การพัฒนาภาษาไทยใน GNOME

เทพพิทักษ์ การุญบุญญานันท์

การสัมมนา “การส่งเสริมซอฟต์แวร์โอเพนซอร์สในประเทศไทย”
28 มีนาคม 2548

1 การจัดแสดงข้อความ

1.1 Pango Rendering Engine

Pango Rendering Engine

- โครงสร้างพื้นฐานสำหรับแสดงข้อความนานาภาษา
- ใช้โดย GTK+, libgnomeprint, libcroco
- ไม่ขึ้นกับ GNOME
- หน้าที่: flow ข้อความ (ยูนิโค้ด) ในคอลัมน์ที่กำหนด
- แยกมอดูลตามภาษา
  - Language Engine: วิเคราะห์จุด break ต่าง ๆ ของข้อความ
  - Shaping Engine: จัดเรียง glyph จากข้อความลงในบรรทัด

1.2 การวาดข้อความ

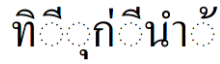
1.2.1 การจัด cluster ตาม วทท 2.0

วทท 2.0

- ลำดับบังคับ: สระบน/ล่าง มาก่อนวรรณยุกต์/ทัณฑฆาตเสมอ
- Input Method ตรวจลำดับ เพื่อรับประกันลำดับที่ถูกต้อง
- Output Method เผื่อลำดับที่ผิด โดยแสดงต่างจากลำดับที่ถูก

→ การจัดการข้อความภาษาไทยทำได้ง่าย ถ้ามีลำดับที่ถูกเพียงแบบเดียว

→ Pango สนับสนุนการวาดข้อความตาม วทท 2.0



รูปที่ 1: ตัวอย่างการแสดงลำดับที่ผิด

1.2.2 การปรับตำแหน่งสระ/วรรณยุกต์

ฟอนต์ที่สนับสนุน

- ฟอนต์ TrueType + Thai Windows PUA
- ฟอนต์ TrueType + Thai MacIntosh PUA
- ฟอนต์ *OpenType*

Angsana New	*Thonburi*	*Norasi*
(Win PUA)	*(Mac PUA)*	*(OpenType)*

รูปที่ 2: ตัวอย่างการวาดฟอนต์แบบต่าง ๆ

1.2.3 การปรับ glyph ด้วย PUA

หลักการ

- ปรับตำแหน่ง glyph โดยแทนด้วย glyph ใหม่ที่เลื่อนเตรียมไว้
- PUA สำหรับภาษาไทย
  - สระบน นิคหิต ไม้ไต่คู้ เพิ่ม 1 ชุด
  - สระล่าง พินทุ เพิ่ม 1 ชุด
  - วรรณยุกต์ ทัณฑฆาต เพิ่ม 3 ชุด
  - ญ หญิง, ฐ ฐาน ตัดเชิง

การปรับ glyph ด้วย PUA

- PUA (Private Use Area):
  - ที่ว่างในตารางยูนิโค้ดช่วง E000-F8FF
  - ไม่กำหนดอักขระ เว้นไว้ให้ผู้ผลิตใช้ *เป็นการภายใน*
- MS และ Apple กำหนด glyph ปรับตำแหน่งของไทยไว้ใน PUA *ซึ่งเป็นคนละเซ็ตกัน* → ใช้ฟอนต์ร่วมกันไม่ได้

- Pango Thai Module: detect ชุด PUA แล้วปรับ glyph ตามประเภท

  → ใช้ฟอนต์ TrueType ทั้งของ Windows และ Mac ที่มีอยู่ในท้องตลาดได้

1.2.4 การปรับ glyph ด้วย OpenType

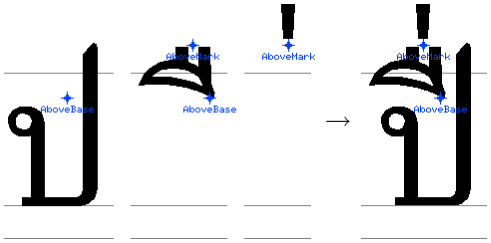
GSUB

- 'ccmp'
  - เลือกชุดวรรณยุกต์ต่ำ-สูง
  - ตัดเชิง ญ ฐ ที่มีสระล่าง
  - แยกส่วนสระอำ และจัดลำดับวรรณยุกต์ที่ประกอบอยู่

รูปที่ 3: กรณีที่จัดการด้วย GSUB

GPOS

- 3 anchor class: AboveBase ('mark'), AboveMark ('mkmk'), BelowBase ('mark')
- จัดการการวางสระ-วรรณยุกต์กรณีต่าง ๆ ด้วย anchor



รูปที่ 4: การจัดตำแหน่ง glyph ด้วย GPOS

ข้อดีของการใช้ OpenType

- **ได้มาตรฐาน:** ไม่บังคับให้ฟอนต์ต้องมี PUA glyph (ซึ่งเป็น proprietary standard)
- **ขยายขอบเขตได้ง่าย:** เพิ่มกรณีภาษาบาลีและกุยได้ทันที โดยไม่ต้องกำหนด PUA ใหม่
- **ปรับละเอียดได้:** ปรับตำแหน่งของเครื่องหมายต่าง ๆ ได้อย่างละเอียด

รูปที่ 5: ตัวอย่างสิ่งที่เพิ่มมาเมื่อใช้ OpenType

1.3 การแบ่งคำ

Pango-LibThai Language Engine

- LibThai: word break โดยอาศัยโค้ดจาก cttex
- `http://libthai.sourceforge.net/`
- Pango: เรียก language engine เมื่อจะวิเคราะห์ข้อความก่อน flow
- Pango-LibThai: Pango language engine ไทย ซึ่งเรียกใช้ LibThai เพื่อตัดคำ

2 การป้อนข้อความ

2.1 โครงสร้าง GTK+ IM

Input Methods Framework สำหรับ GTK+

- gtk-im: IM framework ซึ่งเชื่อมต่อในเทอมของ GDK event, GTK+ signal
- GTK+ IM Module: มอดูลย่อยซึ่งเลือกเปลี่ยนได้จาก context menu ของ text entry
- ตัวอย่าง GTK+ IM:
  - Default: แปลง keyboard symbol เป็นอักขระยูนิโค้ดตรง ๆ
  - IPA: ป้อนอักขระ IPA โดยอาศัยการ compose
  - X Input Method: bridge ระหว่าง XIM กับ GTK+
  - Internet/Intranet Input Method: bridge ระหว่าง IIIMF กับ GTK+

2.2 Input Method ภาษาไทย

IM Module ที่สนับสนุนภาษาไทย

1. Default: IM ทั่วไป แปลง keyboard symbol เป็นอักขระโดยตรง

   → ไม่มีการตรวจลำดับ

2. Thai (broken): IM ไทยใน GTK+ เริ่มแรก แปลง keyboard symbol เป็นอักขระโดยตรง

→ ไม่มีการตรวจลำดับ

3. X Input Method: เรียกใช้ XIM ไทยใน X

   → ตั้งระดับการตรวจลำดับตามข้อกำหนด XIM ไทย + แก้ลำดับได้

4. Thai-Lao: (Bug #81031) เสนอ IM module สำหรับภาษาไทยและลาว

   → ตรวจ-แก้ลำดับได้

5. gtk-im-libthai: third-party IM module โดยใช้ libthai

   → ตรวจ-แก้ลำดับได้

6. Internet/Intranet Input Method: เรียกใช้ IIIM ไทยจาก OpenI18N

   → ตรวจลำดับได้ แต่แก้ไม่ได้

แนวทางพัฒนาในอนาคต

- Default → ไม่ต้องทำอะไร และไม่แนะนำให้ใช้
- Thai (broken) → ขอลบออกจาก GTK+
- X Input Method → คงไว้ *เป็นวิธีที่แนะนำให้ใช้ในขณะนี้*
- Thai-Lao → เสนอให้รวมใน gtk-im-extras (เพื่อภาษาลาว)
- gtk-im-libthai → เผยแพร่เป็นทางเลือกอิสระ (เผื่อระบบอื่นที่ไม่ใช่ X Window)
- Internet/Intranet Input Method → เป็นวิธีที่น่าจะใช้ในอนาคต

2.3 การสนับสนุนใน Application

สิ่งที่โปรแกรมต้องสนับสนุน

- การอ่านตัวอักษรก่อนตำแหน่งเคอร์เซอร์
  - เพื่อพิจารณาความถูกลำดับของอักขระใหม่
  - "`retrieve-surrounding`" signal
- (optional) การลบตัวอักษรในบริเวณที่กำหนด
  - เพื่อแก้ไขลำดับที่ผิดให้ถูกต้อง
  - "`delete-surrounding`" signal

โปรแกรมที่สนับสนุน IM ไทยแล้ว

- GTK+ widgets: GtkEntry, GtkTextView
- Gnumeric: GnmCanvas

- Eel (Nautilus): EelEditableLabel
- AbiWord
- Gal (Evolution): EText

3 งานแปลข้อความ

3.1 ประวัติ

ประวัติ

- เริ่มแปล GNOME 2.4 โดย คุณไพศาข์ สีเหลืองสวัสดิ์
- งานแปลสมทบใน GNOME 2.6 โดย OpenTLE และนิสิต ม.บูรพา
- ปรับปรุงคำแปลใน GNOME 2.8 เล็กน้อย
- แปลเพิ่มใน GNOME 2.10 = 45.46% (2005-03-18)

3.2 สถานะ

เกณฑ์ของ "supported languages":

- ไม่เกิน 50% → Unsupported
- ไม่เกิน 80% → Partially Supported
- เกิน 80% → Supported

→ ภาษาไทย = *Unsupported*

สรุป

สถานะ

- **การแสดงผล**

  → สนับสนุนการจัดแสดงระดับคุณภาพ ด้วยฟอนต์ TrueType ไทยในท้องตลาดแทบทุกแบบ รวมทั้ง *OpenType*

- **การแบ่งคำ**

  → สนับสนุนด้วย Pango-Libthai (third party)

- **การป้อนข้อความ**

  → สนับสนุนด้วย XIM (อนาคตอาจใช้ IIIMF เป็นหลัก)

→ ฝั่ง application สนับสนุนเกือบทั้งหมดแล้ว

- **งานแปล**

  → GNOME 2.10 แปลเกือบถึง 50% แล้ว

งานในอนาคต

- พิจารณา check-in IM Module ไทย-ลาว ใน gtk-im-extras
- สำรวจและแก้ GNOME application ที่ยังไม่สนับสนุน IM ไทย
- ตรวจสอบการตรวจ-แก้ลำดับการป้อนภาษาไทยใน IIIMF
- พิจารณาหา solution สำหรับ language engine แบบไม่ใช่ third party
- ตรวจคำแปลที่ไม่เหมาะสม และแปลเพิ่มเพื่อเขยิบสู่สถานะ partially supported