TCL
Thai Computational Linguistics Laboratory

CRL

Development of Thai Encoding and the Implementations

Virach Sornlertlamvanich
Thai Computational Linguistics Laboratory
Communications Research Laboratory (CRL), Thailand.
virach@crl-asia.org

Thailand Open Source Federation (TOSF.org)

International Symposium on Indic Scripts; Past and Future,
Research Institute for the Languages and Cultures of Asia & Africa, Tokyo University of Foreign Studies,
17-19 December 2003

Introduction on Thai (Morphology)

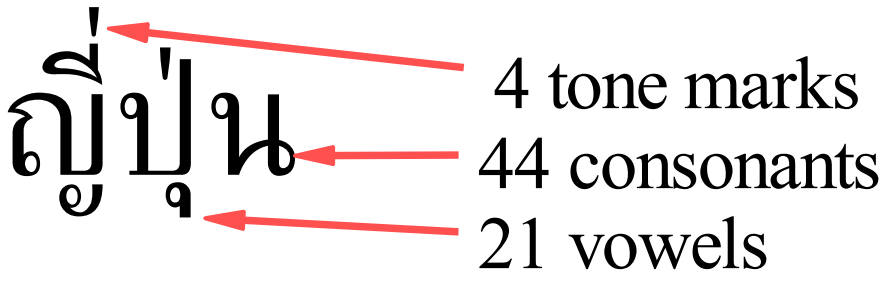
- Paragraph

> วิวัฒนาการทางพันธุวิศวกรรมซึ่งเป็นส่วนหนึ่งของเทคโนโลยีชีวภาพ ได้เจริญรุดหน้าไปอย่างรวดเร็ว จนสามารถทำให้เกิดสิ่งมีชีวิตสายพันธุ์ใหม่ ที่เป็นผลมาจากการตัดต่อยีน ซึ่งเราเรียกเจ้าสิ่งมีชีวิตเหล่านั้นว่าสิ่งมีชีวิตแปลงพันธุ์หรือจีเอ็มโอนั่นเอง ปัจจุบัน ความขัดแย้งทางความคิดเกี่ยวกับจีเอ็มโอยังรุนแรงทั่วโลก การสร้างความเข้าใจในเรื่องนี้จึงมีความสำคัญอย่างยิ่ง

- Alphabetical system
- No word boundary

  Ex: **“GODISNOWHERE”**

  1) God is now here.
  2) God is nowhere.
  3) God is no where.



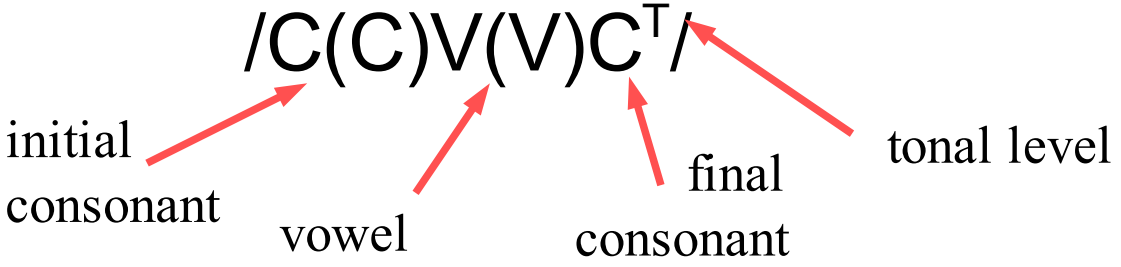
- No explicit sentence marker

Introduction on Thai (Syntax)

- Simple grammar
  - Written and spoken texts are not much different
- Sentence pattern
  - SVO
- No inflection froms for
  - tenses <=> auxiliary verb
  - plurality <=> quantifiers, classifiers, determiners
  - subject-verb agreement
- No syntactic marker
  - word position

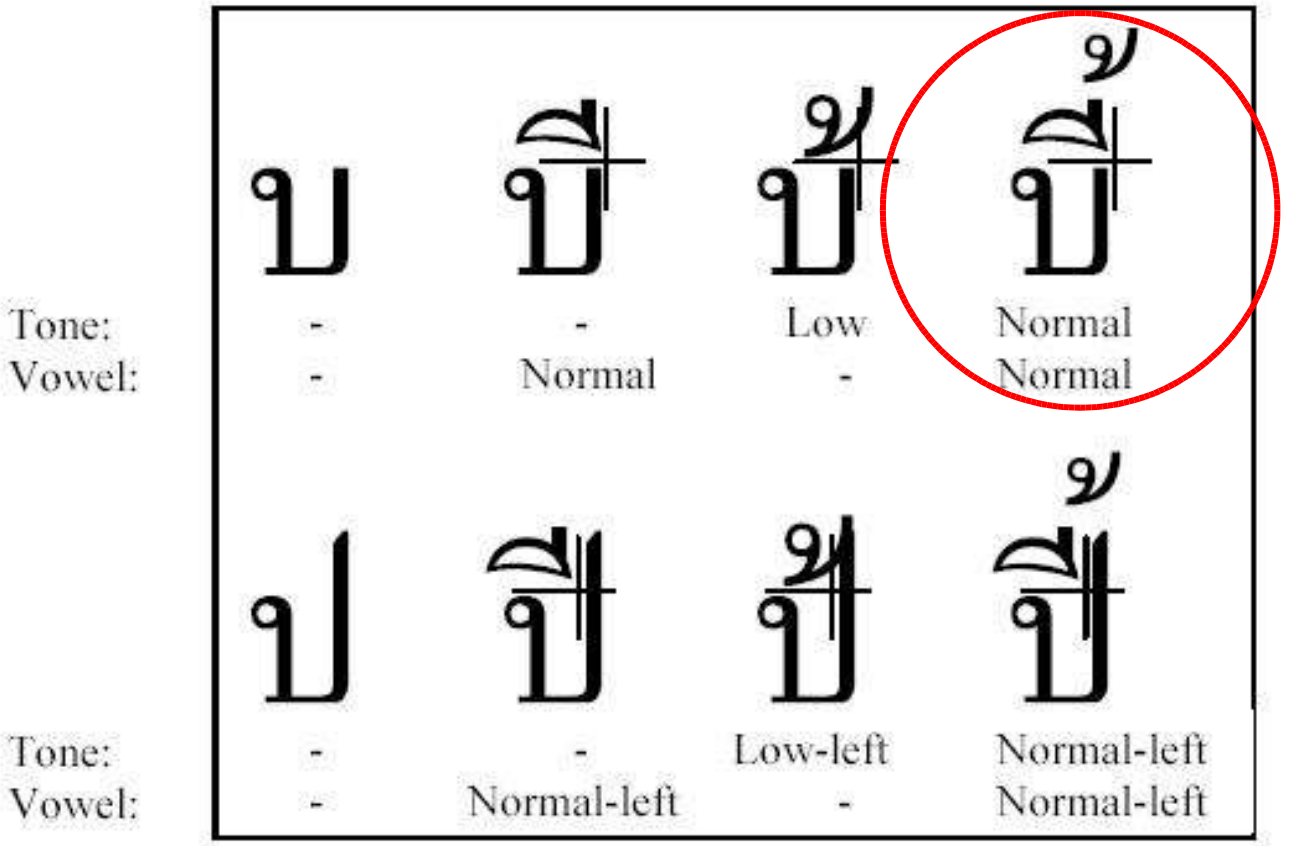
Introduction on Thai (Phonology)

- Syllable



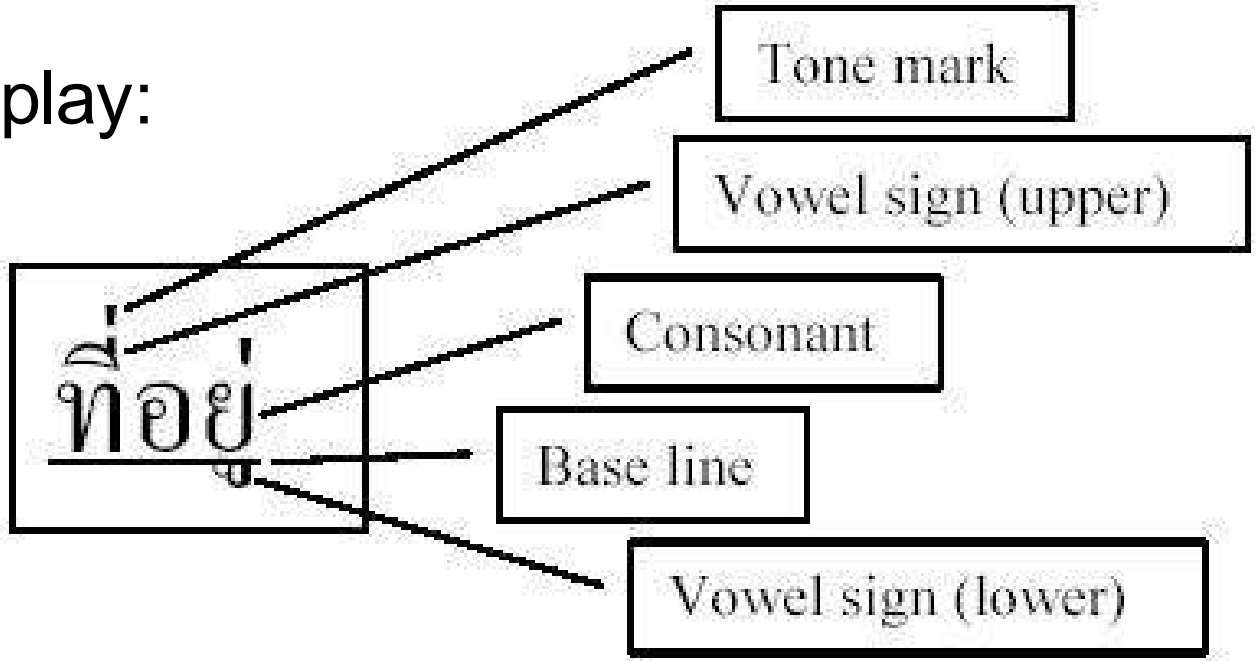
- Different tones convey different meanings
  - /su:aj4/ = beautiful /su:aj0/ = terrible
- No liaison
  - A word has the same pronunciation, no matter where it is.
- No strict pronunciation rule
  - ตุ๊กตา /tuk3/ka1/ta:0/
  - ตุ๊กแก /tuk3/kx:0/

Multiple Glyphs of Vowel Sign and Tone Mark

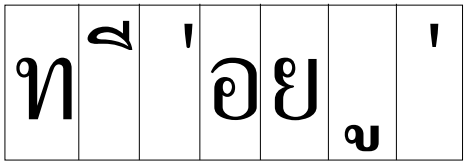


Orthography

Display:

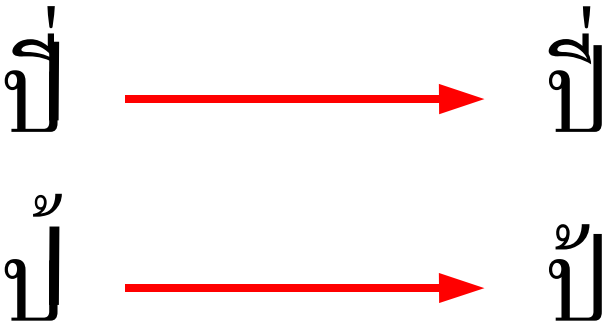


Byte sequence:



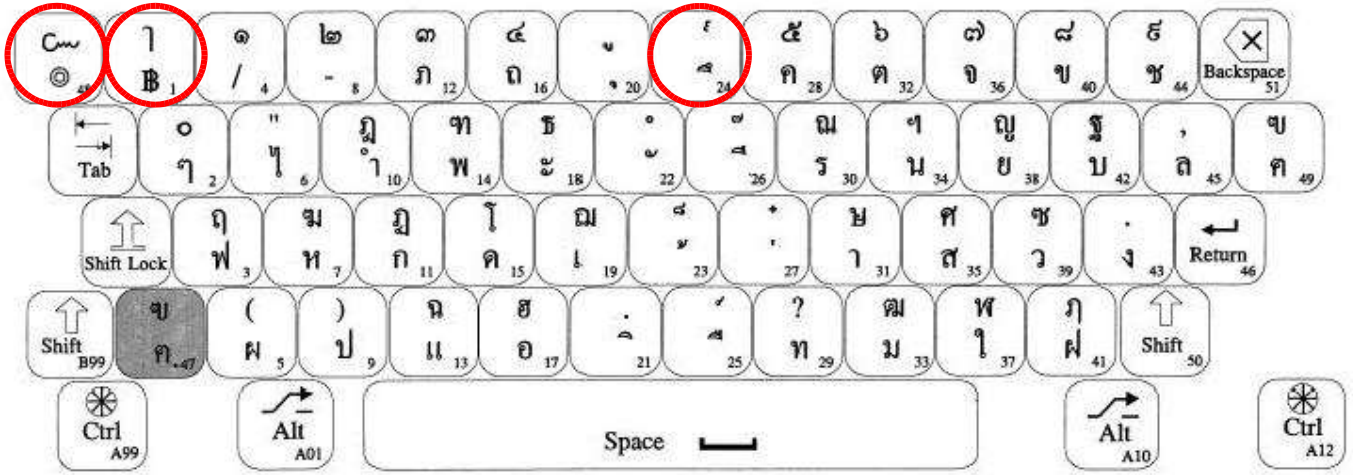
Preprocessing

- Shaping



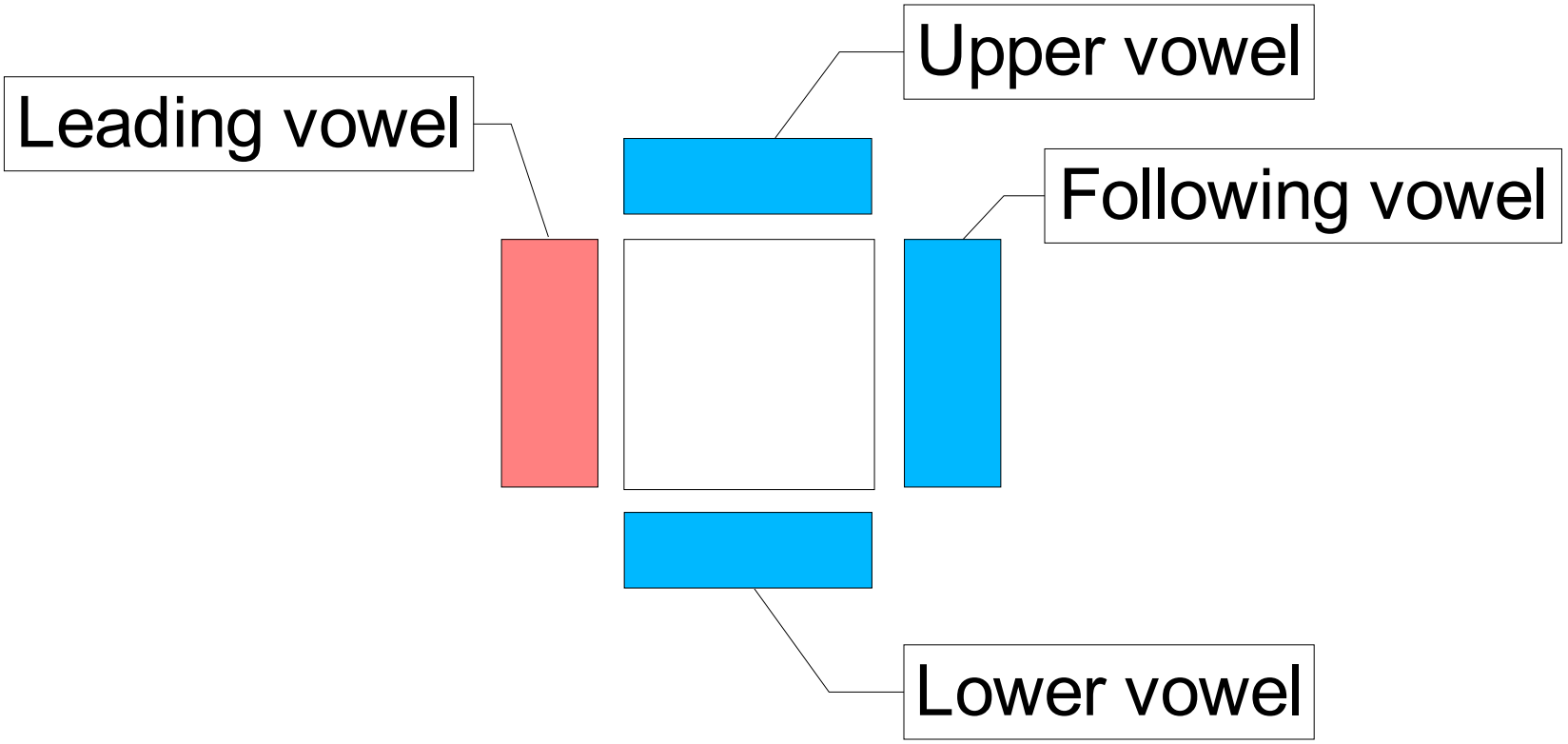
- Normalization

TIS 820-2538 (1995)



De facto standard

Possible Vowel Position



Vowel position

- Above

  กี => ก(k) + ◌ี(ii)

- Under

  กุ => ก(k) + ◌ุ(u)

- After

  กะ => ก(k) + ะ(a)

- Before

  เก => ก(k) + เ(ee)

- Surrounding

  เกา => ก(k) + เ-า(au)

  เกาะ => ก(k) + เ-าะ(oa)

  เกะ => ก(k) + เ-ะ(e)

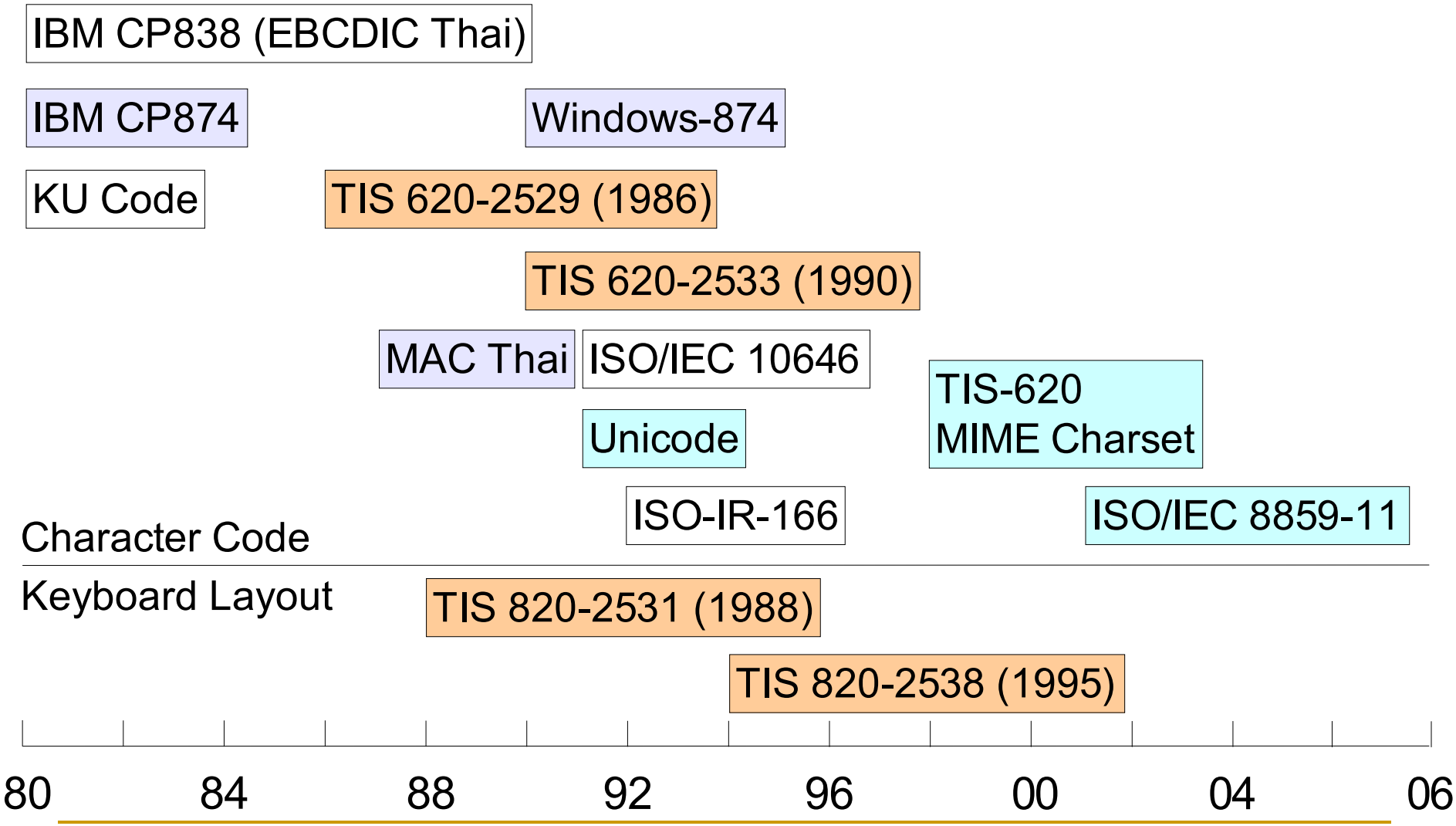
  แกะ => ก(k) +แ-ะ(ea)

Sorting Algorithm

(Theppitak's linux.thai.net/thep/tsort.html)

- For every leading vowel in the string, swap it with the next character.
- Append 2 zero digits to the string.
- Scan the string from left to right. For each of tonal mark, Mai Tai Khoo, and Thantakhat, remove it and append it to the string after the 2 digits representing its original position from the string tail.
- For example,
  - แก่น -> กแน0002 ่
  - อะร้าอร่าม -> อะราอราม0007 ้03 ่

Standardization



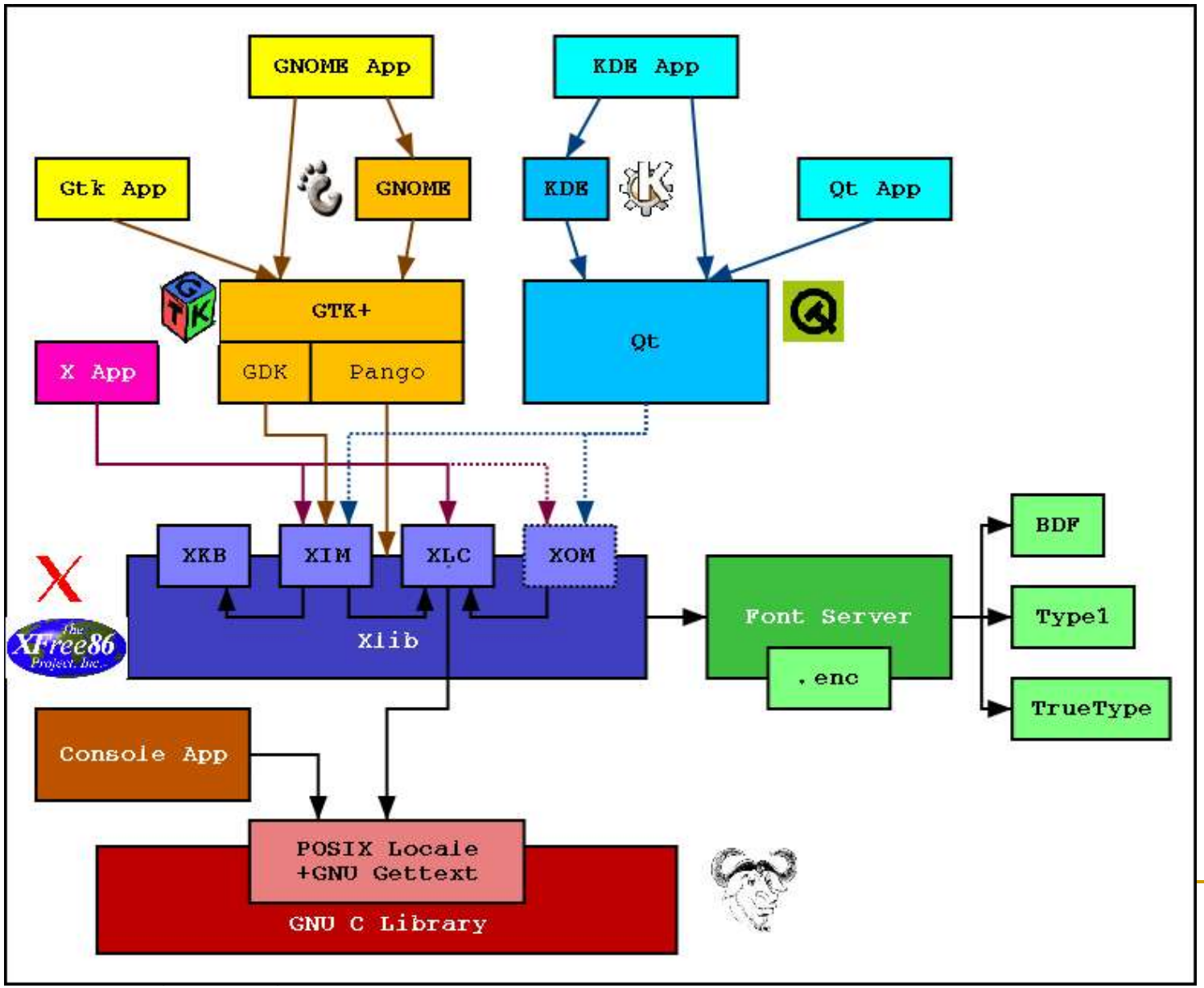
Implementation

- Keyboard, locale, convention
- Vendor standards
  - IBM CP838 (KU Code)
  - IBM CP874 (Extended TIS)
  - Microsoft Windows-874 (Extended TIS)
  - Mac Thai (Extended TIS)
- Current support
  - Data exchange
    - TIS-620
    - Unicode
    - ISO/IEC 8859-11
  - Displaying and printing
    - tis620-0: plain TIS-620
    - tis620-1: Mac Thai
    - tis620-2: Micrsoft Windows-874

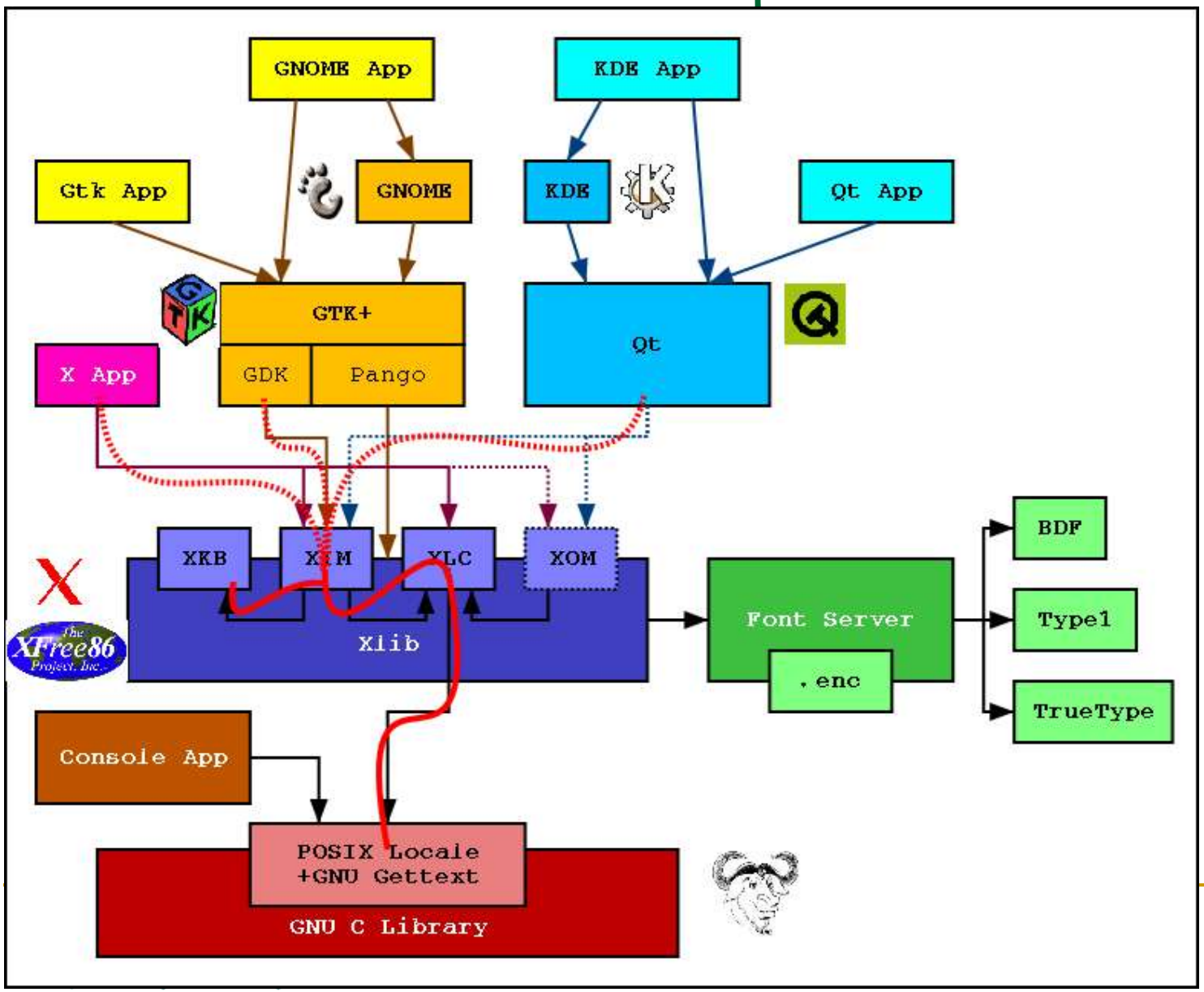
Implementation

- Thai in UNIX/Linux applications
  - Locale: th_TH.TIS-620 is available from glibc 2.1.1
    - LC_COLLATE: sorting
    - LC_CTYPE: character code
    - LC_TIME: calendar
    - LC_MONETARY: unit
    - LC_NUMERIC: number
  - Printing (Fonts + LPRng and CUPS)
- Thai in OpenOffice.org
  - OfficeTLE (word-wrapping + shaping + normalization)
- Thai in Web Browser
  - Mozilla (word-wrapping + shaping + normalization)
- Thai in Database
  - MySQL (sorting)

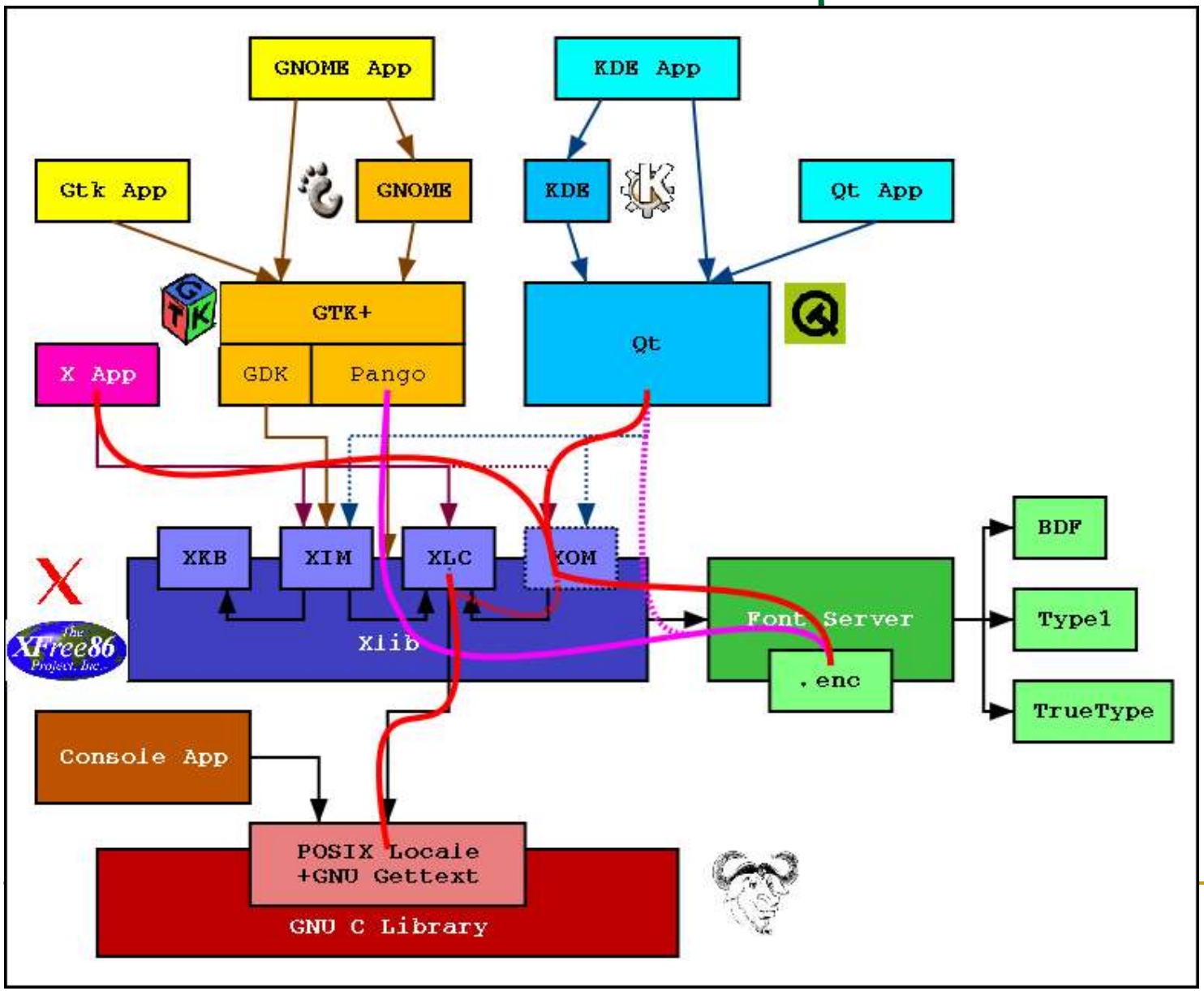
I18N Framework



I18N Framework: Input Method



I18N Framework: Output Method



Defined Charset in Thai Web Pages (.th)

(Oct. 2001, 1310/5272 sites from 8096 domains)

Charset	Site	%	Charset	Site	%
windows-874	682	52.06	**gb2312**	2	0.15
(blank)	519	39.62	**x-user-defined**	1	0.08
tis-620	61	4.66	**windows874**	1	0.08
iso-8859-1	8	0.61	**Thai(tis-620)**	1	0.08
shift_jis	8	0.61	**thai(Windows)**	1	0.08
window-874	6	0.46	**TIS620**	1	0.08
windows-1252	3	0.23	**tis620)**	1	0.08
utf-8	3	0.23	**window**	1	0.08
euc-kr	3	0.23	**windows-128**	1	0.08
iso-8859-11	3	0.23	**windows-847**	1	0.08
x-sjis	2	0.15	**X-MAC-THAI**	1	0.08
			Total	1310	100

Defined Charset in Thai Web Pages (July 2003)

	.th	.com	.net	.org	Total	%
windows-874	5315	3048	223	39	8625	50.55
tis-620	2173	2930	40	25	5168	30.29
iso-8859-1	991	1419	16	38	2464	14.44
utf-8	51	20	1	3	75	0.44
iso-8859-11	2	10	0	0	12	0.07
(blank)	496	182	22	20	720	4.22

Remark

- Font with attaching point concept to resolve the extra-position character problem. OpenType, G raphite, ...
- Standardization and the implementation
  Code set, Glyph, Charset, Input method, ...
- Internationalization and Multilingualization
- Linguistic support.