ยูนิโคดกับการพัฒนาฟอนต์อักษรโบราณอีสาน

เทพพิทักษ์ การุญบุญญานันท์

แนะนำตัว

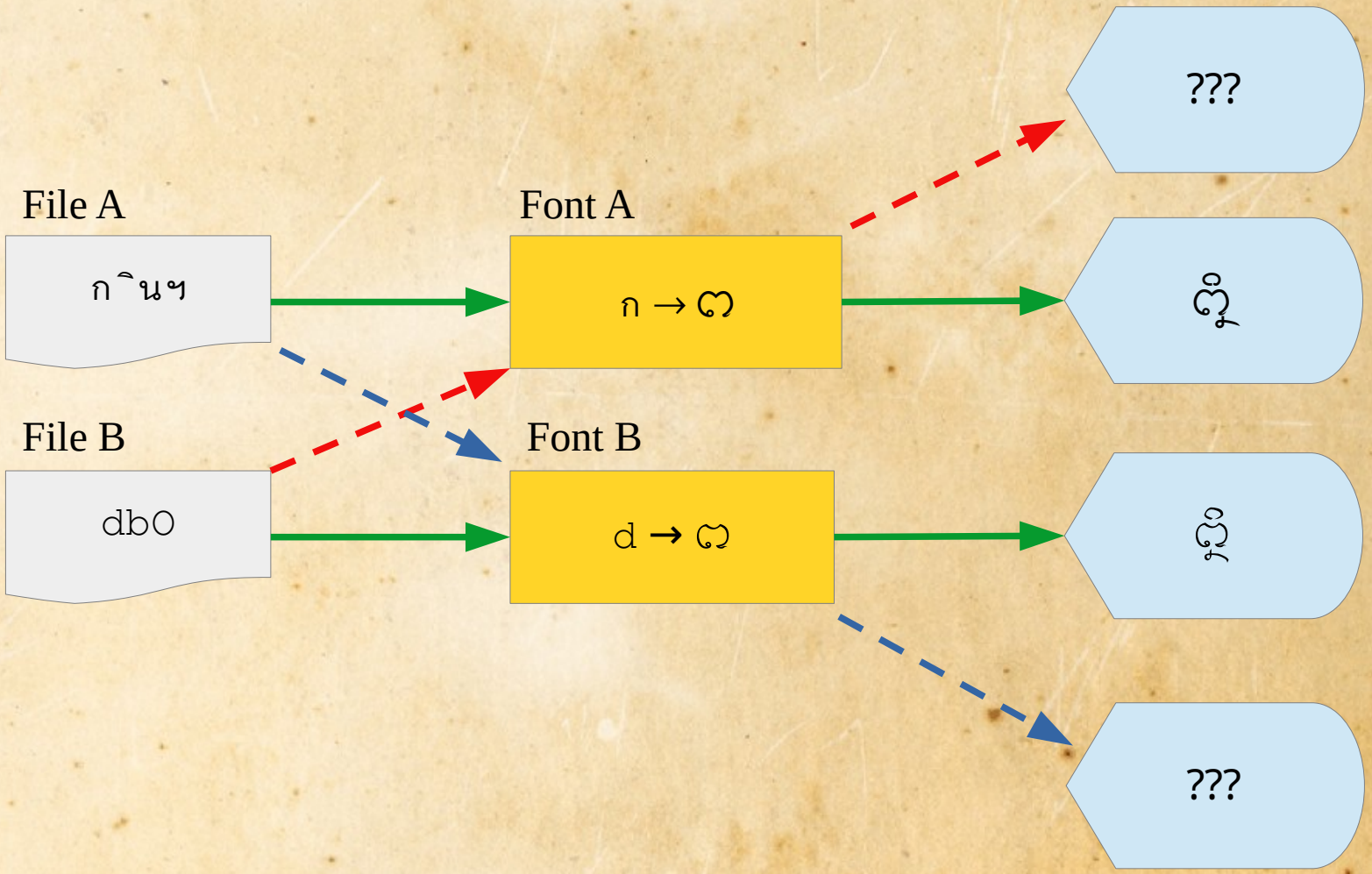
- งานที่ NECTEC ทำให้ได้ทำงานมาตรฐาน
- CICC (ญี่ปุ่น) หนุนสำรวจอักษรท้องถิ่นเสนอยูนิโคด
- รายงานอักษรล้านนา, ยาวี
- อ.ทวี สว่างปัญญางกูร: 13 ฟอนต์อักษรเผ่าไท: ล้านนา ไต้คง ไทดำ อาหม คำตี่ ฯลฯ

แนะนำตัว

- กลับมาทำงานที่ขอนแก่นบ้านเกิด
- สนใจภาษาอีสาน: คำศัพท์, หน่วยเสียง, ไตรยางศ์
- ศึกษาอักษรอีสาน: อักษรธรรม, ไทยน้อย
- พัฒนาฟอนต์ Khottabun (ธรรมอีสาน + ไทยน้อย)
- เบื้องหลัง: เอกสารในเน็ต, หนังสือห้องสมุด, ใบลาน
- สื่อสารกับผู้รู้ ทำงานร่วมกับผู้เชี่ยวชาญยูนิโคดตามโอกาส

ทำไมต้องยูนิโคด?

ก่อนยูนิโคด



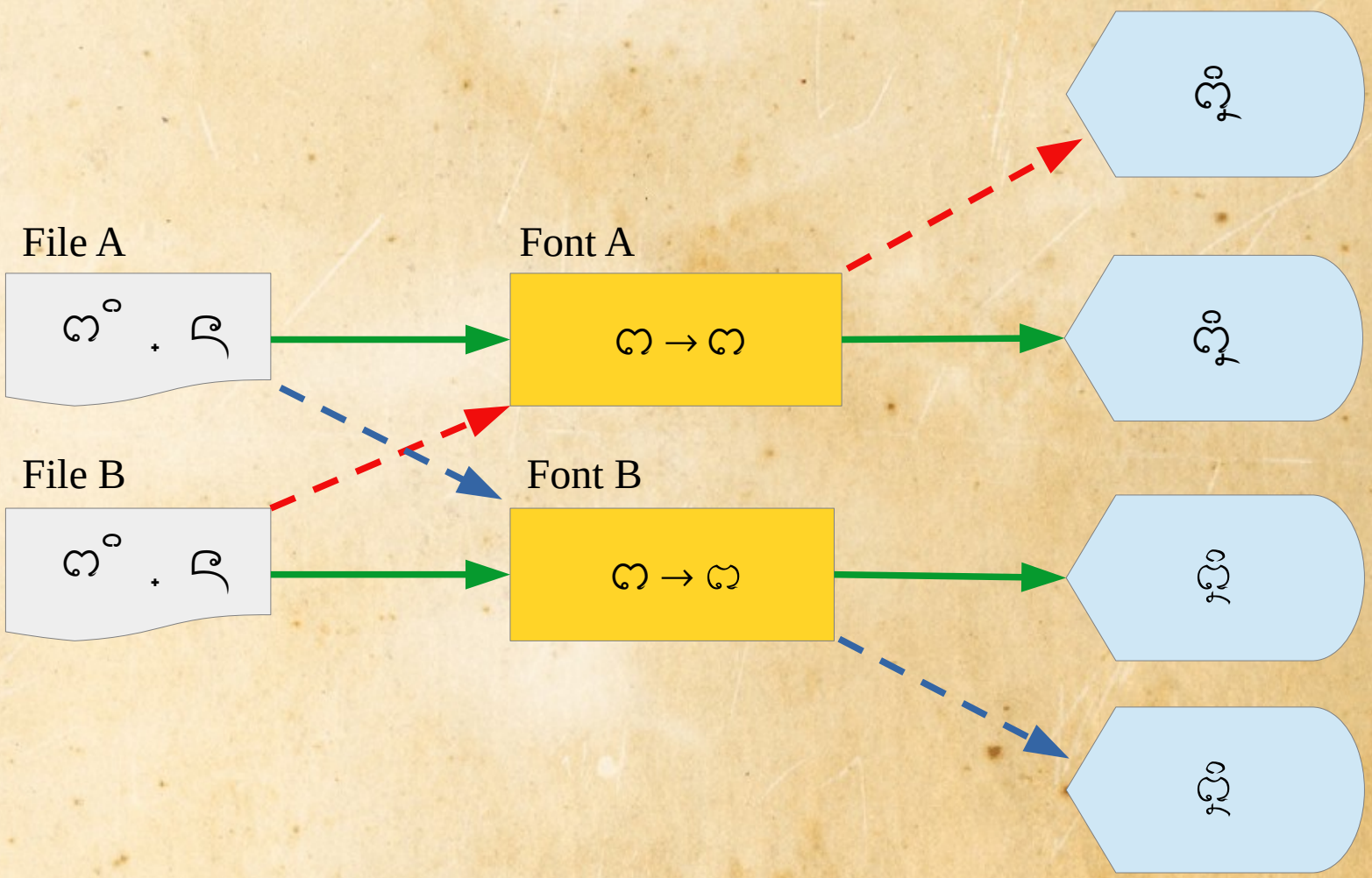
ก่อนยูนิโคด

- ข้อมูลในแฟ้มผูกกับฟอนต์ที่ใช้
- ข้อมูลในแฟ้ม (ทั้ง “ก” และ “d” ฯลฯ) ต่างก็ไม่ใช่อักขระอักษรธรรมหรือไทยน้อย

  → แลกเปลี่ยนข้อมูลกันไม่ได้

  → ประมวลผลไม่ถูกต้อง (ขึ้นอยู่กับอักษรที่ยืมใช้)

เมื่อใช้ยูนิโคด



เมื่อใช้ยูนิโคด

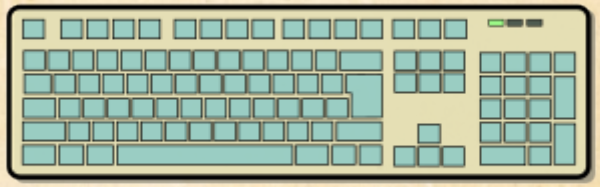
- แฟ้มเดียวกันวาดได้กับทุกฟอนต์

  → แลกเปลี่ยนข้อมูลกันได้

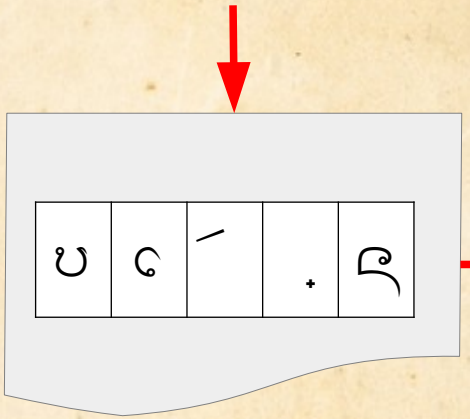
- อักขระทุกตัวเป็นอักขระอักษรธรรม/ไทยน้อย

  → ประมวลผลได้เต็มที่

ส่วนประกอบของระบบ



ระบบป้อนข้อความ
(ยูนิโคด)

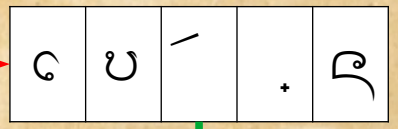


ข้อความ
(ยูนิโคด)

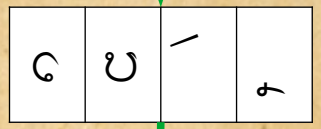
Rendering
Engine

ฟอนต์
(ยูนิโคด)

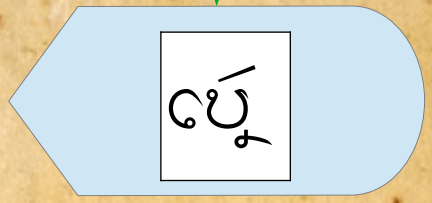
reordering



substitution



positioning



แสดงผล

อักษรธรรม

Timeline ยูนิโคดอักษรธรรม

1992 **Unicode:** สำรวจอักษรไทลื้อ (เชียงใหม่)

2004-2005 **Martin Hosken:** เสนอร่างอักษรล้านนา (Lanna) L2/05-095R

2005 **INALCO (ฝรั่งเศส):** รายงานอักษรธรรมลาว → คล้ายล้านนา

2007 ปรับร่างข้อเสนอล้านนา L2/07-007R

- **จีน:** รายงานอักษรไทลื้อ (สิบสองปันนา)
- **จีน:** ทักท้วงชื่ออักษร Lanna → เปลี่ยนเป็น Tai Tham

2008 ประชุมกลุ่ม Tai Tham → ใช้ฟอนต์อักษรขึน/ลื้อเป็นตัวแทน

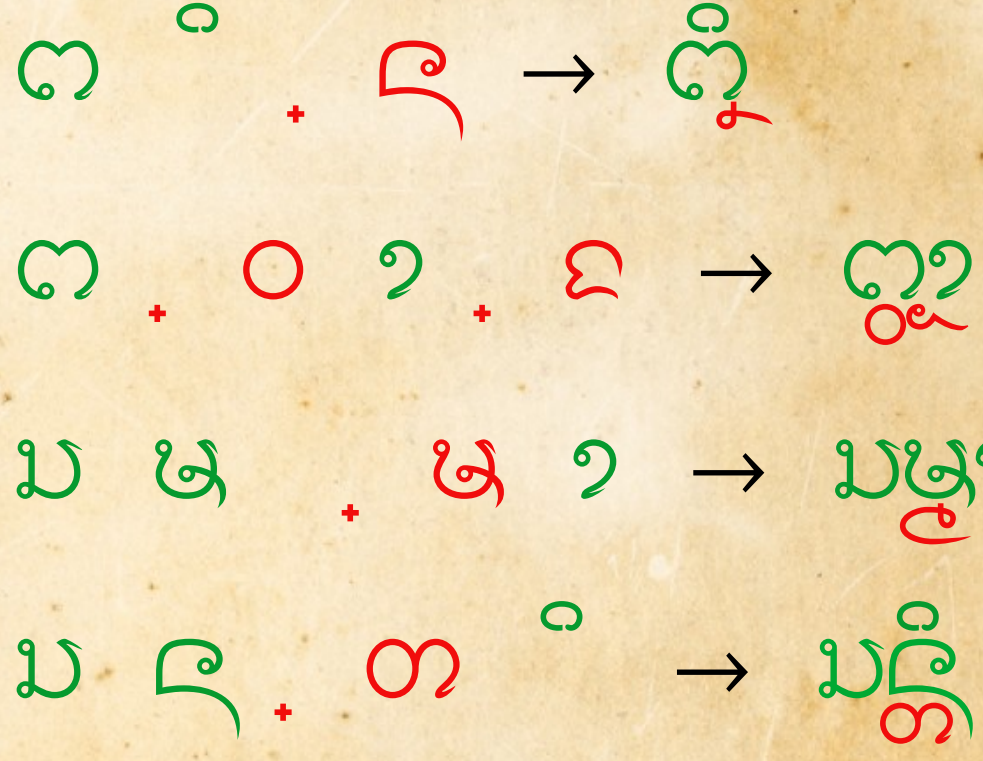
- 3 กลุ่มอักษร: ล้านนา, ขึน-ลื้อ, ลาว-อีสาน

2009 **Tai Tham ใน Unicode 5.2.0** (U+1A20 .. U+1AAF)

2019 **Martin Hosken:** โครงสร้างใหม่ที่เหมาะกับ engine ของไมโครซอฟท์ L2/19-365

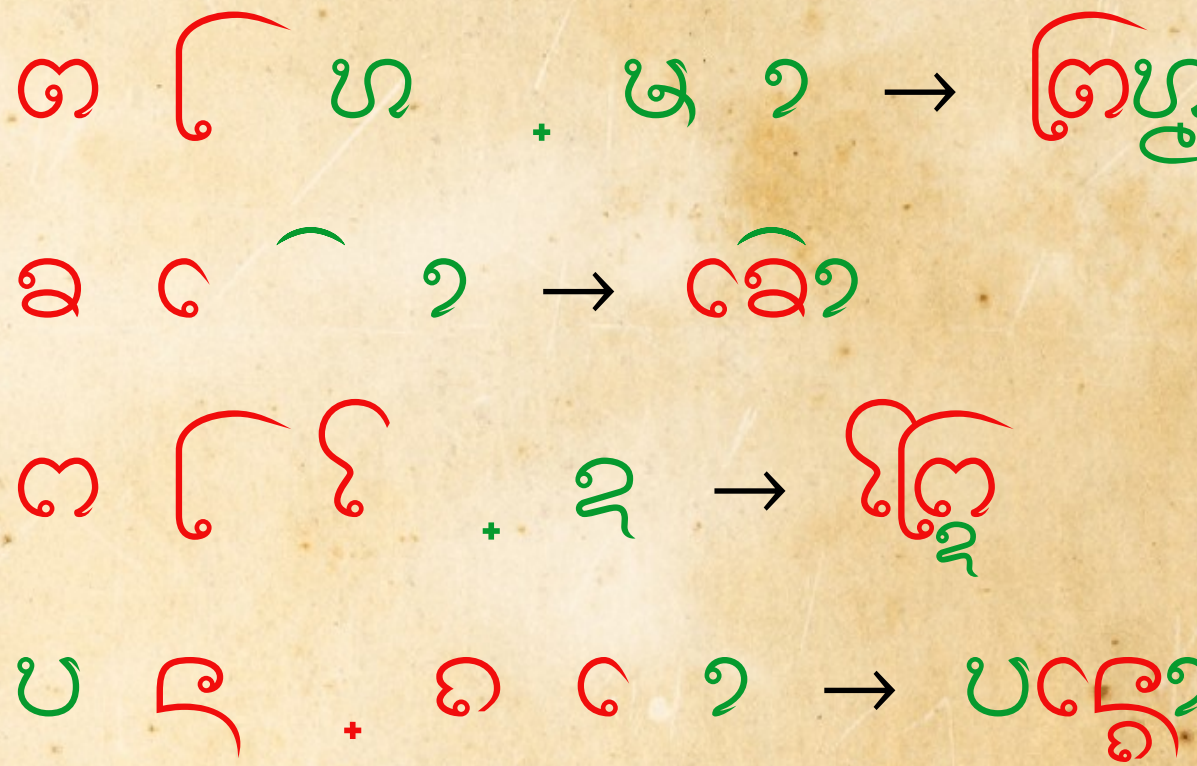
รหัสยูนิโคดอักษรธรรม

- ตัวเชิง ใช้ “วิราม (virama)” แบบอินเดีย (ไทย: พินทุ)



รหัสยูนิโคดอักษรธรรม

- “logical order”: พยัญชนะ ตัวควบ สระ ตัวสะกด



รหัสยูนิโคดอักษรธรรม

- ใช้รหัสรูปแบบเดียว
- สระออล่าง ◌ุ ใช้ U+1A6C ไม่ใช่ ◌ุ + ◌ฺ
- ระวง ◌ั ใช้ U+1A55 ไม่ใช่ ◌ั + ◌ฺ
- สระโอลอย เ◌ว ใช้ U+1A52 ไม่ใช่ เ◌ + ◌ว
- สระอำ ◌ำ ใช้ ◌ + ◌ำ ไม่ใช่ ◌ำ + ◌ และไม่มีสระอำเอกเทศ

รหัสยูนิโคดอักษรธรรม

- อักขระใกล้เคียงกัน
  - เชิง ล ᩠ ใช้ U+1A56; ล เฟื้องข้าง ᩣ᩠ล ใช้ ᩣ + ᩠ล
  - ไม้อังแล่น ᩮ ใช้ U+1A58; ง สะกดบน ᩮ ใช้ U+1A59

ข้อจำกัด ณ ปัจจุบัน

ข้อกำหนด engine ยังไม่ยุติ

- อักขรวิธีซับซ้อน: บาลี, ไทย/ลาว, รูปพิเศษ
- ไม้กังไหล/ไม้อังแล่น: พฤติกรรมหลากหลาย

→ ยังไม่มี engine เฉพาะสำหรับ Tai Tham

การวาดในแอป

Rendering Engines

- MS Office: *Uniscribe*
- LibreOffice, Firefox, Chrome: *Harfbuzz*

การวาดในแอป

Universal Shaping Engine (USE)

- Engine ย่อยของ Uniscribe ของ MS
- ใช้กับอักษรทั่วไปที่ไม่มี engine เฉพาะ (ธรรมด้วย)
- Harfbuzz ก็ทำ engine ที่เข้ากันกับ USE
  โดยมีช่องทางหลบเลี่ยงไปใช้ standard engine
  → ใช้แท็กอักษร 'DFLT' อย่างเดียว

ทางเลือกของฟอนต์

1. ใช้ USE

- มีกลไกใน engine ช่วยทำ reordering
- เอกสารต้องใช้ลำดับอักขระที่ USE ทำงานได้ (ซึ่งขัดสามัญสำนึก)
- ไม้กังไหลไม่ไหล (ไม่ประมวลผลข้ามเซลล์)

ทางเลือกของฟอนต์

2. หลบเลี่ยง USE (เฉพาะ Harfbuzz)

- มีกลไกทำ reordering ให้บ้าง ยกเว้น ระวง
- ไม้กังไหลสามารถไหลได้ (ถ้ามีกฎ)
- วิธีการ: ใช้แท็กอักษร ‘DFLT’ อย่างเดียว
  → ใช้ Standard Shaping Engine แบบ reorder บางส่วน

ทางเลือกของฟอนต์

3. ใช้ Standard Shaping Engine (Windows)

- คล้ายกรณีเลี่ยง USE แต่เป็นภาคบังคับ
- ไม่มี reordering ให้เลย ฟอนต์ต้องทำเองทั้งหมด
- ตัวอย่างแอป: Notepad

กลุ่มของแอป

- **บังคับใช้ USE:** MS Office
- **บังคับใช้ Standard Engine:** Notepad
- **เลือกได้:** LibreOffice, Firefox, Chrome ฯลฯ

เมื่อ MS Office ไม่เข้าพวก

- ถ้า USE รองรับอักษรธรรม → ฟอนต์เรียบง่าย
- ถ้า USE บกพร่อง (ซึ่งขณะนี้เป็นข้อนี้)

  → ฟอนต์ถูกตีกรอบให้แก้ปัญหาได้จำกัด

  → ผู้ใช้ต้องใช้ลำดับอักขระที่ฝืนสามัญสำนึก

  → ต้องรอ MS แก้ปัญหา (ต่างกับโอเพนซอร์ส)

ข้อเสนอแนะ

- ฟอนต์เลือกหลีกเลี่ยง USE ไปก่อน
- จะถึงขั้นรองรับ Notepad ด้วยหรือไม่แล้วแต่ผู้สร้าง
- หลีกเลี่ยง MS Office เมื่อทำงานกับอักษรธรรม
- พิจารณาใช้ LibreOffice

อักษรไทยน้อย

Timeline ยูนิโคดอักษรไทยน้อย

2015 **ราชบัณฑิตยสภา:** ร่าง transliteration อักษรไทยน้อยส่ง ISO

2017 **ISO:** ให้ติดต่อ Unicode เรื่องที่เพิ่มอักษรไทยน้อยในบล็อคอักษรไทย

2018 **ราชบัณฑิตยสภา:** ขอเพิ่มอักษรไทยน้อยในบล็อคอักษรไทย L2/18-041

- **Martin Hosken:** อักษรไทยน้อยเข้ากับอักษรธรรมมากกว่าอักษรไทย
- **Ben Mitchell** (ร่วมกับ Patrick Chew, ผม, อ.ประพันธ์ เอี่ยมวิริยะกุล): ควรเพิ่มอักษรไทยน้อยในบล็อคอักษรลาว L2/18-072
- **Unicode:** แจ้งราชบัณฑิตยสภาพิจารณาเอกสารของ Ben Mitchell

สรุป: ยังไม่กำหนดมาตรฐานในยูนิโคด แต่พอมีแนวโน้ม

รหัสยูนิโคดอักษรไทยน้อย

- ใช้รหัสยูนิโคดอักษรลาวเป็นฐาน

  → ได้อักขระ ◌ิ ◌ั ๅ ๆ ◌ะ ◌ะ โดยปริยาย

  → ใช้ “visual order” เหมือนอักษรไทย

  ◌ิ ◌ั → ◌ั

ตัวเชิงที่ยืมจากอักษรธรรม

◌ฺ◌ฺ◌ ฯลฯ

- [เสนอ] ใช้พินทุลาว (U+0EBA) ตามด้วยอักษรปกติ
  - ◌ฺ + ง → ◌ฺง
  - ◌ฺ + ร → ◌ฺร
  - ◌ฺ + ล → ◌ฺล

ตัวควบ

ඈ ඈ ඇ ඇ ඤ ඥ ඬ ඬ ඳ ඳ (มีอีกไหม?)

ทางเลือก:

1.กำหนดอักขระเอกเทศเหมือน ඹ ඹ

- 👍 สอดคล้องกับ ඹ ඹ
- 👎 ถ้าพบตัวอื่นเพิ่มอีกจะทำอย่างไร?

ตัวควบ

2.ใช้ Zero Width Joiner (ZWJ: U+200D)

ค + ZWJ + ว → ควว

👍 ยืดหยุ่น, ไม่ต้องกำหนดอักขระเพิ่ม

👎 แปลกแยกจาก ฃ ฃ

ตัวควบ

[เสนอ] เลือกใช้ ZWJ

- ไม่ต้องกำหนดอักขระใหม่
- ยืดหยุ่นต่อการพบตัวควบเพิ่ม

อักขระที่ต้องกำหนดเพิ่ม

- สระออย ◌ฺ

  [เสนอ] กำหนด U+0EBE (ถัดจาก ◌ฺ ที่ยังว่าง)

ฟอนต์ไทยน้อย

- เพิ่มกฎ ligature สำหรับตัวเชิง, ตัวควบ
- จัดตำแหน่งการซ้อนสระบน/ล่าง

ขอบคุณครับ