Unicode font challenges in linguistics

Berthold Crysmann

Laboratoire de linguistique formelle
crysmann@linguist.jussieu.fr

1 Scripts

In linguistics, we need support for various scripts, ideally covering the whole of unicode. While scripts from different traditions may follow entirely different design principles (compare Latin/Greek/Cyrillic vs. Arabic vs. CJK) and can therefore be recruited from fonts with different design without too much trouble, the design of scripts within a script family should match.

Thus, for Latin base fonts, we want matching, or better identical glyph designs for Greek and Cyrillic, and, of course the IPA.

The following 2 fonts appear to satisfy these conditions:

- Linux Libertine (French renaissance antiqua)
- FreeSerif (=Times; Baroque antiqua)

Charis SIL has very good unicode coverage for Latin and Cyrillic, including sophisticated accent placement, but, unfortunately no Greek (in text mode).

- Charis SIL (=Charter; Linear Antiqua)

Samples:

- Greek
  - Monotonic

    Γλυκά θροεί η κουκουναριά στης ρεμματιάς το πλάι, όμως και συ, γιδοβοσκέ, γλυκειά φλογέρα παίζεις· δώρο σου πρέπει δεύ- τερο, ύστερ' από τον Πάνα.
  - Polytonic

    Ἄνδρα μοι ἔννεπε, μοῦσα, πολύτροπον, ὃς μάλα πολλὰ πλάγχθη, ἐπεὶ Τροίης ἱερὸν πτολίεθρον ἔπερσεν: πολλῶν δ' ἀνθρώπων ἴδεν ἄστεα καὶ νόον ἔγνω, πολλὰ δ' ὅ γ' ἐν πόντῳ πάθεν ἄλγεα ὃν κατὰ θυμόν, ἀρνύμενος ἥν τε ψυχὴν καὶ νόστον ἑταίρων.
- Cyrillic
  - Bulgarian: Всички хора се раждат свободни и равни по достойнство и права.
  - Russian: Все люди рождаются свободными и равными в своем достоинстве и правах.

– Ukrainian: Всі люди народжуються вільними і рівними у своїй гідності та правах.

- IPA

  ˌɪntəˈnæʃn·l fəˈnɛtɪk ˈælfəˌbɛt

In addition to the IPA, many languages with latin-based scripts contain additional characters that are not merely accented variants of existing base characters. E.g. West-African languages have capitals for e.g. implosives, as in Hausa:

(1) Ɓarawo ya saci nama.
thief 3.sg.compl steal meat/animal
'The thief stole the meat.'

These glyphs are typically found in section Latin Extended B.

Unicode only caters for some precombined letters with diacritics. Combinations without their own code position can be produced with combining diacritics. Precombined and combining diacritics should be optically indistinguishable. This is typically achieved with OpenType rules.

(2) Ɓàrāwō yâ sǎci nāmǎ.
thief 3.sg.pot steal meat/animal
'The thief might steal the meat.'

(3) Ɓàrāwō yâ sǎci nāmǎ.
thief 3.sg.pot steal meat/animal
'The thief might steal the meat.'

Many fonts lack the appropriate mark-to-mark anchors for recursive stacking of diacritics: the SIL fonts are a notable exception. I have created versions of FreeSerif and Linux Libertine that do support this. Accent placement for Linux Libertine is done mostly automatically; for FreeSerif, I heavily take advantage of existing accent marks.

And finally, some Vietnamese:

> Tuyên ngôn toàn thế giớ i về nhân quyên của Liên Hợp Quốc. Tất cả mọi ngườ i sinh ra đều được tự do và bình đẳng về nhân và quyên. Mọi con ngườ i đêu được tạo hoá ban cho ly phâm phải đối xử với nhau trong tinh băng trí và lương tâm và cân hữu.

2 Math

Empirical linguistics and formal linguistics alike need good math support. Modulo kerning (which is necessarily different in math and text modes), multilingual text fonts should have accompanying math fonts. Math support includes not only symbols for statistics, set theory or logic but also stretchable operators for e.g. AVMs.

Owing to the limited availability of unicode math fonts, it seems necessary to rely on non-unicode LaTeX math packages, such as Mathdesign Charter for Charis, and newtxmath for Libertine. For FreeSerif only, there is an equally Times-based unicode math font, i.e. TG Termes Math.

- Comparison of text vs. math italics (compare character weight).

  a*a* b*b* f*f*

  Charter appears slightly lighter than Charis SIL, although the design matches.

- Some inline maths:

(4) a. For any leaf type t_1[mud μ_1,morsyn σ], t_2[mud μ_2,morsyn $\sigma \wedge \tau$] is a morphological competitor, iff $\mu_1 \subseteq \mu_2$.

b. For any leaf type t_1 with competitor t_2, expand t_1's morsyn σ with the negation of t_2's morsyn $\sigma \wedge \tau$: $\sigma \wedge \neg(\sigma \wedge \tau) \equiv \sigma \wedge \neg\tau$.

- AVMs with math operators

(5) $$\text{word} \rightarrow \begin{bmatrix} \text{morphs} & \boxed{e_1} \bigcirc \cdots \bigcirc \boxed{e_n} \\ \text{morsyn} & \boxed{0}\ (\boxed{m_1} \uplus \cdots \uplus \boxed{m_n}) \\ \text{rules} & \left\langle \begin{bmatrix} \text{morphs} & \boxed{e_1} \\ \text{mud} & \boxed{m_1} \\ \text{morsyn} & \boxed{0} \end{bmatrix}, \ldots, \begin{bmatrix} \text{morphs} & \boxed{e_n} \\ \text{mud} & \boxed{m_n} \\ \text{morsyn} & \boxed{0} \end{bmatrix} \right\rangle \end{bmatrix}$$

- Old-style figures (or not) in attribute names:

(6) $$\begin{bmatrix} \text{c-cont} & \begin{bmatrix} \text{hook} & \begin{bmatrix} \text{index} & \boxed{e} \\ \text{ltop} & \boxed{l} \end{bmatrix} \\ \text{rels} & \left\langle \begin{bmatrix} \text{pred} & \text{hortative-rel} \\ \text{lbl} & \boxed{l} \\ \text{arg1} & \boxed{h} \end{bmatrix} \right\rangle \\ \text{hcons} & \left\langle \boxed{h} =_q \boxed{l} \right\rangle \end{bmatrix} \\ \text{dtrs} & \left\langle \begin{bmatrix} \text{ss|l|cont} & \begin{bmatrix} \text{hook} & \begin{bmatrix} \text{index} & \boxed{e} \begin{bmatrix} \text{tam} & \text{subj} \end{bmatrix} \\ \text{ltop} & \boxed{l} \end{bmatrix} \\ \text{rels} & \left\langle \ldots \begin{bmatrix} \text{arg0} & \boxed{e} \\ \text{lbl} & \boxed{l} \end{bmatrix} \ldots \right\rangle \end{bmatrix} \end{bmatrix} \right\rangle \end{bmatrix}$$

Note: Although the unicode-math package permits substitution of math letters with glyphs from the unicode text font, this is typographically questionable, due to incorrect metrics.

3 Summary

If you want widest Unicode coverage, go for FreeSerif. Kerning is sometimes not optimal (e.g. Test).[1] In addition to LGC you even get Ethiopic, Arabic, Hebrew Devanagari, and lots more with an essentially Timesish design. Plus: TG Termes Math gives you a 100% matching unicode math font. Minus: The FreeSans font has some kerning issues (Test). There are no oldstyle figures.

[1] You may like to experiment substituting certain character ranges with TG Termes, but, watch out (!), there are no anchors for diacritics (yet). Of course, these may be added by script at first, but I do not know how this could be fed into the TeX Gyre workflow: in the long run, one would want to fine-tune anchor positions by hand.

Linux Libertine has very broad unicode coverage, although only LGC + Hebrew. The newtx-math package provides adequate math support, though not unicode math. Plus: there are 3 weights (normal semi-bold, extra-bold), a display font and a nice companion sans font. Unicode coverage for extra bold sans is reduced, though. The font provides old-style figures, as well as calligraphic ligatures.

Charis has very good general support for Latin and Cyrillic with sophisticated diacritic placement, but no Greek. Another minus: the Mathdesign Charter fonts are slightly too light, though one might adjust the stem width in FontForge by a few points, and hope that this does not mess up the metrics too much.

As for sans serif and mono fonts with good unicode coverage, there are the DejaVu Sans and DejaVu Sans Mono.[2]. There is even good non-unicode math support via the arev math package. I am not so sure that DejaVu goes that well with Times, though, because of their reduced ascenders.

[2]Some anchors for diacritics are present in DejaVu Sans. I shall provide a version soon with extensive anchors.