Linguistics: special characters and typographic idiosyncrasies

Version 0.3.1, 12 May 2022

Version history:

- 0.1, 22 June 2020
- 0.2, 9 August 2021
- 0.3, 9 May 2022
- 0.3.1, 12 May 2022

Scope of this document

The number of linguistics characters in the Unicode Standard is enormous. No attempt is made here to cover all of them. The following are observations of phenomena that have had an impact on Brill’s treatment of linguistic texts. It should be noted that the term ‘linguistics’ can cover the study of specific languages; the study of ‘language’ as such (sometimes called ‘theoretical linguistics’); comparative linguistics; and philology, which is the study of all sorts of language phenomena within the context of traditional scholarly disciplines, such as Classical Studies, theology, Semitic Studies, Arabic Studies, Sinology, and so on.

Distinguishing ‘twins’ typographically

In linguistic representations – specifically IPA, but not limited to that system – two slightly different forms of the same Latin letter represent different phonemes. The following Latin characters are affected: Latin **a**, **f**, and **g**.

Some Greek characters used as phonetic symbols have a distinct ‘Latin’ shape. Most of them now have a Unicode code point of their own, but not all. Therefore, the current version of the Brill fonts (4.00) has the alternate glyph shapes still at their Greek code points but accessible through the OpenType Stylistic Set 20. These are and . Note that there are also other Latin-shaped Greek characters, among which are and , which have code points of their own (and).

Because of the subtlety of differences in appearance of these characters it is important to check (or spot-check) these characters by code point. The easiest way to do this in MS Office (Windows) is to copy the character whose Unicode value you wish to know from its source and paste it into a Word document. Once pasted, with the insertion point positioned just after the character in question, type **Alt X**, which converts the character to its Unicode hexadecimal value (typing **Alt X** again will toggle this back to the character). On macOS, you can use Character Viewer (sometimes referred to as ‘Emoji & Symbols’): in its Search field, paste the character whose value you wish to determine and it will show the required information instantly next to ‘Unicode’, as a hexadecimal value prefixed with ‘U+’.
For more information, see Using Unicode hexadecimal codes.

Latin twins in the Brill typeface

The letter a can be of the ‘two-storey’ kind, almost always found in serif typefaces; and it can be ‘single-storey’, as in many sans-serif typefaces (this latter, , is also known as ‘script a’ and ‘Latin alpha’). In serif typefaces, the regular or roman style normally has a two-storey design, whereas the italic is normally of the single-storey kind. In non-technical type it does not matter that the two are sightly different depending on the style.

In linguistic contexts, however, the Latin letters a, f, and g have ‘twins’ with subtly different shapes, and these represent different phonemes. In the table below, in the left-hand column the three Latin pairs of twins are listed:

Roman	Unicode	Italic	Italic (Stylistic Set 20)
a	U+0061	*a*	*a*

ɑ	U+0251	*ɑ*	
f	U+0066	*f*	*f*
ƒ	U+0192	*ƒ*	
g	U+0067	*g*	*g*
ɡ	U+0261	*ɡ*	

So what should a typesetter do if IPA text is italicised and you want the 'two-tier' shape of **a** to be retained? Apply the **OpenType Stylistic Set 20** to the character: this has been programmed into the Brill typeface (The Brill Typeface User Guide, p. 4).

Note: Even when the author has correctly applied the correct OpenType stylistic set to characters listed above, editors must still mark them for the typesetters. The OpenType ss20 attribute does not, unfortunately, carry over to most page layout applications such as Adobe InDesign!

Note also the following concerning the Latin twins mentioned above:

Character	Code point	Name	Remarks
	U+0251	Latin alpha or 'script a'	There is a capital, , U+2C6D, but this forms part of several Cameroon language orthographies, and it is not ordinarily used in strictly linguistic contexts. Note also the existence of U+1D45, U+0252, U+1D9B, U+AB64, and U+AB30.
f	U+0192	f with hook or 'script f'	Dutch florin (guilder); uppercase is , U+0191; do not confuse with lowercase abbreviation is, , U+A76D, or with lowercase dotless j with stroke and hook, , U+0284.
	U+0261	'script g'	IPA voiced velar plosive; uppercase is , U+A7AC.

Greek twins in the Brill typeface

In linguistics, the following Greek letters must take on a special 'Latin' shape, and in the Brill typeface these glyph shapes are accessible either via a dedicated Unicode point (which is preferred), or via the OpenType Stylistic Set 20:

Greek	Unicode	Latin shape (Stylistic Set 20)	Latin shape (Unicode)	Unicode Latin shape
β	U+03B2	(*do not use*)	β	U+A7B5
θ	U+03B8	θ		
λ	U+03BB	λ		
χ	U+03C7	(*do not use*)	χ	U+AB53

Important note: Even when the author has correctly applied the correct OpenType stylistic set to characters listed above, editors must still mark them for the typesetters, and the latter must be instructed to replace such characters with dedicated characters whenever available, such as in the case of and (U+A7B5, U+AB53). The OpenType ss20 attribute does not, unfortunately, carry over to most page layout applications such as Adobe InDesign!

Other Latin twins of Greek letters in Unicode

The following Greek letters have Latin twins with Unicode code points of their own, which clearly distinguish them from Greek-language characters. They are used mostly in linguistics contexts, although many of them have subsequently found a place in Latin orthographies of several African languages as well.

Character	Code point	Name	Remarks
	U+0251	**Latin alpha** or 'script a'	There is a capital, , U+2C6D, but this forms part of several Cameroon language orthographies, and it occurs but rarely in strictly linguistic contexts. Note also the existence of U+1D45, U+0252, U+1D9B, U+AB64, and U+AB30.
	U+A7B5	**Latin beta**	There is a capital, , U+A7B4, but this forms part of Gabonese orthographies, and it occurs but rarely in strictly linguistic contexts. Note the availability of the Latin glyph shape of Greek beta U+03B2 in the pre-version-4 Brill fonts through application of OpenType ss20.
	U+0263	**Latin gamma**	There is a capital, , U+0194, but this forms part of some African orthographies, and it occurs but rarely in strictly linguistic contexts. Note also the existence of U+02E0 Superscript Latin gamma, and U+0264 'Baby gamma' or 'ram's horns'.
	U+1E9F	**Latin delta** or 'script d' or 'insular d'	Note also the existence of U+018D turned delta.
	U+025B	**Latin epsilon** or 'open e'	There is a capital, , U+0190, but this forms part of some African (Niger-Congo) orthographies, and it occurs but rarely in strictly linguistic contexts. Note also the existence of U+1D93, U+025C, U+1D94, U+025D, U+1D9F, U+025E, U+029A, U+1D08, U+1D4B, and U+1D4C.
θ	U+03B8	**Latin theta**	This character has not yet been encoded in Unicode. The Latin glyph shape of Greek theta U+03B8 in the Brill fonts is accessible by application of OpenType ss20.
	U+0269	**Latin iota**	There is a capital, , U+0196, but this forms part of some African (Niger-Congo) orthographies, and it occurs but rarely in strictly linguistic contexts. Note also the existence of U+1DA5 and U+1D7C. Do not confuse with Cyrillic iota U+A647.
λ	U+03BB	**Latin lambda**	This character has not yet been encoded in Unicode. The Latin glyph shape of Greek lambda U+03BB in the Brill fonts is accessible by application of OpenType ss20. Note also the existence of U+019B.
	U+028A	**Latin upsilon**	There is a capital, U+01B1, but this forms part of some African (Niger-Congo) orthographies, and it occurs but rarely in strictly linguistic contexts. Note also the existence of U+1D7F and U+1DB7.
	U+0278	**Latin phi**	Note also the existence of U+1DB2 and U+2C77.

	U+AB53	**Latin khi**	Note the availability of the Latin glyph shape of Greek khi U+03C7 in the pre-version-4 Brill fonts through application of OpenType ss20. There is a capital, , U+A7B3, but this is only used in German dialectology. Note also the existence of U+AB54 and U+AB55.
	U+A7B7	**Latin omega**	There is a capital, , U+A7B6. Both are used in African orthographies. Note also the existence of U+0277 and U+AB65.

Confusables in linguistics

In linguistics, the following non-literal symbols are often confused:

WRONG character	Code point	Name	CORRECT character	Code point	Name	Remarks
Ø	U+00D8	Latin capital letter O with stroke		U+2205	Empty set	The 'empty set' is used in linguistics to denote a zero morpheme (null morpheme) or zero-grade *ablaut* (or phonological 'zero'). Often submitted by authors as Capital letter O with stroke.
=	U+003D	Equals sign		U+2E17	Double oblique hyphen	The 'double oblique hyphen' is often used in grammars, as a clitic marker or morpheme boundary marker. Often submitted by authors as Equals sign.
“ `` '' "	U+201C U+0060 (2×) U+0027 (2×) U+0022	Left double quotation mark; Grave accent (2×); Apostrophe (2×); Quotation mark		U+02BA	Modifier letter double prime	To **transliterate the Cyrillic hard sign** (capital *and* lowercase) in the Latin script. Note that the double prime consists of just **one** U+02BA character and that this exhibits no casing behaviour.
‘ ` '	U+2018; U+0060; U+0027	Left single quotation mark; Grave accent; Apostrophe		U+02B9	Modifier letter prime	To **transliterate the Cyrillic soft sign** (capital *and* lowercase) in the Latin script. Note that the single prime U+02B9 exhibits no casing behaviour.