Field Encoding
Determining Character Needs

Martin Hosken
SIL Non-Roman Script Initiative

The Procedure

- Decide which characters are in your encoding
- Look 'em up in Unicode
- Create a nice 1:1 table to convert your data into Unicode
- Use Encore to build fonts, Keyman for keyboards, SILtec for mapping
- Job done!

How Tough Will it Get?

- Can I use an existing industry standard?
  - Do you have an eng in your encoding?
- How complex is the Unicode encoding of my script?
  - Bidi issues? Virama issues? Diacritics? Large character sets?
- How many characters am I using that aren't in Unicode yet?
  - Can I get away without having to use the PUA?

Unicode is not finished

Unicode is not Cleani-code

- Grey areas
- Make mistakes
  - Can't necessarily fix all mistakes
  - Want it now vs get it right
- Not everything encoded yet
- Unicode Technical Committee is final arbiter
  - Need to relate

Unicode is a volunteer organisation

- Industry has paid people's wages and given time
- Academic concerns are valued

NRTC2

Character vs Glyph

Home-grown encodings

- Visually motivated

Ambiguous use of character codes

- What is '!'?
  - Exclamation mark (U+0021)
  - Retroflex click (U+01C3)
- In SILIPA93, code x22 is '1'
  - Dotless i (U+0131)
  - i when with upper diacritic: ì
- Different language different encoding!

NRTC2

Unit vs Sequence

Procedure

- My strange combination character isn't there
- Use a sequence, if you can
- Except for some ‘new letters’
  - When using center diacritics (e.g. Ł)

New Characters

- Different normative properties

NRTC2

Unit vs Sequence

Same Characters

- Same normative properties
- Different informative properties
  - Case mappings
  - Sort order
  - Glyph variant

Well, almost!

- Unicode is not cleani-code!

Stylistic Variants

In Burmese: ဿ vs သ္သ

- Do you encode free variation or not?

In Latin: fi vs fi

- Which one do you map to?

Dumb Renderers

Don't encode new presentation forms

- The longer you wait the easier this is

Keep presentation forms separate

- Mess up analysis, keyboarding, etc.
- Then let them die!

Smart Fonts

- Support presentation forms as dumb glyphs

Unicode is Possible

It is not trivial
- Lot's of work to do the implementation

Can be confusing
- Get help

Encoding choices have far reaching implications
- Need to relate to others
- Takes time to resolve issues
- Don't hack. It will bite you!

Contact Information:

- Non-Roman Script Initiative
  SIL International
  7500 West Camp Wisdom Rd.
  Dallas, TX 75236
  (972) 708-7440
  fonts@sil.org