การรองรับภาษาไทยในซอฟต์แวร์โอเพนซอร์ส

เทพพิทักษ์ การุญบุญญานันท์
theppitak@gmail.com

Thai Linux Working Group

กันยายน 2564

การรองรับขั้นพื้นฐาน

- รหัสอักขระ
  - ยุค 8 บิต: TIS-620, ISO-8859-11
  - ยุค multilingual: Unicode, ISO/IEC 10646
- ข้อกำหนดท้องถิ่น (POSIX locale, ISO/IEC 14652, Unicode CLDR)
  - string collation (`LC_COLLATE`, ISO/IEC 14651, UTS #10)
  - date/time format
  - numeric format
  - currency & format
  - etc.

การแสดงข้อความ

- การจัดเรียงสระและวรรณยุกต์
  - หลบหาง ป ฝ ฟ (ฬ)

    น้านี้พีฟ่าปี่ป้าฝ้าฝุ่น
  - หลบหาง/แปลงรูป ฎ ฏ

    กฎุมพี ตรีกฎก กุฎฎุฏฎุ

    กฎุมพี ตรีกฎก กุฎฎุฏฎุ
  - แปลงรูป ญ ฐ

    กตัญญู ทิฏฐุชุกรรม

การแสดงข้อความ

- คาถาบาลี-สันสกฤต
  - ญ ฐ ตัดเชิง

    ปณฺณายติ ทิฏฐา
  - รองรับการซ้อนนิคหิตเหนือสระอิ (ที่ไม่ใช่สระอึ)

    จกฺขุสมิ̀ปิ

การแสดงข้อความ

- Implementation
  - ยุค Windows XP และก่อนหน้า: PUA glyphs
    - glyph หลายชุดสำหรับวางในตำแหน่งต่างๆ หรือแปลงรูป
    - rendering engine ต้องรู้ว่าจะใช้ชุดไหนเมื่อไร
    - ฟอนต์รองรับได้เท่าที่ rendering engine ทำ
    - Windows กับ Mac แยกใช้รหัสคนละชุด $\rightarrow$ ฟอนต์ใช้ข้ามระบบไม่ได้
    - Pango: detect และรองรับฟอนต์จากทั้งสองแพลตฟอร์ม

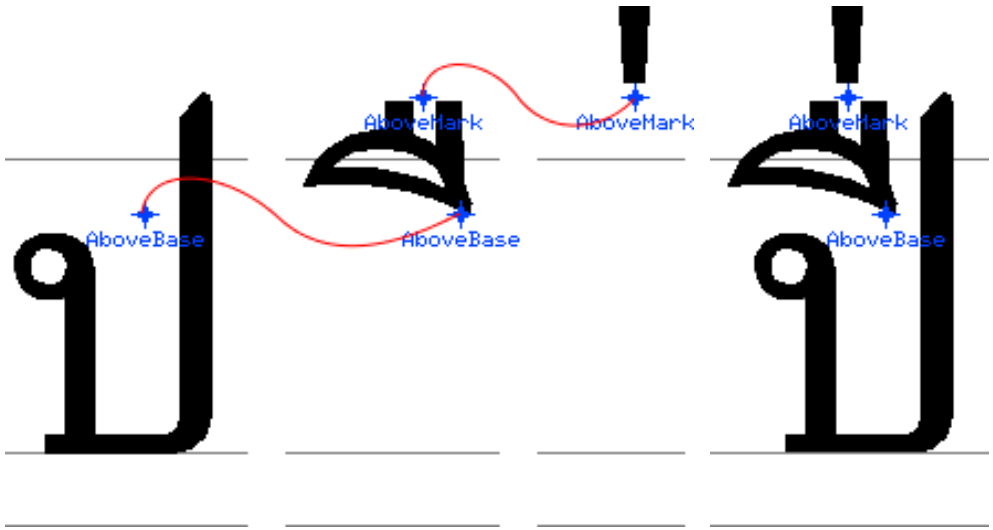
การแสดงข้อความ

- Implementation
  - ยุคหลัง Windows XP: OpenType
    - GSUB เลือก/เรียบเรียง glyph

    - ฟอนต์สามารถเพิ่มรูปแบบได้อย่างอิสระ

การแสดงข้อความ

- Implementation
  - ยุคหลัง Windows XP: OpenType
    - GPOS กำหนด anchor สำหรับวางซ้อนอักขระ



    - ไม่จำเป็นต้องเพิ่ม glyph เพื่อการจัดตำแหน่ง
    - ปรับตำแหน่งโดยละเอียดได้

ประเด็นเปิดของฟอนต์

- ความสูงของตัวอักษร
  - Scaled down (Windows, Mac)
    - TH Sarabun* 16pt + Arial 12pt
    - MS Word, LibreOffice ต้องแยกฟอนต์ CTL, Latin
  - Latin-compatible (Linux, LaTeX, Google Fonts)
    - Kinnari 12pt + Arial 12pt
    - ผสมกับฟอนต์สากลได้โดยไม่ต้องแยกขนาด

ประเด็นเปิดของฟอนต์

- ความสูงของตัวอักษร (ต่อ)
  - นิยามของ point size?

(ที่มา: เพจเฟซบุ๊ก “แกะรอยตัวพิมพ์ไทย”)

ประเด็นเปิดของฟอนต์

- ความสูงของตัวอักษร (ต่อ)
  - OS/2 Metrics: TypoAscent, TypoDescent, WinAscent, WinDescent

  - spec ของไมโครซอฟท์ที่ฟอนต์ไทยในวินโดวส์ไม่ใช้???

ประเด็นเปิดของฟอนต์

- การรองรับคาถาบาลี-สันสกฤต
  - การตัดเชิง ญ ฐ
    - mark up language เป็นบาลีหรือสันสกฤต
    - ใช้ stylistic alternative หรือ stylistic set
    - ปัจจุบัน:
      HTML/CSS, LibreOffice, LaTeX → ทำได้
      MS Office → ไม่รองรับ
  - การซ้อนนิคหิตเหนือสระอิ
    - GPOS
    - ปัจจุบัน:
      HTML/CSS, LibreOffice, LaTeX, MS Office → ทำได้

ประเด็นเปิดของฟอนต์

- การรองรับภาษาชาติพันธุ์
  - การยืมอักขระไปใช้ต่างหน้าที่ → ซ้อนอักขระได้ไม่จำกัด

    ปะเ့ติ็ล โ့ญ็จ ปั็วฮ့ ท็อง เปิ็ว มู่ย
    แต็ง เจ่อ เปริ่ห์ โจ๊ เปิ่ย โ့ทร ม็อง เต็ง อาึ ยาึ
    ปิ่ปี็ปี่ปั่ป่
    จือร့ การ့
  - ฟอนต์:
    - เตรียม anchor สำหรับการซ้อน
    - รองรับอักขระพิเศษ เช่น ตัวขีดเส้นใต้ (Macron)

Input Method

- Frameworks
  - XIM/XKB
  - GTK/Qt IM Module
  - SCIM
  - iBus
  - uim
  - fcitx

Input Method

- keyboard layout
  - เกษมณี, มอก.820-2538
  - ปัตตะโชติ
  - มนูญชัย!
- sequence check/correction
  - วทท 2.0
  - ภาษาชาติพันธุ์

Lexical Analysis

- Line break
  - word boundary
    - UAX #14 Unicode Line Breaking Algorithm
    - dictionary-based engine
  - hyphenation
    - TEX hyphenation patterns
- Word break
  - cursor by word
  - text double-click

งานแปลข้อความ

- กระบวนการ
  - peer review ก่อน submit
  - อภิปรายเพื่อเลือกคำแปล
  - พัฒนา glossary เพื่อคำแปลที่สอดคล้องกัน
  - ผู้ใช้รายงานปัญหาเพื่อปรับคำแปล
- ข้อสังเกต
  - มีผู้เปิดใช้คำแปลไทยมากขึ้น
  - งานแปลที่ไม่ใช่โอเพนซอร์สก็ดูมีคุณภาพมากขึ้น ;-)

Applications

- Spell check
- Soundex
- ...