L2/21-229

Exploratory document on a problematic casing pair used by some African orthographies

Eduardo Marín Silva

16/05/2021

Introduction. In this document I discuss the following casing pair (Ƀƀ), that according to Wikipedia, is used by some African orthographies. I myself, couldn't confirm the veracity of those claims, but this document is written by taking their word for it.
As far as I know, this casing pair isn't supported because the casing of the homoglyphs, point to entirely different characters. I explore a total of five possible encoding models to support this pair.

Status Quo. Here is a chart of the status quo giving the codepoints, glyphs (in serif and sans serif presentations), names and casing relations, as well as an explanation of their original function.
The last row contains illustrations of the first four encoding models, with the necessary glyph alternates highlighted in yellow.

Uppercase	Lowercase	Case Pairs	Original Function
0182 Ƀ-Ƀ Latin B with Topbar	0183 ƀ-ƀ	Ɓƀ-Ɓƀ	Part of the old Zhuang orthography, that mixed Latin and Cyrillic characters. It couldn't be unified with the Cyrillic Be, due to the different glyph for the lowercase, that is more similar to the Latin Small Letter 'ƀ' than to the Cyrillic Small Be 'б'. So a new pair of Latin letters were encoded due to technical considerations; creating an homoglyph with the Capital Cyrillic Be 'Б'. It represents the 'Voiced Bilabial Implosive'.
0181 Ɓ-Ɓ Latin B with Hook	0253 ɓ-ɓ	Ɓɓ-Ɓɓ	Part of many African orthographies, including the International African Alphabet (AIA), the African Reference Alphabet, the Pan-Nigerian Alphabet, as well as orthographies for Fula and Hausa. The orthographies, tend to use them for the same sound and the small letter is used by IPA for the same sound too.
0411 Б-Б Cyrillic Be	0431 б-б	Бб-Бб	Part of the Modern Cyrillic Alphabet; it typically represents the 'Voiced Bilabial Plosive'. Not to be confused with the 'Cyrillic Ve' (0412 В 0432 в) which is an homoglyph with the Capital Latin 'B' and its small capital, but does represent the 'Voiced Bilabial Implosible'.
A1) 0182 Б-Б A2) 0182 Б-Б A3) 0181 Ɓ-Б B) 0411 Б-Б	A1) 0253 ɓ-ɓ A2) 0183 ƀ-ɓ A3) 0253 ɓ-ɓ B) 0431 ƀ-ɓ	Ɓƀ-Бɓ	According to the French[1] and English[2] Wikipedia, the orthographies that use this pair are: a 1928 version of the AIA, plus some orthographies of some Mande languages (Dan, Kpelle and Loma), as well as old orthographies for Shona (1931-1935) and the Ndau dialect, and for a 1971 Liberian New Testament. Also, according to the Spanish Wikipedia[3], it was at least proposed for Xhosa and Zulu. It seems that this casing pair is more popular in Liberia overall, and it also seems to represent the 'Voiced Bilabial Implosive' too. Whether it's in current use is unknown.

In addition, the linguist Clement Doke used yet another different glyph for the capital for writing Shona: Ɓ[1]. In summary:

- The uppercase is identical to:
  - **0182 Ƀ LATIN CAPITAL LETTER B WITH TOPBAR**
  - **0411 Б CYRILLIC CAPITAL LETTER BE**
- The lowercase is identical to:
  - **0253 ɓ LATIN SMALL LETTER B WITH HOOK**
- And is somewhat similar to the letters:
  - **0183 ƀ LATIN SMALL LETTER B WITH TOPBAR**
  - **0431 б CYRILLIC SMALL LETTER BE**

Encoding models. Here we discuss different encoding models to support this orthography. The first 3 are labeled: **A1)**, **A2)** and **A3)**, because they all use already existing Latin characters, while model **B)** uses already existing Cyrillic characters. Model **C)** is the only one that requires new characters to be encoded.

- **A1) Use 0182 Ƃ and 0253 ɓ:** This is perhaps the most intuitive solution, the only downside is that these characters are already in casing relations with different characters; so a complicated ad-hoc method would be needed to implement it. This may not be viable without defining a new locale on the CLDR data, but I'm not too familiar with the logistics of the library.
- **A2) Use 0182 Ƃ and 0183 ƃ but make 'ɓ' an acceptable glyph variant of the latter:** This uses two letters that are already in a casing relation; however, making it so 0183 ƃ is only sometimes an homoglyph of 0253 ɓ, but in a completely different linguistic context (Zhuang language vs several African languages); is definitely bound to be problematic. We must also consider, that while 0183 ƃ currently has a very narrow scope, 0253 ɓ is way more broad in its applications.
- **A3) Use 0181 Ɓ and 0253 ɓ but make 'Ƃ' an acceptable glyph variant of the former:** This has the same downside as the previous model, but is different in two ways. First when comparing the pairs of **A2)** and **A3)** on their default glyphs, with respect to the proposed alternates, it's obvious that **A2)** has an edge when it comes to visual similarity (Ƃ-ɓ are more similar to each other, than Ɓ-Ƃ). However, 0181 Ɓ and 0253 ɓ were meant to represent the same African languages in question, and not an unrelated East-Asian language; this makes them much closer linguistically. Considering that the relevant users, must likely already use these characters makes them seem like an ideal choice. But I must repeat, making 0181 Ɓ only sometimes be an homoglyph, with two other linguistically distant characters (0182 Ƃ and 0411 Б) is doubly problematic in that respect (as compared with A2).
- **B) Use 0411 Б and 0431 б but make 'ɓ' an acceptable glyph variant of the latter:** This makes use of the Cyrillic casing pair. While б is visually similar to ɓ, that's about where the upsides end. Most obviously, they are not Latin characters, and it's almost certain the users perceive and have perceived them as such. While sporadic characters from other scripts inserted into Latin texts is not unheard of, this specific application would be unprecedented; because such mixture has only been attested in phonetic notation, as well as the old Zhuang orthography. Not to mention that the Cyrillic letters represent a different sound. While the Zhuang implemented Cyrillic characters, due to the proximity they had with the Soviets on the 30's. The same could not be said for the African Mande or Shona communities, that have only been exposed (historically) to Latin characters, the Arabic script and possibly other African-original writing systems. Furthermore, if changing the glyph of '0431 б' to 'ɓ' is acceptable, then changing the glyph of '0431 б to Ƃ' should have also been acceptable; invalidating the rationale behind the disunification of the 'Latin B with Topbar' pair, in the first place.
- **C) Encode two new Latin characters:** While this is perhaps the most elegant solution, it also seems like the most extreme. These new characters would each be confusable with other characters, however that hasn't stopped other disunifications under a similar rationale before. I have attached a chart, illustrating cases where problematic casing relations have necessitated new homoglyphs. The letter between square brackets, is the respective upper- or lowercase; [.] means the letter is caseless. Important glyph variants are placed next to each other with a hyphen between them.

Base character	Homoglyph(s)	Notes
006A j LATIN SMALL LETTER J [J] 0237 ȷ LATIN SMALL LETTER DOTLESS J [.]	03F3 j-ȷ GREEK LETTER YOT [J]	Originally, the Greek yot was conceived as a caseless version of the Latin small 'j', to be used on phonetic notation within the Greek script. It was not until later, that a capital version of the yot was conceived. Since it harmonized better with the shape of the letter iota 'ι', showing the yot without the dot became acceptable. The Latin dotless 'ȷ' is considered a different, caseless letter.
004A J LATIN CAPITAL LETTER J [j]	037F Ϳ GREEK CAPITAL LETTER YOT [j-ȷ]	
00D0 Đ LATIN CAPITAL LETTER ETH [ð]	0110 Đ LATIN CAPITAL LETTER D WITH STROKE [đ-ɖ] 0189 Ɖ LATIN CAPITAL LETTER AFRICAN D [ɖ]	These are a set three Latin homoglyphs, that have been disunified due to the different appearances of their lowercase letters [ð-[đ-ɖ]-ɖ] in different orthographies.
0259 ə LATIN SMALL LETTER SCHWA [Ə]	01DD ǝ LATIN SMALL LETTER TURNED E [Ǝ]	These two pairs tend to represent the same sound, but the glyphs for their capitals differ dramatically [Ə-Ǝ], with each one being preferred by different communities.
0294 ʔ LATIN LETTER GLOTTAL STOP [.]	0241 Ɂ LATIN CAPITAL LETTER GLOTTAL STOP [ʔ]	The glottal stop was originally a caseless letter for phonetic notation, but certain 'First Nations' orthographies treat it as a casing letter. Associating a casing relation to the already existing letter would have been technically problematic, so they were disunified.
0298 ʘ LATIN LETTER BILABIAL CLICK [.]	A668 Ꙩ CYRILLIC CAPITAL LETTER MONOCULAR O [ꙩ]	The origins of these two letters are completely different, with one being a caseless phonetic symbol, while the other is a variant of the Cyrillic 'O' used in certain liturgical texts. This demonstrates that when it comes to letter-shapes, confusables often emerge by accident.
03B4 δ GREEK SMALL LETTER DELTA [Δ]	1E9F ẟ LATIN SMALL LETTER DELTA [.]	The Latin delta is used for transcriptions of certain medieval manuscripts and it is caseless. It's unclear what necessitated the creation of that letter, instead of just using the Greek one.
03B5 ε GREEK SMALL LETTER EPSILON [Ɛ]	025B ɛ LATIN SMALL LETTER OPEN E [Ɛ] 0511 ԑ CYRILLIC LETTERS SMALL REVERSED ZE [Ɛ]	The Latin pair 'Ɛɛ' is based on the Greek pair 'Ɛε' and the letters are also called 'epsilon' by the user community. They represents a different sound form the Latin 'Ee'. The shape of the letters are distinct [Ee-Ɛε] in order to avoid confusability, since they appear concurrently. The glyph for the capital 'Ɛ' is just an enlarged version of the Small Greek Letter 'ε', and at the same time could be considered an allograph of the Latin 'E', called a 'script E'. While the Cyrillic pair 'Ɛԑ' is derived from the regular ze 'Ꙁꙁ' and is therefore paleographically distinct, it's not clear why the Latin pair couldn't have been used in Cyrillic text.
0395 Ε GREEK CAPITAL LETTER EPSILON [ε]	0045 E LATIN CAPITAL LETTER E [e] 0415 Е CYRILLIC CAPITAL LETTER IE [e]	
03B9 ι GREEK SMALL LETTER IOTA [I]	0269 ɪ LATIN SMALL LETTER IOTA [ɩ] A647 ꙇ CYRILLIC SMALL LETTER IOTA [ɩ]	There seems to be three pairs of 'iotas' [Iɪ-ɩı-ꙇι] for the Greek, Latin and Cyrillic scripts respectively. The Latin pair 'ɩı' only differs from the Greek 'Iι', because the Latin capital is just an enlarged lowercase, while the Greek capital is an homoglyph with the Latin Capital 'I'. The Cyrillic pair 'ꙇι' seems to be rendered differently, but only in some fonts. Interestingly, the Greek capital iota is also a homoglyph with two other Cyrillic characters: the capital old I [I[i-ɪ]] and the palochka [I[.-l]]. The former has a lowercase that is often shown without a dot (specially in sans serif presentation), and the latter was originally made to be caseless, but was later assigned a lowercase, that is a homoglyph to the Latin Small Letter L 'l'.
0399 Ι GREEK CAPITAL LETTER IOTA [ι]	0049 I LATIN CAPITAL LETTER I [i] 0406 І CYRILLIC CAPITAL LETTER *old* I [i-ɪ] 04C0 Ӏ CYRILLIC LETTER PALOCHKA [.-l]	

03C9 ω GREEK SMALL LETTER OMEGA [Ω]	A64D ꝍ CYRILLIC SMALL LETTER BROAD OMEGA [Ꝍ] A7B7 ꝱ LATIN SMALL LETTER OMEGA [ꝱ]	Like the iota, there seem to be three pairs of 'omegas' [Ωω-Ꝍꝍ-ꝱꝱ], for Greek, Cyrillic and Latin respectively. Once again, the Latin and Cyrillic versions of their capitals, are just their respective lowercases but enlarged. While the Cyrillic pair 'Ꝍꝍ' unlike the Latin pair 'ꝱꝱ', show up without the middle loop, they can regain it sans serif presentation [Ωω-ɯɯ-ɯɯ]. Further complicating things, is the fact that there is a 'regular' version of the Cyrillic Omega (0460-0461), with this one sometimes also having an enlarged lowercase as the capital 'Ѡѡ' and other times having a glyph more like from the Latin pair, but only for the uppercase 'ꝱѡ'.
0411 Б CYRILLIC CAPITAL LETTER BE [б]	0182 Ƃ LATIN CAPITAL LETTER B WITH TOPBAR [Ƃ]	As explained above, the Latin pair 'ƂƄ' was encoded for representing the old Zhuang orthography. It was only disunified from Cyrillic due to the different shapes of their lowercase letters [б-ƃ], otherwise the Cyrillic pair 'Бб' could have been used.

If new characters were to be encoded, I would name them like so:

XXXX Б LATIN CAPITAL LETTER BE

XXX1 ƃ LATIN SMALL LETTER BE

These names would mirror the ones for Cyrillic, which is fitting and overall the most elegant option I could come up with.

Personal preference. Out of all the models, **B)** would be the one I would dismiss outright. Model **A1)** seems like it could work, but it would necessitate new annotations to the code-charts. I suggest the following:

0181 Ɓ LATIN CAPITAL LETTER B WITH HOOK

• African languages

• Zulu, Pan-Nigerian alphabet *((replaced it with a more general note above, since more African alphabets include it))*

• lowercase is 0253 ɓ

0182 Ƃ LATIN CAPITAL LETTER B WITH TOPBAR

• in some African orthographies, this letter is the preferred uppercase of 0253 ɓ instead of 0181 Ɓ

→ 0411 Б cyrillic capital letter be

0183 ƃ LATIN SMALL LETTER B WITH TOPBAR

• Zhuang (old orthography)

• former Soviet minority language scripts

→ 0253 ɓ latin small letter b with hook

→ 0411 Б cyrillic capital letter be *((moved to the capital counterpart))*

→ 0431 б cyrillic small letter be

⋮

0253 ɓ LATIN SMALL LETTER B WITH HOOK

• implosive bilabial stop

• African languages

• Pan-Nigerian alphabet *((replaced it with a more general note above, since more African alphabets include it))*

• uppercase is 0181 Ɓ

• in some African orthographies, 0182 Ƃ is the preferred uppercase of this character

→ 0182 Ƃ latin capital letter b with topbar

→ 0431 б cyrillic small letter be

What would need to happen in the CLDR side of things is alien to me though. In any event, there needs to clear guidelines over how to handle problematic casing pairs like this one, because this is bound to keep happening.

Figures.

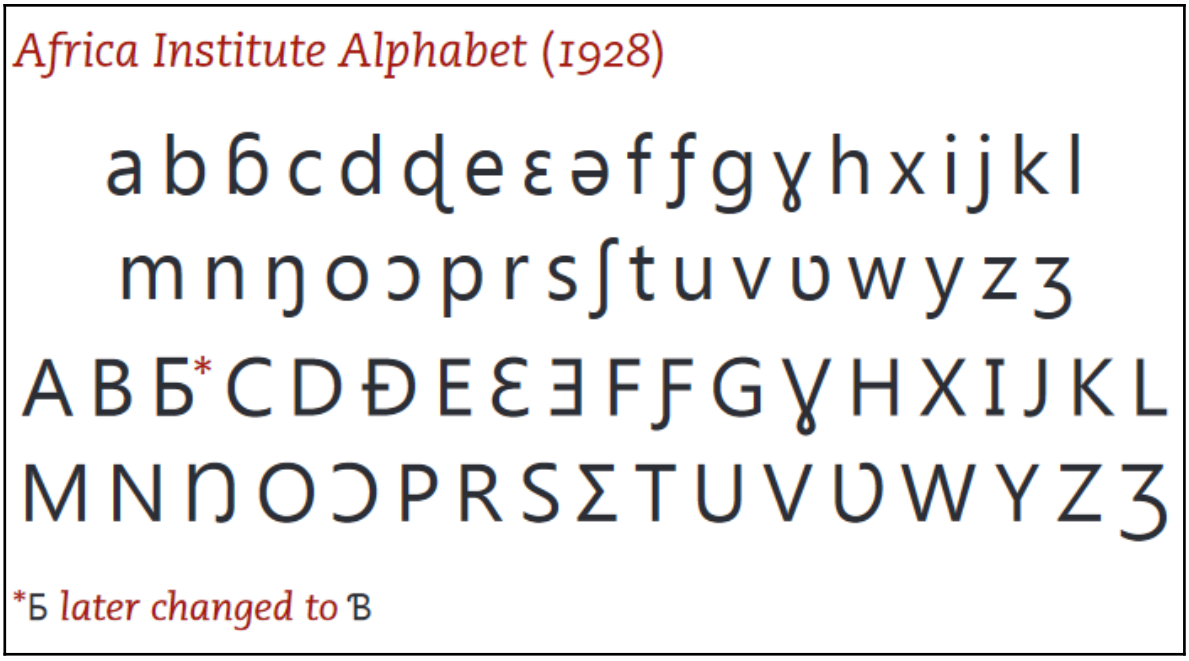

Figure 1. PDF illustrating different African alphabets (page 4)[4]

Formes du B crocheté

Majuscule	Minuscule	Description
Ɓ	ɓ	Forme majuscule basée sur le B majuscule.
Ƃ	ɓ	Forme majuscule sans panse supérieure ni crochet mais avec une barre horizontale ; notamment utilisée au Liberia et aussi dans l'orthographe du shona utilisée de 1931 à 1955.
Ƃ	ɓ	Forme majuscule sans panse supérieure avec un crochet droit comme partie supérieure, notamment utilisée par Clement Doke (en) dans l'orthographe du shona utilisée de 1931 à 1955.

Figure 2. French Wikipedia explaining glyph differences[1]

Sources.

[1] French Wikipedia page on Ɓ
[2] English Wikipedia page on Ɓ
[3] Spanish Wikipedia page on Ɓ
[4] PDF illustrating the different African alphabets