

# Thai in Transition *and the Thai* Gigaword/Terabyte Web Corpus

Rikker Dockum and Doug Cooper  
Center for Research in Computational Linguistics  
CRCL Inc is a US 501(c)3 nonprofit



<http://sealang.net/archives/ri-tera-2008.pdf>

# Today's agenda

- Thai in transition
- Trying to influence the change
- An observation tool: the corpus
- Thai corpus projects
- A Web-based corpus
- In conclusion



**SEALang Web Corpus**

100 results

any date

match phrase

don't sort results

visit all available sites

**Yahoo** **Google**

**Search** **Sample** **Analyze**

**Define** **Predict** **news2000** **Churn 0**

แอบแป้ว

New Words (Royal Inst. 2007)

New Words (Royal Inst. 2007)

Foreign Words (Royal Inst. 2006)

Slang (Min. of Ed. 2000)

Thai Slang (Jintana 2003)

Merged list

กํกํ  
กํ  
กํกํกํกํ

Title Bevariety Forum :: อ่าน - วิธีการถ่ายรูปแก้ว  
Url <http://www.thaifooddirect.com/...> Size:  
Date: (UTC) 6/30/2008

# Thai in transition

- **Types of change:**
  - lexical (new words)
  - grammatical / syntactic  
(how they fit together)
  - phonemic (how they sound)
- **A constant state of affairs**
- **A natural state of affairs**



# Influences include

- **Indo-European:**
  - Sanskrit, Pali, English
- **Austroasiatic:**
  - Khmer, Mon
- **Sino-Tibetan:**
  - Chinese, Chinese, Chinese ...



# Can language be planned?

- **King Mongkut**
  - ‘linguistic’ edicts
  - support for Pallegoix
- **The Royal Institute**
  - ‘keeper of the language’
  - sometimes
    - leading the charge
    - holding back the tide





# Can change be controlled?

- **the RI Word Coining Committee**
  - Prince Wan, chief establishment spokesman
- **the word-coining public**
  - Prince Wan, chief anti-establishment spokesman!

“[my teacher] taught me that in Thai we can mix Pali and Sanskrit together, so I need not write *patikamm* in Pali fashion but could write *patikarm*, half Pali, half Sanskrit.”

“The sound and rhythm of a word are most important...”  
*Prince Wan*



# Can change be observed?

- **A text corpus:**
  - a substantial collection of text
  - (usually) a set of search tools
  - may be *genre*, *balanced*, or *open*
  - shows the living language for descriptive applications
  - shows the ‘approved’ language for prescriptive applications

“You can observe a lot  
just by watching”  
*Yogi Berra*





# Some observations worth making

- do foreign loans have any impact on Thai syntax or grammar?
- what semantic and phonological nativization processes 'Thai-ize' loans?
- how effective is the Royal Institute's prescriptive orthography for slang and loans?
- are there typical phonological processes that tend to generate new slang?
- does the distribution of slang characterize Thai Web space, or the slang itself?
- can dictionaries developed without text corpora hope to get it right?
- what semantic niches do foreign borrowings fill?
- what is the grammatical distribution of new and borrowed words?
- how widely used are suggested 'traditional Thai' alternatives?



# Typical Thai corpus projects

- **Thai National Corpus**
  - <http://www.arts.chula.ac.th/~ling/TNC/>
- **Orchid Corpus**
  - <http://www.links.nectec.or.th/orchid/>
- **Chulalongkorn Corpus**
  - <http://www.arts.chula.ac.th/~ling/ThaiConc/>
- **SEAlang Thai Corpus**
  - <http://sealang.net/thai/corpus.htm>



**Varied goals, design, content**

# An atypical corpus: Thai Epigraphy

**Corpus of Sukhothai Inscriptions**  
Maintained by Doug Cooper (bugs to [doug@th.net](mailto:doug@th.net))  
Center for Research in Computational Linguistics, Bangkok <http://crcl.th.net>

[\[popup\] INSTRUCTIONS](#)   [MAIN](#)   [DETAILS](#)   [PROJECTS](#)   [IMPLEMENTATION ISSUES](#)   [BIBLIOGRAPHY](#)

Details of the inscriptions by number and date:  [More details](#) [Database stats](#)

Enter inscription number(s)   All by ☒ number   ☐ date   ☐ custom  

1 2 3 5 7 8 9 10 11 13 14 15 16 37 38 40 45 46 49 52 54 62 63 64 86  
90 93 94 95 98 102 106 107 108 1000

Lines ☐ raw   ☒ segmented   Highlight ☒ tags   ☐ unique   ☐ first

---

  Thai search

☒ segmented (no wildcards): display context =  words   ☒ Image if available

☐ unsegmented; display context =  chars, near =  chars.

Wildcards: (a**b** or a or b) (a~b or a near b) (~~ or near near for reverse order)

---

For 3 and 4 ( ☐ show only   ☒ exclude only ) these items: (click   )

☒ cmpds   ☒ names   ☒ places   ☒ titles   ☒ elabs   ☒ unsures   ☒ frags

   columns,   ☐ order by length   ☐ mark '?'s

☒ Label (eg. 107:1.26)   ☐ list, with:   ☒ word counts   ☐ inscription #'s

---

  Show ( ☒ 1st   ☒ 2nd   ☒ both ) sets

Compare entries between inscription sets, eg. 90 107 X 3 7. "X" can be any letter.



# SEAlang Thai Epigraphy Project

Inspection order: 1 107 1000 90 11 3 2 5 7 8 62 102 106 94 64 45 38 95 93 46 10 9 40 49 52 16 63 37 13 98 1  
226 entries.

Study line

Highlight or doubleclick any I:F.L (inscription:face:line) number, then click here to see it in context.

Study word

Highlight or doubleclick any word, then click here to see each line it occurs on. The word may be a compound (but only spaces).

กุศลผลบุญ-E 15:1.33  
บ่ฆ่าบ่ตี-E 1:1.31 5:1.22  
บ่ชอบบ่ยำ-E 3:1.23  
มีไรมีนา-E 1:2.34 1:3.3  
ใครใจในใจ-E 1:4.10  
จงคงตรงต่อ-E 15:3.7  
ในน้ำในถ้ำ-E 45:1.2  
ภายนอกในใจ-E 40:1.17  
40:2.6  
มาคว้ามลาก-E 38:2.46  
มานบมาเห็น-E 9:2.5  
ได้คาดได้พบ-E 45:1.21  
ทำเนื้อทำตน-E 3:2.23  
ในหล้าฟ้าใส-E 15:1.26  
ประกิดชิดชน-E 2:2.34  
ไปจอดไปกราย-E 38:1.41

ผิดแพกแสกว้าง-E 1:1.25  
ผู้เฒ่าผู้แก่-E 2:1.52 14:2.15  
ผู้เฒ่าผู้แก่-E 106:2.36  
ผู้รู้ผู้เห็น-E 38:2.30  
ผู้ใหญ่ผู้ราม-E 40:1.2  
พ่อแม่พี่น้อง-E 106:1.24  
พ่อลูกพี่น้อง-E 7:4.20  
พายุลุนปุนหลัง-E 2:1.40  
ไพร่ฟ้าข้าไทย-E 1:1.23 3:2.32  
3:2.43 5:1.16  
ไพร่ฟ้าหน้าปก-E 1:1.33  
ไพร่ฟ้าหน้าใส-E 1:1.21 1:1.6  
พายุลุนปุนหลัง-E 11:1.17  
มั่นคงตรงอยู่-E 40:1.10  
มีคมนท่อนธการ-E 40:1.23  
เยียดคเยียด-E 5:1.24

หมอนนั่งหมอนนอน-E 1:2.15  
หมากส้มหมากหวาน-E 1:1.13  
หินแลงแปลงเลียด-E 62:2.9  
อันเหง้าเจ้าไทย-E 106:1.25  
คุณในปลาในข้าว-E 64:1.2  
ใจรักภักดีไมตรี-E 64:1.9  
ช้อยปลดช้อยมล้าง-E 64:1.11  
ด้วยเกล้าด้วยหาญ-E 1:4.15 3:2.5  
ด้วยรู้ด้วยหลัก-E 1:4.15  
ด้านใต้ด้านเหนือ-E 62:2.15  
ทั้งสิ้นทั้งหลาย-E 1:2.11  
บ่มีเงื่อนบ่มีทอง-E 1:1.30  
บ่ให้จับบ่ให้หาย-E 2:1.49  
บ่ให้ลืบบ่ให้ตาย-E 2:1.50  
บ้านใหญ่บ้านเล็ก-E 1:2.35 1:3.3  
พาทยพิณเตรสังข์-E 102:1.36

# Comparing modern corpora

- **Thai National Corpus**
  - follows British National Corpus criteria
  - specifies domain, time, medium
  - Internet less than 5%
  - fixed and repeatable
  - hand segmented and tagged
  - prescriptive – quality, not size
- **SEAlang Thai Corpus / WebCorpus**
  - the more, the merrier
  - not segmented or tagged
  - Internet up to 100%, not fixed
  - descriptive – size, not quality



# An intensively tagged corpus

**Thai Word-Transcription Tagger 1.0**

Tag Input Text: D:\wire\TNC\tagged\MRP01-report2.txt

Save Output

*Thai National Corpus Project*

Written by Wirote Aroonmanakorn, Dept. of Linguistics Chulalongkorn University

Output

```
<p n="1"><s n="1"><w tran="raaj0Naan0kaan0pra1 chum0">ราชบัณฑิตยสถาน</w></s></p>
<p n="2"><s n="2"><w tran="kha3na3">ขนม</w><w tran="tham0Naan0">ทำขนม</w><w tran="phUua2">เพื่อ</w><w
tran="tit1 taam0">ติดตาม</w><w tran="truuat1 sOOp1">ตรวจสอบ</w><w tran="pan0haa4">ปัญหา</w><w tran="saan4">สาร</w><w
tran="pal1rOOt1">ประธาน</w></s></p>
<p n="3"><s n="3"><w tran="khraN3">ครั้ง</w><w tran="thii2">ที่</w></s><s n="4"><w tran="1/2539">1/2539</w></s></p>
<p n="4"><s n="5"><w tran="wan0">วัน</w><w tran="can0">กัน</w><w tran="thii2">ที่</w></s><s n="6"><w>22</w></s><s n="7">
<w tran="mee0saa4jon0">เมษายน</w></s><s n="8"><w>2539</w></s><s n="9"><w tran="wee0laa0">เวลา</w></s><s n="10"><w
tran="14.00">14.00</w></s><s n="11"><w tran="nOO0">น</w></s></p>
<p n="5"><s n="12"><w tran="na3">น</w></s><s n="13"><w tran="hOON2">ห้อง</w><w tran="pra1 chum0">ราชบัณฑิตยสถาน</w><w
tran="chan3">จีน</w></s><s n="14"><w>8</w></s><s n="15"><w tran="krom0">กรม</w><w tran="khuuap2khum0">ควบคุม</w><w
tran="mon0phit3">มาพิชัย</w></s></p>
<p n="6"><s n="16"><c>&tab;</c><c>&tab;</c><c>&tab;</c><c>&tab;</c><c>&tab;</c><c>&tab;</c></s></p>
<p n="7"></p>
<p n="8"></p>
<p n="9"><s n="17"><w tran="phuu2">ผู้</w><w tran="khaw2">เข้า</w><w tran="nuuam2">ร่วม</w><w tran="pra1 chum0">ราชบัณฑิตยสถาน</w></s></p>
<p n="10"><s n="18"><gap desc="name_list"/></s></p>
<p n="11"><s n="19"><w tran="raaj0chUU2">ราชบัณฑิต</w><w tran="phuu2">ผู้</w><w tran="maj2">ไม่</w><w tran="maa0">มา</w><w
tran="pra1 chum0">ราชบัณฑิตยสถาน</w></s></p>
<p n="12"><s n="20"><gap desc="name_list"/></s></p>
<p n="13"></p>
<p n="14"><s n="21"><w tran="p@@t1">เปิด</w><w tran="pra1 chum0">ราชบัณฑิตยสถาน</w><w tran="wee0laa0">เวลา</w></s><s n="22">
```



# SEAlong Thai WebCorpus

- **experimental research tool**
  - study Thai-language change and growth
  - assist in Thai dictionary design / lexicographic research
  - evaluate new techniques in web-as-corpus design



# Thai WebCorpus size / coverage

- **100% Web-based, not tagged**
- **ที่ = 100,000,000 + hits**
  - 1,000 chars (250 words) / page
  - 100,000,000,000 characters
  - 25,000,000,000 words
- **a gigaword corpus now**
- **a terabyte corpus in (3? years)**



# Thai WebCorpus design

- **built ‘on the fly’**
  - no stored data
- **powered by Google, Yahoo ...**
  - use API, screen-scraping
  - multiple parallel searches
- **some genre control**
  - informal: blogs, chat ...
  - semi-formal: news, /or.th ...
  - formal: /go.th, /ac.th ...
- **some challenges**
  - outwitting page ranking
  - on-the-fly segmentation



# Thai WebCorpus home page

**SEAlang Web Corpus**

100 results   
any date   
match phrase   
don't sort results   
visit all available sites

**Yahoo** **Google**

**Search** **Sample**  
**Analyze** **Define**  
**Predict** **news2000**  
**Churn 0**

ที่

New Words (Royal Inst. 2007)

Predict ☒ on ☐ off

< ||| >

**New Words (Royal Institute, 2007)**

ก๊งก๊ง

## About the SEAlang Thai Web Corpus

The Thai Web Corpus is an experimental research tool built to:

- support research studying Thai-language change and growth,
- assist in Thai dictionary design and lexicographic research,
- evaluate new techniques in web-as-corpus design.

The Thai Web Corpus supports SEAlang's **Thai in Transition** research project, which is investigating Thai perceptions of and responses to language change. Questions include (*see more ...*):

- do foreign loans have any discernible impact on Thai syntax or grammar?
- what semantic and phonological nativization processes serve to 'Thai-ize' loans?
- how effective is the Royal Institute's prescriptive orthography for slang and loans?
- are there any characteristic phonological processes that tend to generate new slang?
- does the distribution of slang characterize Thai Web space, or the slang itself?
- can dictionaries developed without text corpora hope to get it right?
- what semantic niches do foreign borrowings fill?
- what is the grammatical distribution of new and borrowed words?



# Specifying text genre

## SEAlang Web Corpus

100 results

any date

match phrase

don't sort results

visit all available sites

- visit all available sites
- edu
- organizations (.org)
- Thai**
- \*news sites
- \*message boards
- \*literature
- \*reference sites
- \*blogs
- webpace (.th)
- academic (ac.th)
- government (go.th)
- organizations (or.th)
- Wikipedia

New Words (Royal Inst. 2007)

## About the SEAlang Thai Web Corpus

The Thai Web Corpus is an e

- support research study
- assist in Thai dictionary
- evaluate new technique

The Thai Web Corpus support research project, which is inv

responses to language change

- do foreign loans have a or grammar?
- what semantic and phon to 'Thai-ize' loans?
- how effective is the Roy for slang and loans?
- are there any characteri to generate new slang?
- does the distribution of

# A simple search

### SEAlang Web Corpus

100 results

any date

match phrase

don't sort results

visit all available sites

Yahoo

Google

Search

Sample

Analyze

Define

Predict

news2000

Churn 0

แอ็บแบ๊ว

New Words (Royal Inst 2007)

New Words (Royal Inst 2007)

Foreign Words (Royal Inst 2006)

Slang (Min. of Ed. 2000)

Thai Slang (Jintana 2003)

Merged list

New Words (Royal Institute, 2007)

ก่งกึ่ง

กึ่ง

ก่งโก่งกอก

Query "แอ็บแบ๊ว" Yahoo found 504,000 pages (per hit), 100 returned, 93 actual sightings

**Title** แอ็บแบ๊ว - วิกีพีเดีย  
**Url** <http://th.wikipedia.org/...> **Size:** 31289 bytes (US) 6/15/2008  
**Summary** ปฏิวัติ "แอ็บแบ๊ว" 6 ชั่วโมงที่โรงแรมบอลของสุทธิชัย หยุน ... ข้อมูลเกี่ยวกับ แอ็บแบ๊ว อาจสามารถหาอ่านได้จากเมนู ภาษาอื่น ด้านซ้าย .

**Title** แอ็บแบ๊ว on Flickr - Photo Sharing!  
**Url** <http://www.flickr.com/...> **Size:** 34119 bytes (US) 6/20/2008  
**Summary** Flickr is almost certainly the best content management and sharing ... · Everyone's Upload Members For a Location badzboy's Photostory

**Title** แอ็บแบ๊ว คืออะไร ?  
**Url** <http://www.inboxstory.com/...> **Size:** 24 bytes (US) 5/24/2008  
**Summary** เคยได้ยินคำว่า "แอ็บแบ๊ว" กันมั้ยคะ? เอ็นทรี่นี้ไม่ได้เขียนโดยใช้ reference จากตัวเอง คนที่แอ็บแบ๊วจนชำนาญก็จะขนาดแก้มที่ป่องกำลัง

**Title** Bevariety Forum :: อ่าน - วิธีการถ่ายรูปแอ็บ  
**Url** <http://www.thaifooddirect.com/...> **Size:** 11 bytes (US) 6/30/2008



# A simple sample

### SEAlang Web Corpus

100 results ▼

any date ▼

match phrase ▼

don't sort results ▼

visit all available sites ▼

Yahoo
Google

Search
Sample

Analyze
Define

Predict
news2000

Churn 0

แฉับแฉับ

<
|||
>

Query "แฉับแฉับ" Yahoo found **504,000** pages (about **98 pages per hit**), 100 returned actual sightings. Suppressing likely duplicates.

	เขตปลอดคน	แฉับแฉับ	
๕ ก๊อด้วจิ ขอเสนอ วิธีการถ่ายภาพ	แฉับแฉับ	!!(แบบมือโปร)	เหมาะสำหรับสาวทุ
วิธีการถ่ายภาพ	แฉับแฉับ	!!(แบบโปร)	การถ่ายรูป
'วไป เรื่องทั่วไปเกี่ยวกับบอล	แฉับแฉับ	!!แอบให้เล็กๆน้อยๆ "	... หัวขั
อ:	แฉับแฉับ	!!แอบให้เล็กๆน้อยๆ (อ่าน 621 ค	
เมื่อ หมา. "	แฉับแฉับ	" - Liverpool Thailand Fanclu	
assword : เรื่อง : คู่มือสอน "	แฉับแฉับ	"... อีกอีก แต่คนนี้อะ "	<
เตรียมตัวสำหรับถักตุ๊กตา "หมี	แฉับแฉับ	" 2 ครั้ง. เริ่มต้นถักจริงๆกันช	
ปฏิวัติ "	แฉับแฉับ	" 6 ชั่วโมงที่โรงแรม 5 ดาวกลาง	
ube - 14.10.07 Micky Say "อย่า	แฉับแฉับ	" [Thai Word]	
เคยได้ยินคำว่า "	แฉับแฉับ	" กัน ...	
'า ว้ายๆ นี่มัน "	แฉับแฉับ	" กันนี่หว่า หรือแกนนำบางคนจะ	
เคยได้ยินคำว่า "	แฉับแฉับ	" กันมั้ยคะ? เอ็นทรี่นี้ไม่ได้	
วันนี้คุณ "	แฉับแฉับ	" กันหรือยัง?!	
ทราบได้ไม่ ก็ให้นึกถึงคำว่า "	แฉับแฉับ	" ขึ้นมาอย่างจับใจว่า. ประมาณ	
มีริมฝีปาก ... ยังบอกวิธีฝึก "	แฉับแฉับ	" ง่ายๆคือยืนหน้ากระจก ฝึกทำหน	
"วันก่อนเรายังงงคำว่า "	แฉับแฉับ	" จากกิจกรรมที่ต้องโฟกัสในมหา	
เกาหลีแบน "เหม่ม	แฉับแฉับ	" จากรายการทีวี   BLike.net -	
คลับอร์ด 16 ข้อ. ช่วยเหลือ ...	แฉับแฉับ	" ตัวเลือก. UnDer_DiNaMiC. วั	
... สำหรับคำว่า "	แฉับแฉับ	" นั้น ผู้สื่อข่าวรายงานว่า มี	
เทรนด์ "	แฉับแฉับ	" น่ารักหรือต๊อดจิด	
... ยากที่จะคาดเดาว่ากระแส "	แฉับแฉับ	" มีต้นกำเนิดมาจากที่ใดกันแน	

### New Words (Royal Institute, 2007)

ก่งก้ง

ก้ง

กงโก้งกกงก

กฎบัตรกฎหมาย

กลืนทัวร์

กดจุด

กฉับ

# Analyzing collocates

Deciding how to search for แอ๊บแบ๊ว ... 543000 hits, so launching Yahoo parallel search ... 1, 101, 201, 301, 401, 501, 601, 701, 801, 901, Yahoo returned 783 items from 1000 pages (claimed from 515000 to 543000 available).

Target "แอ๊บแบ๊ว" Searched 97305 characters. **Highlighted** items are dictionary entries. Results: left, **177 distinct leading collocates**, right: **122 distinct trailing collocates** (both from 776 contexts) 'Show leading ...' and 'show trailing ...' display up to 5 typical contexts for every collocate.

SHOW LEADING COLLOCATES W/CONTEXTS

SHOW COLLOCATES WITHOUT CONTEXTS

SHOW TRAILING COLLOCATES W/CONTEXTS

(227/29.2%) \_แอ๊บแบ๊ว (100/12.8%) "แอ๊บแบ๊ว  
(20/2.5%) หื่นแอ๊บแบ๊ว (18/2.3%) สาวแอ๊บแบ๊ว  
(17/2.1%) การถ่ายรูปแอ๊บแบ๊ว (17/2.1%) .แอ๊บแบ๊ว  
(16/2%) หราแอ๊บแบ๊ว (14/1.8%) ไม่แอ๊บแบ๊ว  
(11/1.4%) ขอแอ๊บแบ๊ว (10/1.2%) หัวใจแอ๊บแบ๊ว  
(10/1.2%) 'แอ๊บแบ๊ว (7/0.9%) อย่าแอ๊บแบ๊ว (7/0.9%)  
ที่แอ๊บแบ๊ว (6/0.7%) คนแอ๊บแบ๊ว (6/0.7%)  
มาแอ๊บแบ๊ว (6/0.7%) ต้องแอ๊บแบ๊ว (6/0.7%)  
แบบแอ๊บแบ๊ว (6/0.7%) แต่แอ๊บแบ๊ว (6/0.7%)  
(แอ๊บแบ๊ว (6/0.7%) จะแอ๊บแบ๊ว (5/0.6%)  
ภาพแอ๊บแบ๊ว (5/0.6%) บล็อกแอ๊บแบ๊ว (5/0.6%)  
หมีแอ๊บแบ๊ว (5/0.6%) การแอ๊บแบ๊ว (4/0.5%)

(247/31.8%) แอ๊บแบ๊ว\_ (134/17.2%)  
แอ๊บแบ๊ว. (131/16.8%) แอ๊บแบ๊ว" (27/3.4%)  
แอ๊บแบ๊วาว (18/2.3%) แอ๊บแบ๊ว! (11/1.4%)  
แอ๊บแบ๊วเป็น (10/1.2%) แอ๊บแบ๊ว) (10/1.2%)  
แอ๊บแบ๊ว' (10/1.2%) แอ๊บแบ๊วของ (9/1.1%)  
แอ๊บแบ๊วแบบ (6/0.7%) แอ๊บแบ๊วมา (6/0.7%)  
แอ๊บแบ๊วจะ (5/0.6%) แอ๊บแบ๊วจน (5/0.6%)  
แอ๊บแบ๊วเหมือนกัน (5/0.6%) แอ๊บแบ๊วแล้ว  
(4/0.5%) แอ๊บแบ๊วเหมือน (3/0.3%)  
แอ๊บแบ๊วกัน (3/0.3%) แอ๊บแบ๊ว? (3/0.3%)  
แอ๊บแบ๊วนะ (3/0.3%) แอ๊บแบ๊วอะ (3/0.3%)  
แอ๊บแบ๊วได้ (3/0.3%) แอ๊บแบ๊วเลย (3/0.3%)



# Collocates in context

(18/2.3%) สาวแอ๊บแบ๊ว

ทริปหน้าหนาว...กับ <	สาวแอ๊บแบ๊ว	... .. ด้าน "สุรยุทธ์" ทันสมัย บอก
ความรัก, ซอปปิ้ง, งาน , ... ไซ <	สาวแอ๊บแบ๊ว	(อย่างแน่นนอน) และ ... หวานใจ, หาก
ตัเลย ... สวยใส ไร้มารยา ชอบ <	สาวแอ๊บแบ๊ว	ผิวขาวปานหยวก น้องอีฟจิง. นั
ภาพลักษณ์การเป็นนักร้อง <	สาวแอ๊บแบ๊ว	ที่ไม่เปิดตัวเรื่องราวความ
ียดา หรือจอยรินลณี ปากของ <	สาวแอ๊บแบ๊ว	จะถูกกำหนดให้มีริมฝีปากบมน

(17/2.1%) การถ่ายรูปแอ๊บแบ๊ว

ถ่ายรูปแอ๊บแบ๊ว... <	การถ่ายรูปแอ๊บแบ๊ว	เป็นส่วนผสมของการถ่ายรูป
ลายใจเจี๊มา ... วิธี <	การถ่ายรูปแอ๊บแบ๊ว	แบบมืออาชีพ โดย Krubaball. วันนี้ เ
พแอ๊บแบ๊ว!!(แบบโปร) <	การถ่ายรูปแอ๊บแบ๊ว	เป็นส่วนผสมของการถ่ายรูป
<	การถ่ายรูปแอ๊บแบ๊ว	เป็นส่วนผสมของการถ่ายรูป
ู้ทรงคุณค่า ... วิธี <	การถ่ายรูปแอ๊บแบ๊ว	แบบมืออาชีพ โดย ป้าแล่ม. วันน

(17/2.1%) .แอ๊บแบ๊ว

ังกร กั้ว กั้ว อะ. กิ่งก็ออออ... <	.แอ๊บแบ๊ว	...อม...รียอะ. กั้ว. กั้ว. ... ประกาศผ
บอกได้คำเดียว...ว่า...น่ารัก... <	.แอ๊บแบ๊ว	. Interests. หนังสือวรรณกรรมแฟนตาซี.
นายคุณลัดใบ..แวน..สก็อย..อีโม. <	.แอ๊บแบ๊ว	! ... วินโดว กับแมค ใครกันที่เจ
กับ designer จาก Jaspal ... สไตล์แต่งหน้า <	.แอ๊บแบ๊ว	1. ก่อนแต่งหน้าทามอยซ์เจอไร
ะวังสารเคมี. เชื้อ ... มาอีกแระ <	.แอ๊บแบ๊ว	.ว.ว.ว. จาก blue_sushi พุธ, 24/1/2550 เวลา : 15:30
		...

(16/2%) หราแอ๊บแบ๊ว

หัวใจได้ฝัน. เพลง : ... จินต <	หราแอ๊บแบ๊ว	. อัลบั้ม : จินตหรา ครบเครื่อง
จินต <	หราแอ๊บแบ๊ว	. เพลง จี๋หอย - พี ... พ ว ก เ ร า ช า
ลาก) ถ่มหนอยคดอยด้วย จินต <	หราแอ๊บแบ๊ว	(จินตหรา พนลาก) หัวใจเรือกนา

# Looking ahead

- Corpora of the past?
- Corpora of the future?
- Yearly Thai language ‘DNA sample’
  - true cost:
    - 1998: \$100,000 / terabyte
    - 2008: \$200 / terabyte (= year)
    - 2018: \$0.50 / terabyte?
  - true value:
    - 2008: little
    - 2058: much
    - 2108: enormous!



# In conclusion

- **Thai is always in transition**
  - it's the secret of every successful language!
- **Corpora help us understand change**
  - and every kind of corpus helps
- **SEAlang Thai WebCorpus**
  - ongoing development
  - <http://sealang.net/webcorpus/thai>



# Thai in Transition *and the Thai* Gigaword/Terabyte Web Corpus

Rikker Dockum and Doug Cooper  
Center for Research in Computational Linguistics  
CRCL Inc is a US 501(c)3 nonprofit

<http://sealang.net/webcorpus/thai>

<http://sealang.net/archives/ri-tera-2008.pdf>

