

## Hierarchical Exception-Marking for Automated Transcription of Thai

Glenn Slayden  
December 4, 2008

### Phonemic transcription (cont.)

- ▶ Romanization: a (phonemic) transcription rendered into a Latin alphabet
- ▶ Minor technical distinction:
  - Although the input to our system is Thai script, we use this input method only as a *proxy* for the pronunciation of a native speaker.
  - The true objective is to generate output which enables a person who understands the target representation format to pronounce the word correctly.

### Overview

- ▶ Phonemic Transcription
- ▶ Thai language: 30 seconds, 30,000 ft.
- ▶ Phonemic Transcription of Thai
- ▶ Issues in Thai orthography
- ▶ History of transcription at [thai-language.com](http://thai-language.com)
- ▶ Exception-notation hierarchy implementation
- ▶ Evaluation
- ▶ Future Work
- ▶ Live Web Demonstration

### Thai language in 30 seconds

- ▶ Monosyllabic language
- ▶ Subject-Verb-Object word order
- ▶ 5 spoken tones: low, mid, rising, high, falling
- ▶ Indic alphabetic script derived from Khmer
- ▶ “44” consonant glyphs; “32” vowel glyphs (pre- & post-posed, super- & sub-script), 4 superscript tonal markers
- ▶ Left to right; no spaces between written words
- ▶ Morphology: no inflection, minimal derivation
- ▶ Particles mark tense, mode, formality, aspect, etc.

### Phonemic transcription

- ▶ Rendering the semantically significant phonology of a word or phrase in a (typically) non-native writing system
  - Phonetic features which do not convey meaning are not rendered. IPA is typically not used.
  - Silent letters are not rendered...
    - หนู            /nú:/            “mouse”
  - ...and inherent vowels must be rendered
    - ถนน        /tʰàñǎn/        “road”
  - Pali geminates became unwritten linking syllables
    - ผลไม้        /pʰǎnlámá:j/        “fruit”

### Phonemic transcription of Thai

- ▶ There is no de-facto or consensus standard transcription system for Thai.
- ▶ Proliferation of transcription schemes in published references. At least 14 in my bookshelf.
- ▶ RTGS, AUA, ALA-LC, Haas, FSI, Paiboon, Bua Luang, t-l Enhanced, Paiboon, Smythe, FSI, Kuo,...

## Phonemic transcription of Thai

- ▶ A credible system must represent tones
  - otherwise, we get...
    - ไหมไหมไหม ? “mule, mule, mule”
    - mai mai mai ? “mule, mule, mule”
    - /mǎj mǎj mǎj/ “New silk burns.”
- But the Thai government-sanctioned “Royal Thai General System” (RTGS) and many others do not indicate tones.

phak mi khwam samkhan kae ahan chin pen yang ying  
/pʰàk mi: kʰwɑ:m sǎmkʰan kǎe: ʔa:hǎ:n tɕi:n pen jà:ŋ jǐŋ/  
“Vegetables are very important in Chinese food.”

## Pitfalls of phonemic transcription

- ▶ Source language dialects, registers, and regional variation
  - The field generally posits four dialects (central, northern, southern, and Isaan) and five registers (street, elegant, rhetorical, religious, and royal)<sup>1</sup>
- ▶ Result depends on the reader’s dialect and pronunciation of the target language
  - Pronunciations of English: American, British, Canadian, Australian, etc.
  - This is a true problem for an Internet resource aimed at an international audience of non-linguist travelers.

1. wikipedia

## Phonemic transcription of Thai

- ▶ A credible system must render vowel duration:

- ข้าวขาว “white rice”

- /kʰā:w kʰā:w/

? khaw khaw ? khaaw khaaw

- /kʰǎw kʰǎw/

- เขาเข้า “He goes in.”

## Pitfalls of phonemic transcription

- ▶ Gives the layperson false confidence
  - Unintelligible result
    - ฉันไปภูเก็ต → “I’m going to Phuket.”
    - chan pay phuket → ?\*/tʰæn pei fɔkɪt/
    - /tɕʰǎn paj pʰu:kèt/
  - or worse: unintentional malapropism (especially in tonal languages like Thai)
    - คนขายตั๋วอยู่ที่ไหน “Where is the ticket seller?”
    - /khon khai tua yu thi nai/
    - /kʰon kʰǎj tu:a jù: tʰi: nǎj/ (should be /tʰi:a/)
    - คนขายตั๋วอยู่ที่ไหน “Where are the prostitutes?”<sup>1</sup>

(1. www.learningthai.com)

## Pitfalls of phonemic transcription

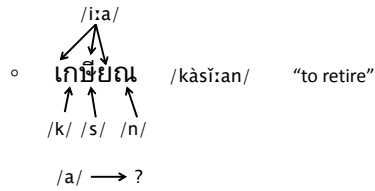
- ▶ Mutually-exclusive phoneme sets between languages
  - งู /ŋu:/ “snake”
  - เฝื้อง /pʰen/ “full moon”
  - เป็น /pen/ “to be”
- ▶ Original Thai spelling cannot be recovered
- ▶ Geographical and proper names have widely accepted transcriptions which may adhere to no particular system
  - ราชดำริ Rajdumri, Rajadamri, Ratchadamri, etc.

## Alternatives to phonemic transcription/Romanization for language students

- ▶ Learn the language from a native speaker!
  - extra-credit: learn Thai script
- ▶ Use audio clips of a native speaker
  - thai-language.com has 13,135 (34% coverage)
- ▶ IPA?
  - [narrow], /broad/
- ▶ Phonemic Thai transcription (upcoming slide)
  - Avoids some of the pitfalls of phonemic transcription

### Orthographic issues in phonemic transcription of Thai

► Preposed vowels



### Example of an exception rule not covered by the tone rule summary

Clustered consonant tone rule:

- In most clusters of phonologically incompatible initial consonants (“non-conforming initial consonant clusters”):
  - the first consonant is mid- or high-class, **and**
  - the second consonant is a sonorant.
- In such cases, the cluster follows the same rule as all clusters; that is, the spoken tone of the syllable as a whole is determined according to the consonant class of the **first** consonant in the cluster.

However, in the case that the second consonant in the cluster is not a sonorant (non-sonorant low-class consonants included) then that **second** consonant is used to determine the spoken tone of the syllable.

### Orthographic issues in phonemic transcription of Thai

► Tone determination

- Syllable tone is determined by:
  1. Consonant ‘class’ of initial consonant (i.e. sonorant, aspirate, plain)
  2. Presence of written tone mark
  3. Long or short duration vowel in syllable
  4. Dead or live syllable ending
  5. Open or closed syllable

บ้าน /bâ:n/ “house”

### Orthographic issues (cont.): Consonant clusters

ถนน /t<sup>h</sup>ànǎn/ “road”

- In native Thai sensibility, this is a monosyllabic word with:
  - initial consonant cluster
  - sub-syllable inherent (unwritten) /-a/

► Ambiguous initial clusters

ปริมาณ /pàrī?ma:n/ “amount”

- Note: for purposes of phonemic transcription, whether it’s considered one “syllable” or two doesn’t matter.
  - (The new system is currently stuck with the “wrong” design)

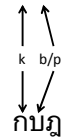
### Orthographic issues in phonemic transcription - tone rule summary:

initial consonant class	no tone mark		tone mark	
	live syllable	dead syllable	live syllable	dead syllable
low sonorant:	M	H	F	F H
mid:	M	L	L	F H R
high aspirate:	R	L	L	F

### Orthographic issues (cont.): “Inherent” (unwritten) vowels

► Unwritten ใ-ะ /-o-/

กบ /kòp/ “frog”



กบฏ /kàbòt/ “rebellion”



## Why wasn't the FST good enough for this task?

- ▶ Exceptions in Thai pronunciation
  - Many driven by etymology (Pali, Sanskrit)
  - Combining forms
  - Ambiguous consonant clustering
  - Vowel duration irregularities in common words
  - Established colloquial pronunciations
  - Inability to represent multiple pronunciations corresponding to register, formality, or acceptable stylistic variation by speaker
- ▶ It became clear that the FST would benefit from *guidance* in handling the wide variety of exceptions

## How the FST handles inherent vowels

- ▶ Up to 3 passes through FST, if required
- ▶ System accepts the first parse of the Thai script stream which completes successfully (i.e. on a legitimate syllable boundary)
  - 1<sup>st</sup> pass: inherent vowels disabled
  - 2<sup>nd</sup> pass: inherent /-oh-/ enabled
  - 3<sup>rd</sup> pass: inherent /-a/ and /-oh-/ enabled
  - Still no parse? Fail.

## What did it do well?

- ▶ Empirically proved against entire 38,000 word lexicon
  - Each entry in the lexicon yielded some result (termination)
    - (FST proves the lexicon)
  - FST was developed until lexicon was satisfied
    - (Lexicon proves the FST)
    - This was a significant result in itself: a comprehensive table of 73 vowel and diphthong types in the Thai language which is minimal against 38,000 word lexicon <http://www.thai-language.com/ref/vowels>
    - Example:
      - เ-อว in the rare word ไก่กว /kɔ:w/

## Building on these strengths

- ▶ Enable volunteer editors to mark exceptions without programming
- ▶ Provide a range of marking options for capturing different classes of exceptions
- ▶ Allow alternate transcription output formats
  - This was a common user request, especially for IPA
  - How to handle existing 'hard override' transcriptions?
- ▶ Maximize lexicon normalization

## What did it do well? (cont.)

- ▶ Identify complex multi-grapheme diphthong patterns
- ▶ No programming required when enlarging the lexicon (unsupervised results for new Thai entries)
- ▶ Broad classes of exception and irregularity could be identified and the best candidates programmed
- ▶ There was provision for a 'hard override'

## Phonemic Thai output

- ▶ All these factors were pointing towards an overhaul of the system
- ▶ But then, one more obvious benefit that had previously been overlooked was recognized:
  - When allowing for multiple output formats, one of the new *output* formats could be phonemic Thai!

สวัสดีครับ → สะ-หวัด-ดี-ครับ "Hello."

- ▶ *This feature was motivating for the overhaul*

## An aside:

- ▶ In (naïve) theory, phonemic transcription can be performed syllable-by-syllable, without regard for word constituency
- ▶ In practice, this isn't adequate:
  - Etymology drives exceptions
  - Loanwords have special considerations
- ▶ An ideal system can process an arbitrary stream of script, but *appreciates hints from the lexicon*
- ▶ All this this leads us to...

## Transcription assistance bit flags

- ▶ 32 flags defined, they indicate:
  - Applicability of certain rules available in the FST
    - i.e. clustered consonant tone rule shown earlier
  - General etymological class (loanword, Pali, etc.)
  - Consonant clustering
    - Presence of specific consonant clusters
    - Enabling/disabling of clusters
  - Forcing of short or long vowel
  - Silent letter disambiguation
  - Miscellaneous
- ▶ Currently used for 953 cases (2.5%)

## A lexicalist approach to exception marking

- ▶ Each entry in the lexicon may have:
  - Inhibit transcription flag
    - e.g. for abbreviations
  - One or more bit flags which "assist" the FST engine
  - Inline syllabification assistance
  - Inline indication of linking syllables
  - Manual "hard override" – now specified in phonemic Thai!
- ▶ For compounds:
  - Transcription selection index for each constituent
  - "Use full transcription" flag (infrequently used)

## Inline syllabification assistance

- ▶ Special inline character • may be inserted in the Thai script of the lexical entry; it is never displayed on the website

Thai word	Unassisted result	Assisted entry	Assisted result	Meaning
กรุณา	krù'ŋka:	ก•รณา	kàr'ŋka:	please
ไหม	k'ò:m	ไช•ม•	k'ò:má	white linen
เทา	t'he:w	เท•ว•	t'he:wá	[Pali] double
เนา	?	น•ว•	náwá	[Pali] new
บุคคลากร	bù'kk'há:ko:n	บค•ค•ลา•กร	bù'kk'há:ko:n	personnel
วีร	wi:n	วี•ร•	wi:rá	[Pali] brave
ไฮโดรคลอริก	hòjdrò:klo:rík	ไฮ•ดร•คลอ•ริก	hòjdro:k'hlo:rík	hydrochloric

## Preferred hierarchy of exception marking

Each lexical entry has independent provision for:

1. Unmarked lexical entry
  - Consider modification to FST engine
2. Set one or more assistance flag bits
3. Inline syllabification assistance
4. Phonemic Thai Hard override

## Inline syllabification assistance

- ▶ Special character • may be doubled •• to indicate unwritten linking syllable (Pali geminate)

กฤษฎีกา	kritdi:ka:	กฤษ•ฎีกา	kritsàdi:ka:	decree
ชินวัตร	tʂ'hinwát	ชิน••วัตร•	tʂ'hinnáwátrá	Shinawatra
ธรรมดา	t'hamda:	ธรรม••คา	t'hammáda:	ordinary
พรหม	pho:nhóm	พรหม••	phommá	Brahma
มฤคทายวัน	m'rú'k'há:jwan	ม•ฤค••ทาย••วัน	má'rú'kk'há'thá:jjáwan	Marukathayawan
ศาสนา	sà:t'na:	ศาสน••นา	sà:tsáná:	religion

- ▶ •• and • currently used for 892 cases (2.3%)

## Phonemic Thai hard override

- ▶ As a last resort, free-form exceptions can be entered *in phonemic Thai script*
- ▶ This change allowed a modular architecture for multiple output formats (upcoming slide)
- ▶ Used to handle digits, abbreviations, truly exceptional or complicated cases, irregular tone, etc.
  - วร (วอ-ระ) \*/wɔːn/ /wɔːráʔ/ “benefit”
  - 555 (ห้าห้าห้า) /hâ:hâ:hâ:/ “Ha, ha, ha!”
  - Currently used in 215 cases (0.56%)

## Multiple output formats

- ▶ The FST was reengineered to manipulate phonemic tokens on a stack, rather than generating a transcription string directly.
- ▶ Output “modules” render the completed phonemic token stack into the desired output format
- ▶ Insight: phonemic Thai “hard-overrides” are “free” for all other output formats by treating them as canonically-spelled *input* to the FST!

Partial screenshot of editing tool showing some exception marking features for a lexical entry

## IPA

- ▶ IPA output is potentially phonetic [narrow], not just phonemic /broad/
- ▶ Although our token stack is largely phonemic, the FST does emit a glottal stop token
  - emitted for silent อ
  - emitted after short vowel, open syllables except inherent -๕
  - not used by other output modules (except AUA)
- ▶ Current area of investigation: characterizing glottal stop from Thai script input.

## Multiple output formats

- ▶ The system now outputs in multiple user-selectable transcription formats:

t-l Enhanced	thai-language.com Phonemic Transcription	saːˈwatː dee <sup>M</sup> khráp <sup>M</sup>
Phonemic Thai	Phonemic spelling in Thai script	สะ-หวัด-ดี-ครับ
IPA	International Phonetic Alphabet (broad with phonemic tones)	/sàwàt diː kʰràp/
RTGS	Royal Thai General System	sawat di khrap
AUA	American University Alumni/Peace Corps	sàwàt dii khráp
ALA-LC	American Language Association-Library of Congress	sawat dī khrap
ISO 11940	An International transliteration standard	Ṣwàṣ dī khráp
Bua Luang	“What You See is What You Say” (E.G. Allyn, et al.)	sáˈwàtː dee khrápː

(สวัสดีครับ “Hello.”)

## Rearchitecting a fragile system

- ▶ Comparison report feature is essential to preserving carefully tuned results when making changes to the FST engine
- ▶ Essentially a unit-test for the transcription of every word in the lexicon.
- ▶ HTML report is a diff of the “before” and “after” snapshots

## Evaluation

- ▶ Designating a gold-standard is a huge problem rooted in many of the inherent pitfalls of phonemic transcription and prescriptivism.
- ▶ For lack of a more credible alternative, we resort to the Royal Institute of Thailand and their 1982 dictionary.
  - machine-readability problems
  - partial phonetic rendering (only the “hard” part)
  - because of these issues, only 33.1% of the gold-standard is directly usable without further processing (11730 of 38572 entries)

## Demonstration

reverse phonemic transcription

Q & A

## Evaluation

- ▶ supervised corrections distribution
- ▶ N = 11727 (usable gold-std intersection with 38572 entries)

Hard	Flags	Correct	Errors	Precision
N	N	10123	1604	0.841
N	Y	10674	1053	0.901
Y	N	10887	840	0.922
Y	Y	11569	158	0.986

## Future Work

- ▶ Characterization of glottal stop
- ▶ Push Pali generalizations into FST
- ▶ FST token stack mechanism is fragile
- ▶ Cache token result rather than multiple output formats
- ▶ Engine is TIS-620 input, Unicode output
  - Convert to all-Unicode
- ▶ Web 2.0
  - Web users as knowledge resource. Need system for maintaining standards/ensuring credibility