อดีต ปัจจุบัน อนาคต ของการรองรับอักษรไทยใน FOSS

เทพพิทักษ์ การุญบุญญานันท์
theppitak@gmail.com

Thai Linux Working Group

มีนาคม 2568

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: Requirements

- การจัดเรียงสระและวรรณยุกต์
  - หลบหาง ป ฝ ฟ (ฬ)

    น้านี้ฬีฬ่าปี่ป้าฝ้าฝุ่น
  - แยกส่วนสระอำ

    ต่ำป่ำ
  - หลบหาง/แปลงรูป ฎ ฏ

    กฎุมพี ตรีกฎุก กุฎฎุฏฎุ

    กฎุมพี ตรีกฎุก กุฎฎุฏฎุ
  - แปลงรูป ญ ฐ

    กตัญญู ทิฏฐุชุกรรม

การแสดงข้อความ: Requirements

- คาถาบาลี-สันสกฤต
  - ญ ฐ ตัดเชิง

    ปณฺณายติ ทิฏฺฐา
  - รองรับการซ้อนนิคหิตเหนือสระอิ (ที่ไม่ใช่สระอึ)

    จกฺขุสมิ̀ปิ vs. จกฺขุสมีปิ

การแสดงข้อความ: Implementation

- ยุค Windows XP และก่อนหน้า: PUA glyphs
  - glyph หลายชุดสำหรับวางในตำแหน่งต่างๆ หรือแปลงรูป
    คู่คี่ปู่ปี่ฎูฐูีญูญี
  - rendering engine ต้องรู้ว่าจะใช้ชุดไหนเมื่อไร
  - ฟอนต์รองรับได้เท่าที่ rendering engine ทำ
    e.g. ยังไม่รองรับคาถาบาลี
  - Windows กับ Mac แยกใช้รหัสคนละชุด → ฟอนต์ใช้ข้ามระบบไม่ได้
  - Pango: detect และรองรับฟอนต์จากทั้งสองแพลตฟอร์ม

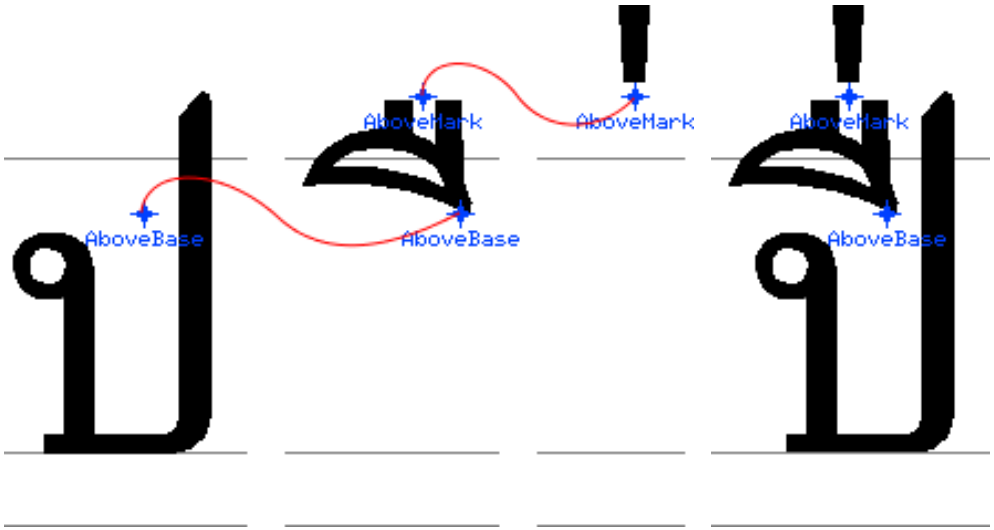
การแสดงข้อความ: Implementation

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

  - ฟอนต์สามารถเพิ่มรูปแบบได้อย่างอิสระ

การแสดงข้อความ: Implementation

- ยุคหลัง Windows XP: OpenType (ต่อ)
  - GPOS กำหนด anchor สำหรับวางซ้อนอักขระ



  - ไม่จำเป็นต้องเพิ่ม glyph เพื่อการจัดตำแหน่ง
  - ปรับตำแหน่งโดยละเอียดได้

การแสดงข้อความ: Issues

- การกำหนดขนาดตัวอักษร
  - Scaled down (Windows, Mac)
    - ฟอนต์ไทย-สากลขนาดเท่ากัน → ตัวหนังสือสูงๆ ต่ำๆ

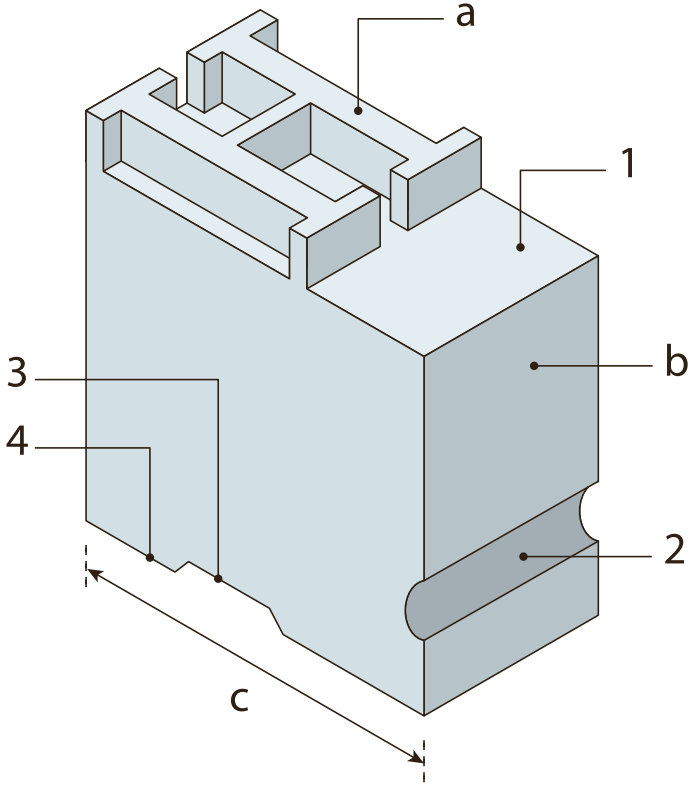
      ผู้ที่ใช้ Thai ก็ Up/Down អក្សរខ្មែរ ไม่ต้อง
    - ต้องจับคู่ฟอนต์ไทย 16pt + สากล 12pt ฯลฯ

      ผู้ที่ใช้ Thai ก็ Up/Down អក្សរខ្មែរ ไม่ต้อง
    - MS Word, LibreOffice ต้องแยกฟอนต์ CTL, Latin
  - Latin-compatible (Linux, LaTeX, Google Fonts)
    - ฟอนต์ไทย 12pt + สากล 12pt (ไม่ต้องแยกขนาด)

การแสดงข้อความ: Issues

- การกำหนดขนาดตัวอักษร (ต่อ)

point size ละติน



(ที่มา: Wikipedia)

ตัวพิมพ์ตะกั่วไทย

(ที่มา: เพจ “แกะรอยตัวพิมพ์ไทย”)

การแสดงข้อความ: Issues

- การกำหนดขนาดตัวอักษร (ต่อ)
  - ถ่างบรรทัดด้วยการแทรกตะกั่ว (leading)

(ที่มา: มูลนิธิเล็ก-ประไพ วิริยะพันธุ์)

การแสดงข้อความ: Issues

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

  - spec ของไมโครซอฟท์ที่ฟอนต์ไทยในวินโดวส์ไม่ใช้???
  - ปัจจุบัน: อักษรไทย = World's Last Exception!!!

การแสดงข้อความ: Issues

- การรองรับคาถาบาลี-สันสกฤต
  - การตัดเชิง ญ ฐ
    - mark up ภาษาเป็นบาลีหรือสันสกฤต (HTML/CSS, LibreOffice, LaTeX)
    - ใช้ stylistic set (HTML/CSS, MS Office)
  - การซ้อนนิคหิตเหนือสระอิ
    - GPOS
    - ปัจจุบัน:
      HTML/CSS, LibreOffice, LaTeX, MS Office → ทำได้

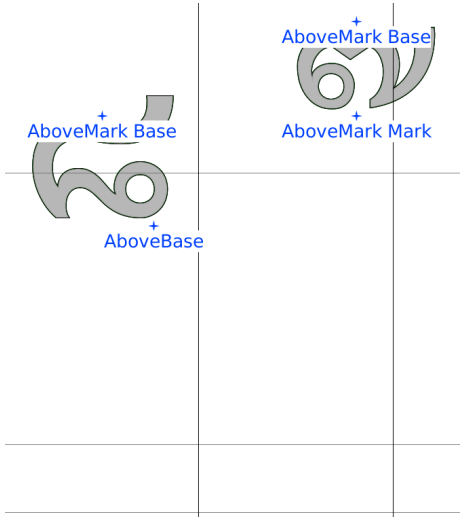
การแสดงข้อความ: Issues

- การรองรับภาษาชาติพันธุ์
  - กูย, เขมรถิ่นไทย, บรู, โส้, ช์อง, ญัฮกุร, ละว้า, มลายูปาตานี ฯลฯ
  - การยืมอักขระไปใช้ต่างหน้าที่ → ซ้อนอักขระได้ไม่จำกัด

    ปะเฺติ็ลฺ โฺญ็จฺ ปั็วฮฺ ท็อง เปิ็ว มู่ย
    แต็่ง เจฺอ เปริ่ห์ โจ๊ เปิ่ย โฺทร ม็อง เติ็ง อาึ ยาึ
    ปิ๎ปี็ป่ปั๋ปฺ่
    จือรุ๎ การุ๎

การแสดงข้อความ: Issues

- การรองรับภาษาชาติพันธุ์ (ต่อ)
  - ฟอนต์:
    - เตรียม anchor สำหรับการซ้อน



    - รองรับอักขระพิเศษ เช่น เครื่องหมายนาสิก (tilde), ตัวขีดเส้นใต้ (below macron):
      ปู๎ซะ ปา๎แง๎

การแสดงข้อความ: Issues

- การรองรับภาษาชาติพันธุ์ (ต่อ)
  - Rendering engine:
    - ยกเลิก วทท. 2.0 → ใช้ข้อกำหนด Unicode

แต็่ง เจุ๋อ โจ๊่ โ๋ทร อาื จือรุ๋

แต็่ง เจุ๋อ โจ๊่ โ๋ทร อาื จือรุ๋

การแสดงข้อความ: Issues

- ฟอนต์ IT9
  - 11 ส.ค. 2553 ก. ICT ผลักดัน 13 ฟอนต์ DIP-SIPA เป็นฟอนต์ราชการ
  - 31 ส.ค. 2553 ก. ศึกษาฯ เลือก 3 ฟอนต์:
    TH Sarabun PSK, TH Niramit AS, TH Chakra Petch
  - 3 ก.ย. 2553 สำนักนายกฯ เลือก TH Sarabun PSK
    และ *ให้ใช้เลขไทยเขียน พ.ศ.*
  - เกิดฟอนต์ TH Sarabun IT9 ที่แสดงเลขอารบิกเป็นเลขไทย
  - ราชการนึกว่าให้ใช้เลขไทยทั้งเอกสารด้วย IT9
    - เลขไทยใน URL
    - เลขไทยใน e-mail address
    - เลขไทยในสูตรเคมี
    - ฯลฯ

การแสดงข้อความ: Issues

- ฟอนต์ IT9 (ต่อ)
  - ปัญหา: สิ่งที่แสดง $\neq$ ข้อมูลจริง
    - การ search
    - การ คำนวณ
    - URL, e-mail address ป้อนตาม hard copy ไม่ได้
  - ปัญหาของการใช้เลขไทย:
    - เลขไทยอ่านยาก
    - ป้อนเลขไทยอย่างถูกวิธีได้ยาก
  - ข้อเท็จจริง:
    - ฟอนต์ IT9 ไม่ใช่ฟอนต์ตามมติ ครม. (มติคือ TH Sarabun PSK)
    - License: ผู้ดัดแปลงแจ้งเป็นลายลักษณ์อักษร แต่ไม่ได้รับอนุญาต
    - มติ ครม. ให้ใช้เลขไทยเขียน พ.ศ. เท่านั้น ไม่ใช่ให้ใช้ทั่วไป

การแสดงข้อความ: Agendas

1. เลิกย่อขนาด ใช้ขนาดสากล
   - Fonts-TLWG
   - GNU Freefonts
   - Google Fonts (Noto, Cadson Demak)
   - จุฬาภรณ์ลิขิต
2. รองรับการซ้อนตามข้อกำหนด Unicode (สำหรับภาษาชาติพันธุ์)
3. รองรับคาถาบาลี-สันสกฤต
4. ฟอนต์ IT9 จงพินาศ!
   - เข้าใจนโยบายให้ถูกต้อง *หรือ* ยกเลิก!
   - กำจัดฟอนต์ IT9
   - ใช้ Sarabun จาก Google Fonts
   - รองรับการป้อนเลขไทยด้วย NumPad

Input Method

- Frameworks
  - **XIM/XKB:** client-server (C, for X)
    ยกเว้น: local, ไทย ฝังมาใน Xlib!
  - **GTK/Qt IM Module:** loadable modules (C/C++)
    ไทย: `gtk-im-libthai`
  - **SCIM:** loadable modules (C++, for KDE, dead)
    ไทย: `scim-thai`
  - **IBus:** client-server (C/Python, GNOME default)
    ไทย: `ibus-libthai`
  - **uim:** loadble modules (C, for X, GTK, Qt, console, Emacs, Mac OS X)
    ไทย: ยังไม่มีมอดูลสำหรับ libthai, อาจใช้ XIM ผ่าน bridge ได้
  - **Fcitx 5:** loadble modules (C++)
    ไทย: `fcitx-libthai` (by upstream, based on `ibus-libthai`)

Input Method

- keyboard layout
  - เกษมณี, มอก.820-2538
  - ปัตตะโชติ
  - มนูญชัย!
- sequence check/correction
  - Client-side requirement: surrounding text retrieval/substitution
  - ภาษาไทย: วทท 2.0
  - ภาษาชาติพันธุ์: ยืดหยุ่น หรือ ตามภาษา?
- วาระ:
  - ป้อนเลขไทยด้วย NumPad

Lexical Analysis

- Line break
  - word boundary
    - UAX #14 Unicode Line Breaking Algorithm
    - dictionary-based engine: LibThai, ICU
  - hyphenation
    - TEX hyphenation patterns from `thailatex`
    - รองรับใน LibreOffice (deb: `hyphen-th`)
- Word break
  - cursor movement
  - text double-click

LibThai

- รวมโค้ดพื้นฐานของการรองรับภาษาไทย
  - `thctype`: แยกประเภทอักขระ
  - `thstr`: normalize string
  - `thcoll`: เรียงลำดับข้อความตามพจนานุกรม
  - `thcell`: แยกเซลล์แสดงผลตาม วทท. 2.0
  - `thrend`: เรียงข้อความเพื่อแสดงผลด้วย PUA glyphs ของ Win/Mac
  - `thinp`: input sequence check/correct ตาม วทท. 2.0
  - `thbrk`: แบ่งคำด้วยพจนานุกรม

LibThai

- ที่มีการใช้งานจริง
  - `thctype`: เรียกใช้จากภายใน
  - `thinp`: input method ต่างๆ
  - `thbrk`: Pango (Mozilla, LibreOffice ใช้ ICU)
- ที่เหลือ
  - `thstr`: ใช้ Unicode normalization
  - `thcoll`: ใช้ `strcoll()`, `strxfrm()` ของ glibc
  - `thcell`, `thrend`: ตกสมัยแล้ว
- วาระ:
  - ตัดโค้ดที่ไม่ได้ใช้
  - รองรับ Unicode โดยตรง

Mozilla

- Firefox 3: Platform word break engine
  - Linux: Pango (ผ่านไปยัง LibThai)
  - Windows: Uniscribe
  - Mac: Carbon
- Firefox 4: IME Surrounding text support
- Firefox 122: ICU BreakIterator แทน platform engine
  - ตัดคำเหมือนกันทุกแพลตฟอร์ม
  - รองรับทุกภาษาพร้อมกัน (จีน, ญี่ปุ่น, ไทย, ลาว, เขมร, พม่า)
  - ภาษาไทย: คุณภาพลดลง, word list ไม่มีการอัปเดต

- ยุคแรก (0.3.x)
  - ฟอนต์ PostScript Type1 + TeX Virtual Font (shaping)
  - Babel definition (คำแปล บทที่, สารบัญ ฯลฯ, เลขไทย, `\wbr`)
  - ตัดคำด้วย `swath` หรือ `cttex`
- ยุคที่สอง (0.4.x)
  - CTAN upload (`thai.dtx`, TDS)
  - แยกฟอนต์ไป build ในแพกเกจฟอนต์

- ยุคที่สาม (0.5.x)
  - hyphenation support
    - Manually hyphenated LibThai word list
    - Semi-automated script for `patgen` processing
- ยุคที่สี่ (สลายตัว)
  - ส่ง hyphenation pattern เข้า `hyph-utf8` (และ LibreOffice)
  - ส่ง Babel definition เข้า `babel-contrib`
  - ไม่มี `thailatex` แต่ยังมี `swath`

- TeX engines
  - pdfTeX
    - ใช้ทรัพยากรจากผลงาน `thailatex` ใน TDS
    - typeset ภาษาไทยด้วยเทคนิค PUA ผ่าน TeX virtual font
    - ตัดคำด้วยโปรแกรมระหว่างกลาง เช่น `swath`
  - XeTeX
    - ใช้ฟอนต์ OpenType จาก OS
    - ตัดคำด้วย ICU
  - LuaTeX
    - ใช้ฟอนต์ OpenType ใน TDS
    - สามารถเข้าถึงกระบวนการภายในผ่าน Lua (ยังไม่เริ่มงานรองรับภาษาไทย)

Thesaurus

- mythes
  - วรรณพงษ์ ภัททิยไพบูลย์ (PyThaiNLP): thai-synonym
  - me:
    - เพิ่มรายการ synonym
    - เขียนสคริปต์แปลงข้อมูลเป็น thesaurus
    - ส่งเข้า LibreOffice dictionary
    - deb: `mythes-th`

Spell Check

- hunspell
  - ศิลา ชุณห์วิจิตรา (NECTEC): รวบรวม word list
    - LibThai
    - LEXiTRON
    - NECTEC Corpus
    - ชื่อประเทศ ชื่อเมือง (ราชบัณฑิตยสถาน)
  - deb: `hunspell-th`
  - me:
    - กำจัดคำสะกดผิด
    - ปรับปรุงการแนะนำตัวสะกด
    - (อนาคต) ปรับอัลกอริทึม

Grammar Check

- ตัวอย่าง
  - การเว้นวรรคหลังไม้ยมก, จุด, จุลภาค (e.g. ต่างๆนานา, ดร.สลัมป์)
  - การใช้ไม้ยมกซ้ำคำผิดที่ (e.g. นาๆ, 10 ชิ้นๆ ละ 2 บาท ฯลฯ)
  - การใช้เลขไทยปนอารบิก (มรดก IT9)
  - ฯลฯ
- TBD

งานแปลข้อความ

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

Debian

- me: Debian Developer ตั้งแต่ปี 2552
- ผลักดันทรัพยากรภาษาไทย
- รับ feedback ปรับปรุงคุณภาพ
- หน้าที่อื่นๆ
  - sponsor uploads
  - ช่วยเหลือผู้สนใจเข้าร่วม (key signing/endorsement, advocate ฯลฯ)
  - evangelist
  - ร่วมกิจกรรมต่างๆ (voting, bug squashing ฯลฯ)