			EGYPTIAN HIEROGLYPH VERTICAL JOINER	#N/A	[:]	[:]	#N/A	#N/A	EHVJ	OK
13431	Cf			EGYPTIAN HIEROGLYPH HORIZONTAL JOINER	#N/A	[*]	[*]	#N/A	#N/A	EHHJ	OK
13432	Cf			EGYPTIAN HIEROGLYPH INSERT AT TOP START	#N/A	[<◰>]	[<◰>]	#N/A	#N/A	EHITS	OK
13433	Cf			EGYPTIAN HIEROGLYPH INSERT AT BOTTOM STA	#N/A	[<◱>]	[<◱>]	#N/A	#N/A	EHIBS	OK
13434	Cf			EGYPTIAN HIEROGLYPH INSERT AT TOP END	#N/A	[<◳>]	[<◳>]	#N/A	#N/A	EHITJ	OK
13435	Cf			EGYPTIAN HIEROGLYPH INSERT AT BOTTOM END	#N/A	[<◲>]	[<◲>]	#N/A	#N/A	EHIBE	OK
13436	Cf			EGYPTIAN HIEROGLYPH OVERLAY MIDDLE	#N/A	[+]	[+]	#N/A	#N/A	EHOM	OK
13437	Cf			EGYPTIAN HIEROGLYPH BEGIN SEGMENT	#N/A	[(]	[(]	#N/A	#N/A	EHBS	OK
13438	Cf			EGYPTIAN HIEROGLYPH END SEGMENT	#N/A	[)]	[)]	#N/A	#N/A	EHES	OK
1BCA0	Cf			SHORTHAND FORMAT LETTER OVERLAP	#N/A	<none>	[<arrow>]	#N/A	#N/A	SHLO	OK
1BCA1	Cf			SHORTHAND FORMAT CONTINUING OVERLAP	#N/A	<none>	[<arrow>]	#N/A	#N/A	SHCO	OK
1BCA2	Cf			SHORTHAND FORMAT DOWN STEP	#N/A	<none>	[↓]	#N/A	#N/A	SHDS	OK
1BCA3	Cf			SHORTHAND FORMAT UP STEP	#N/A	<none>	[↑]	#N/A	#N/A	SHUS	OK
1D159	So			MUSICAL SYMBOL NULL NOTEHEAD	#N/A	ULL NOTE HEA	ULL NOTE HEA	#N/A	#N/A	MNNUL	OK
1D173	Cf			MUSICAL SYMBOL BEGIN BEAM	#N/A	[BEGIN BEAM]	[BEGIN BEAM]	#N/A	#N/A	MNBB	OK
1D174	Cf			MUSICAL SYMBOL END BEAM	#N/A	[END BEAM]	[END BEAM]	#N/A	#N/A	MNEB	OK
1D175	Cf			MUSICAL SYMBOL BEGIN TIE	#N/A	<none>	[BEGIN TIE]	#N/A	#N/A	MNBT	OK
1D176	Cf			MUSICAL SYMBOL END TIE	#N/A	<none>	[END TIE]	#N/A	#N/A	MNET	OK
1D177	Cf			MUSICAL SYMBOL BEGIN SLUR	#N/A	<none>	[BEGIN SLUR]	#N/A	#N/A	MNBS	OK
1D178	Cf			MUSICAL SYMBOL END SLUR	#N/A	<none>	[END SLUR]	#N/A	#N/A	MNES	OK
1D179	Cf			MUSICAL SYMBOL BEGIN PHRASE	#N/A	<none>	[BEGIN PHR.]	#N/A	#N/A	MNBP	OK
1D17A	Cf			MUSICAL SYMBOL END PHRASE	#N/A	<none>	[END PHR.]	#N/A	#N/A	MNEP	OK