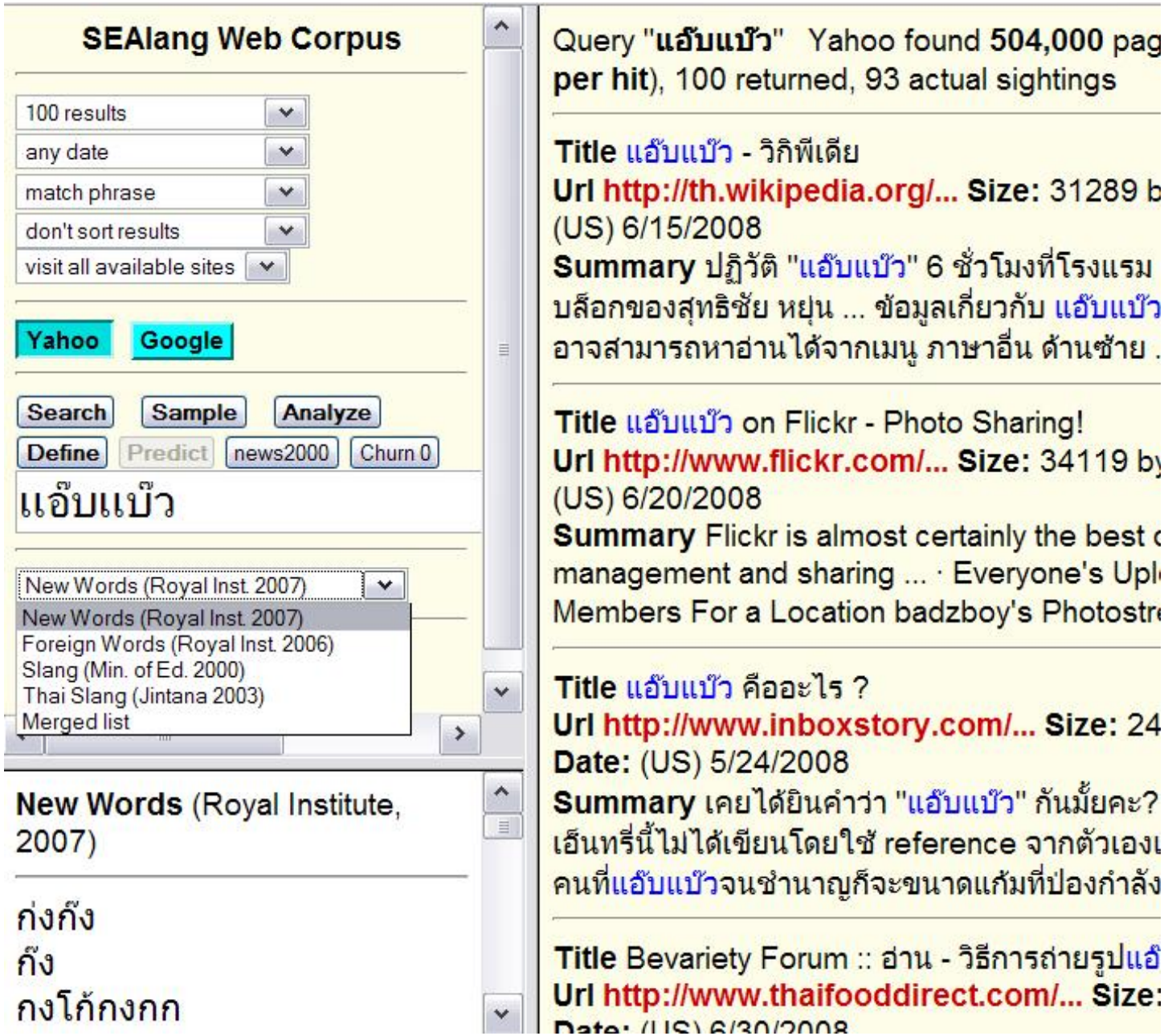
Thai in Transition
and the Thai
Gigaword/Terabyte Web Corpus

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
แอ๊บแบ๊ว
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)
ก่งก๊ง
ก๊ง
กงโก้กงกก
Query "แอ๊บแบ๊ว" Yahoo found 504,000 pag
per hit), 100 returned, 93 actual sightings
Title แอ๊บแบ๊ว - วิกิพีเดีย
Url http://th.wikipedia.org/... Size: 31289 b
(US) 6/15/2008
Summary ปฏิวัติ "แอ๊บแบ๊ว" 6 ชั่วโมงที่โรงแรม
บล็อกของสุทธิชัย หยุน ... ข้อมูลเกี่ยวกับ แอ๊บแบ๊ว
อาจสามารถหาอ่านได้จากเมนู ภาษาอื่น ด้านซ้าย .
Title แอ๊บแบ๊ว on Flickr - Photo Sharing!
Url http://www.flickr.com/... Size: 34119 by
(US) 6/20/2008
Summary Flickr is almost certainly the best
management and sharing ... · Everyone's Upl
Members For a Location badzboy's Photostr
Title แอ๊บแบ๊ว คืออะไร ?
Url http://www.inboxstory.com/... Size: 24
Date: (US) 5/24/2008
Summary เคยได้ยินคำว่า "แอ๊บแบ๊ว" กันมั้ยคะ?
เอ็นทรี่นี้ไม่ได้เขียนโดยใช้ reference จากตัวเอง
คนที่แอ๊บแบ๊วจนชำนาญก็จะขนาดแก้มที่ป่องกำลัง
Title Bevariety Forum :: อ่าน - วิธีการถ่ายรูปแอ๊
Url http://www.thaifooddirect.com/... Size:
Date: (US) 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

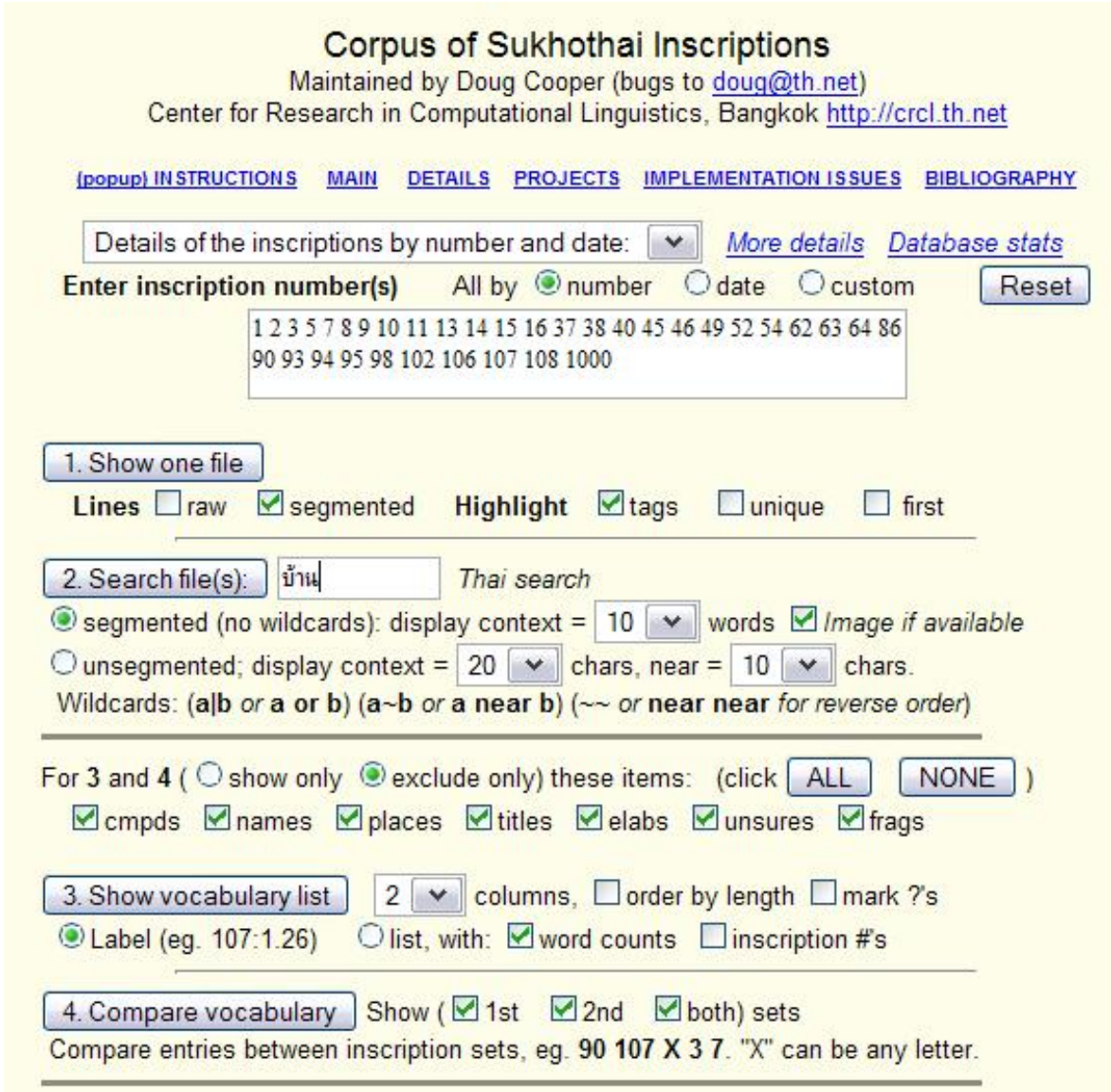
- do foreign loans have any impact on Thai syntax or grammar?
- what semantic and phonological nativization processes 'Thai-ize' loans?
- how effective is the Royal Institute's prescriptive orthography for slang and loans?
- are there typical phonological processes that tend to generate new slang?
- 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



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.

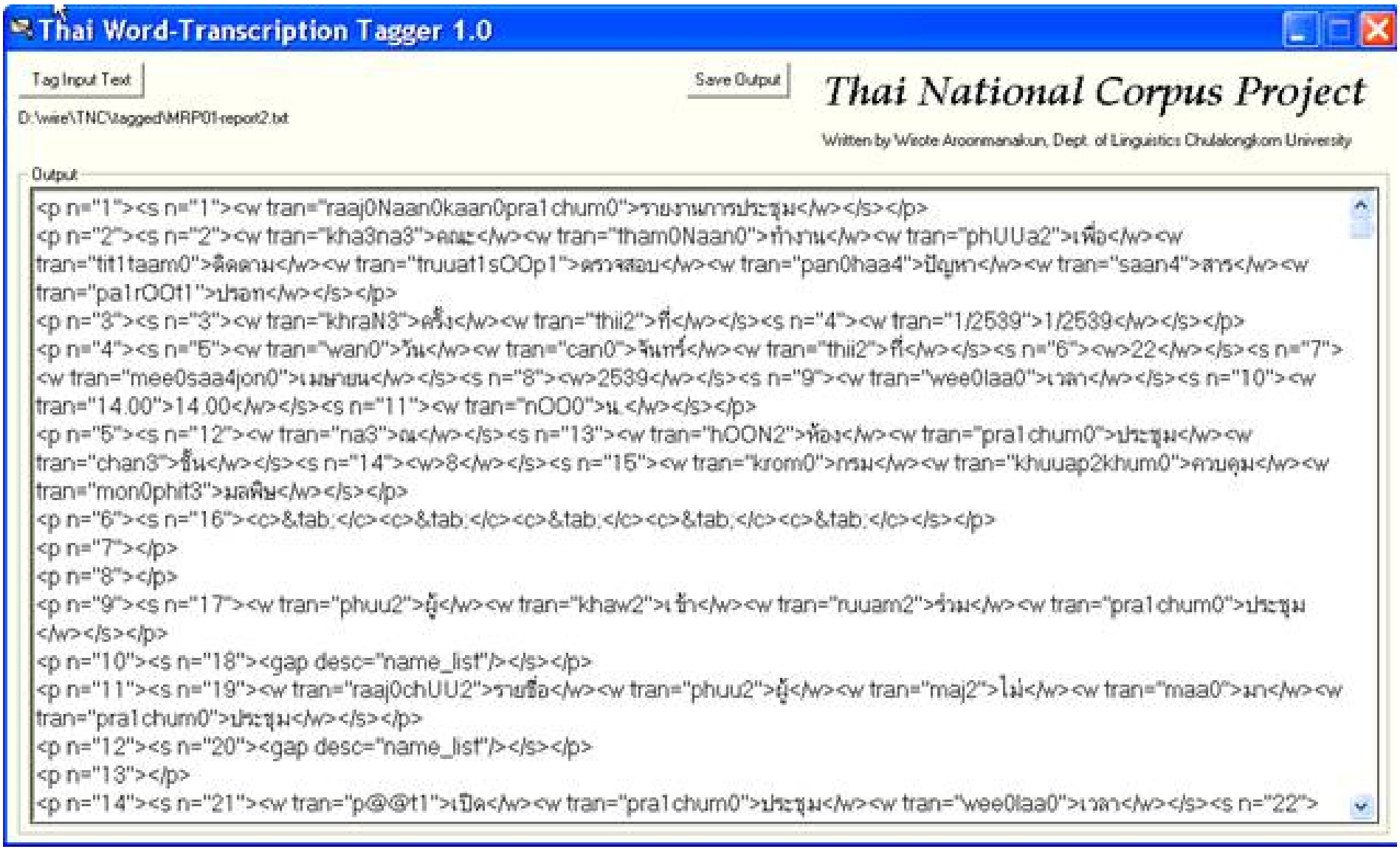
Study word Highlight or doubleclick any word, then click here to see each line it occurs on. The word may be a compound (but only spaces).

กุศลผลบุญ-E 15:1.33
บ่ฆ่าบ่ตี-E 1:1.31 5:1.22
บ่ชอบบ่ยำ-E 3:1.23
มีไร่มีนา-E 1:2.34 1:3.3
ไคร่ใจในใจ-E 1:4.10
จงคงตรงต่อ-E 15:3.7
ในน้ำในถ้ำ-E 45:1.2
ภายนอกในใจ-E 40:1.17 40:2.6
มาคว้ามาฉก-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
ผิดแผกแสกว้าง-E 1:1.25
ผู้เฒ่าผู้แก่-E 2:1.52 14:2.15
ผู้เถ้าผู้แก่-E 106:2.36
ผู้รู้ผู้เห็น-E 38:2.30
ผู้ใหญ่ผู้ราม-E 40:1.2
พ่อแม่พี่น้อง-E 106:1.24
พ่อลูกพี่น้อง-E 7:4.20
พายลูนปูนหลัง-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
ภายลุนปุนหลัง-E 11:1.17
มั่นคงตรงอยู่-E 40:1.10
มืดมนท์อนธการ-E 40:1.23
เยียคดเยียคู้-E 5:1.24
หมอนนั่งหมอนโนน-E 1:2.15
หมากส้มหมากหวาน-E 1:1.13
หินแลงแปลงเลียด-E 62:2.9
อันเหง้าเจ้าไทย-E 106:1.25
คุณในปลายาในข้าว-E 64:1.2
ใจรักภักดีไมตรี-E 64:1.9
ช่อยปลดช่อยมล้าง-E 64:1.11
ด้วยแกล้วด้วยหาญ-E 1:4.15 3:2.5
ด้วยรู้ด้วยหลวัก-E 1:4.15
ด้านใต้ด้านเหนือ-E 62:2.15
ทั้งสิ้นทั้งหลาย-E 1:2.11
บ่มีเงือนบ่มีทอง-E 1:1.30
บ่ให้ฉิบบ่ให้หาย-E 2:1.49
บ่ให้ล้มบ่ให้ตาย-E 2:1.50
บ้านใหญ่บ้านเล็ก-E 1:2.35 1:3.3
พาทย์พิณแตรสังข์-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



SEAlang 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

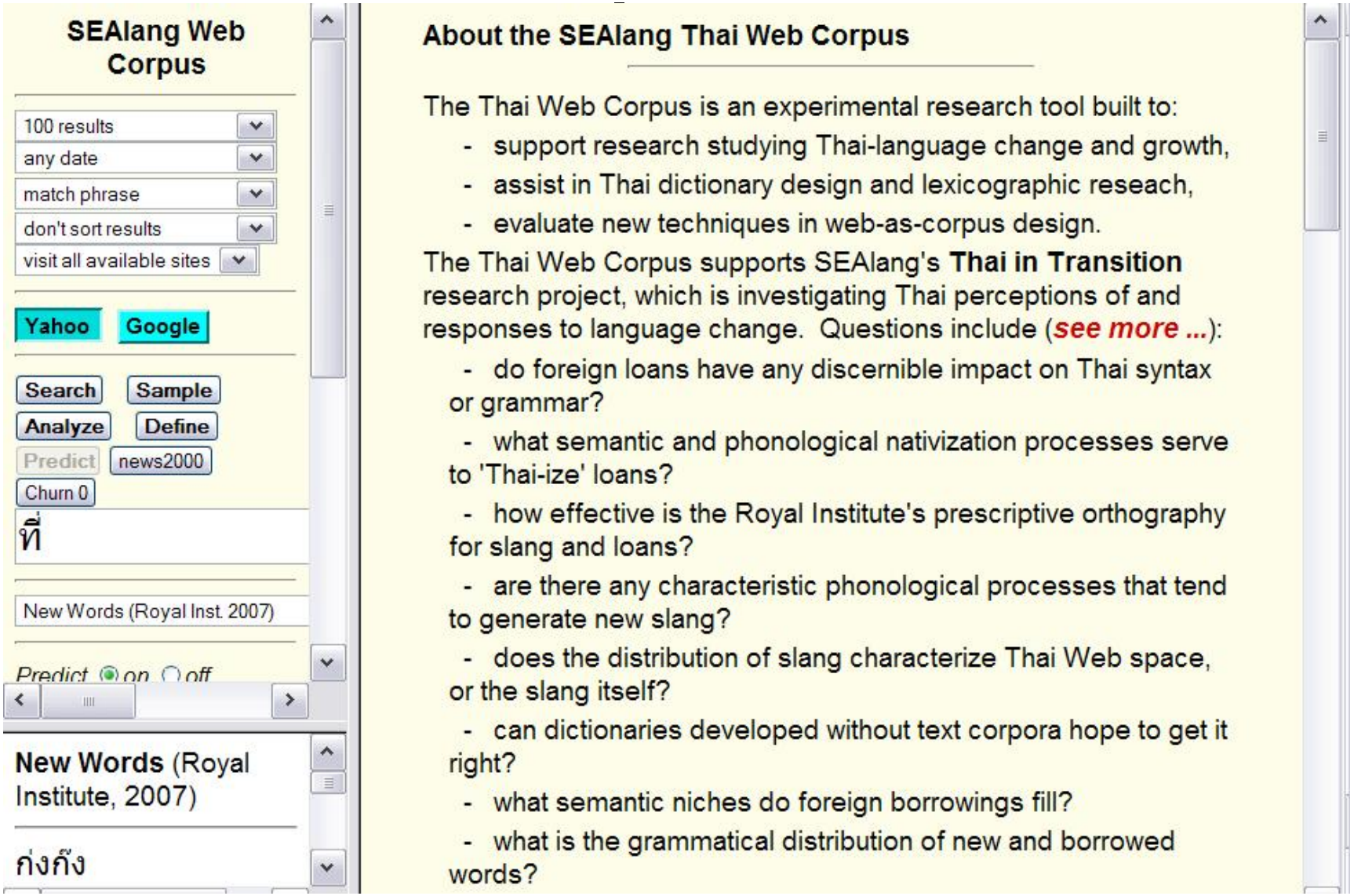
Thai WebCorpus size / coverage

- **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)**

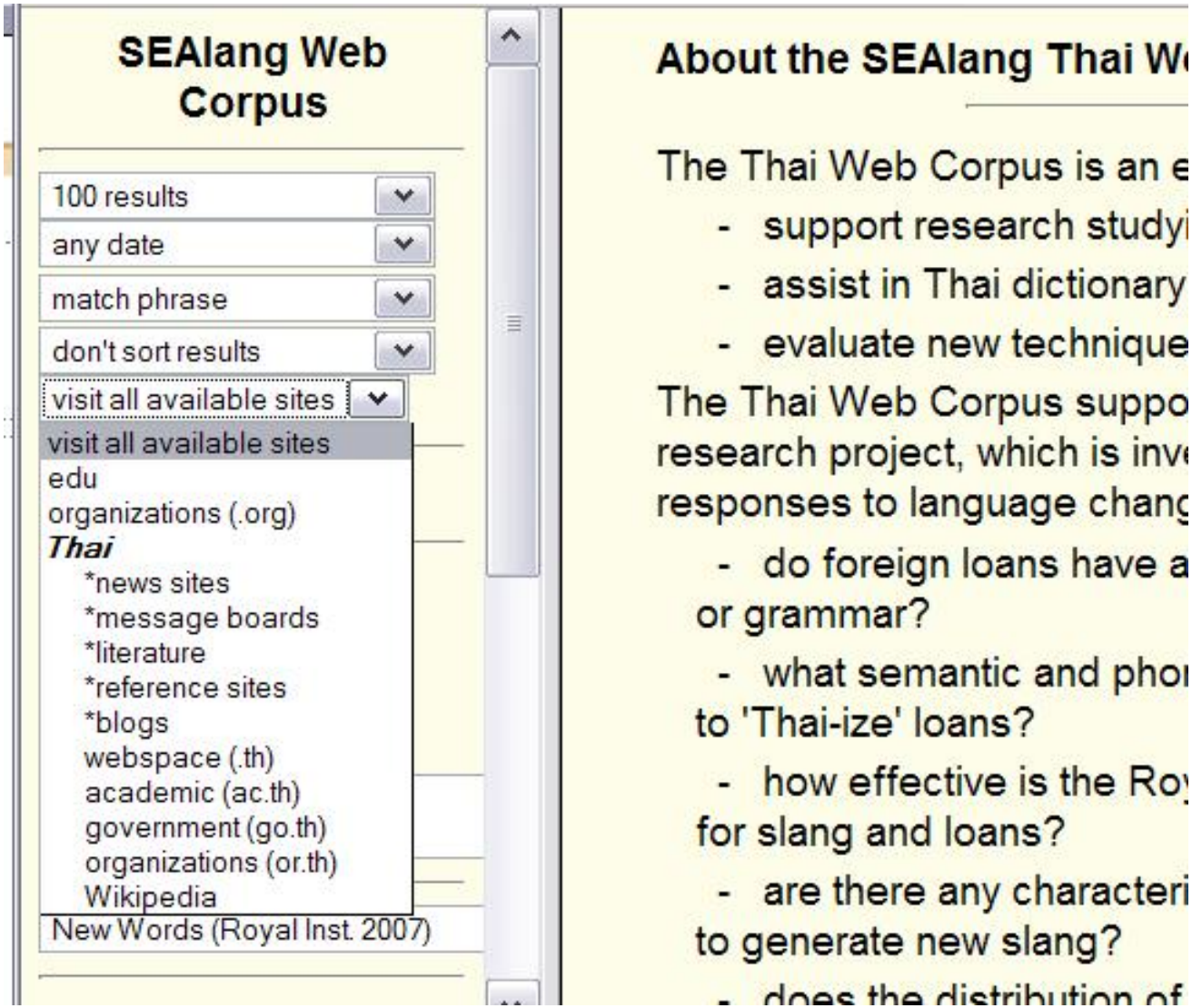
Thai WebCorpus design

- 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

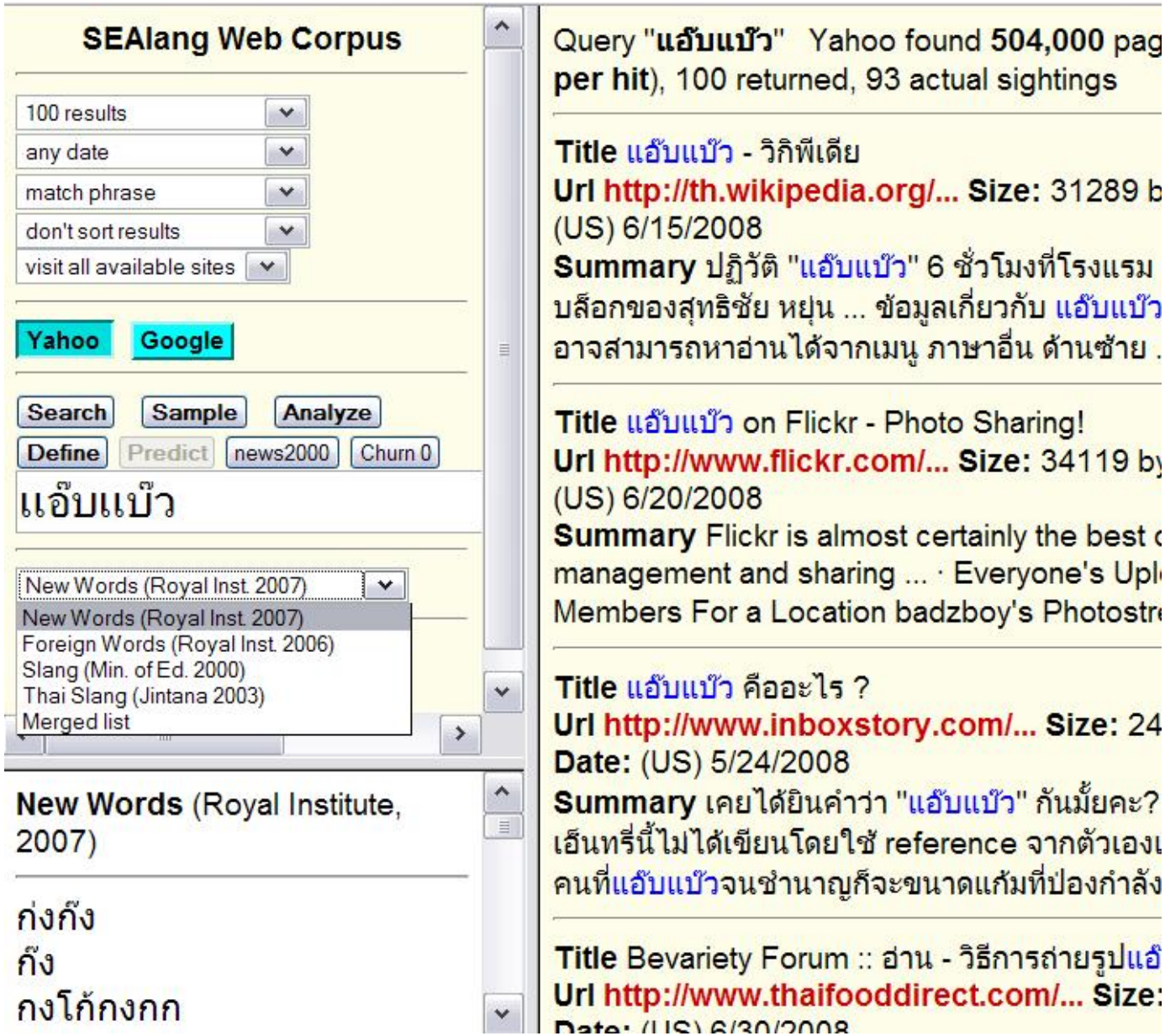
Thai WebCorpus home page



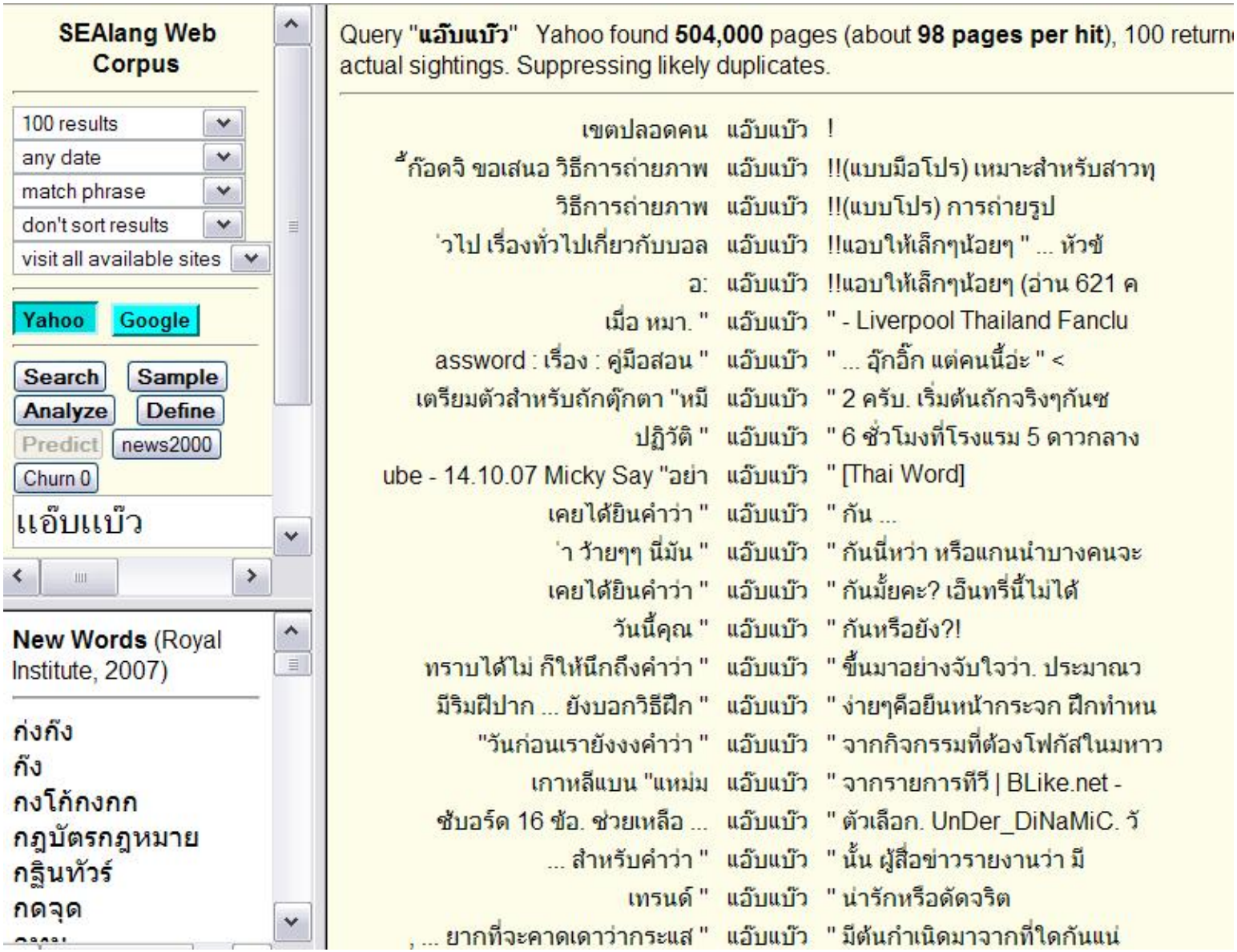
Specifying text genre



A simple search



A simple sample



Analyzing collocates

Deciding how to search for แอ๊บแบ๋ว ... 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 "**แอ๊บแบ๋ว**" 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%) _แอ๊บแบ๋ว (100/12.8%) "แอ๊บแบ๋ว
(20/2.5%) หื่นแอ๊บแบ๋ว (18/2.3%) สาวแอ๊บแบ๋ว
(17/2.1%) การถ่ายรูปแอ๊บแบ๋ว (17/2.1%) .แอ๊บแบ๋ว
(16/2%) หราแอ๊บแบ๋ว (14/1.8%) ไม่แอ๊บแบ๋ว
(11/1.4%) ขอแอ๊บแบ๋ว (10/1.2%) หัวใจแอ๊บแบ๋ว
(10/1.2%) 'แอ๊บแบ๋ว (7/0.9%) อย่าแอ๊บแบ๋ว (7/0.9%)
ที่แอ๊บแบ๋ว (6/0.7%) คนแอ๊บแบ๋ว (6/0.7%)
มาแอ๊บแบ๋ว (6/0.7%) ต้องแอ๊บแบ๋ว (6/0.7%)
แบบแอ๊บแบ๋ว (6/0.7%) แต่แอ๊บแบ๋ว (6/0.7%)
(แอ๊บแบ๋ว (6/0.7%) จะแอ๊บแบ๋ว (5/0.6%)
ภาพแอ๊บแบ๋ว (5/0.6%) บล็อกแอ๊บแบ๋ว (5/0.6%)
หมีแอ๊บแบ๋ว (5/0.6%) การแอ๊บแบ๋ว (4/0.5%)

(247/31.8%) แอ๊บแบ๋ว_ (134/17.2%)
แอ๊บแบ๋ว. (131/16.8%) แอ๊บแบ๋ว" (27/3.4%)
แอ๊บแบ๋ววว (18/2.3%) แอ๊บแบ๋ว! (11/1.4%)
แอ๊บแบ๋วเป็น (10/1.2%) แอ๊บแบ๋ว) (10/1.2%)
แอ๊บแบ๋ว' (10/1.2%) แอ๊บแบ๋วของ (9/1.1%)
แอ๊บแบ๋วแบบ (6/0.7%) แอ๊บแบ๋วมา (6/0.7%)
แอ๊บแบ๋วจะ (5/0.6%) แอ๊บแบ๋วจน (5/0.6%)
แอ๊บแบ๋วเหมือนกัน (5/0.6%) แอ๊บแบ๋วแล้ว
(4/0.5%) แอ๊บแบ๋วเหมือน (3/0.3%)
แอ๊บแบ๋วกัน (3/0.3%) แอ๊บแบ๋ว? (3/0.3%)
แอ๊บแบ๋วนะ (3/0.3%) แอ๊บแบ๋วอ่ะ (3/0.3%)
แอ๊บแบ๋วได้ (3/0.3%) แอ๊บแบ๋วเลย (3/0.3%)

Collocates in context

(18/2.3%) สาวแอ๊บแบ๊ว

ทริปหน้าหนาว...กับ <	สาวแอ๊บแบ๊ว	... ... ด้าน "สุรยุทธ์" ทันสมัย บอก
ความรัก, ชอปปิ้ง, งาน , ... ใช่ <	สาวแอ๊บแบ๊ว	(อย่างแน่นอน) และ ... หวานใจ, หาก
ด้เลย ... สวยใส ไร้มารยา ชอบ <	สาวแอ๊บแบ๊ว	ผิวขาวปานหยวก น้องอีฟจัง. น้
ภาพลักษณ์การเป็นนักร้อง <	สาวแอ๊บแบ๊ว	ที่ไม่เปิดตัวเรื่องราวความ
ิยดา หรือจอยรินลณี ปากของ <	สาวแอ๊บแบ๊ว	จะถูกกำหนดให้มีริมฝีปากบนบ

(17/2.1%) การถ่ายรูปแอ๊บแบ๊ว

ถ่ายรูปแอ๊บแบ๊ว... <	การถ่ายรูปแอ๊บแบ๊ว	เป็นส่วนผสมของการถ่ายรูปบุ
ลายใจเจ๊มาก ... วิธี <	การถ่ายรูปแอ๊บแบ๊ว	แบบมืออาชีพ โดย Krubaball. วันนี้ เว
พแอ๊บแบ๊ว!!(แบบโปร) <	การถ่ายรูปแอ๊บแบ๊ว	เป็นส่วนผสมของการถ่ายรูปบุ
<	การถ่ายรูปแอ๊บแบ๊ว	เป็นส่วนผสมของการถ่ายรูปบุ
ู้ทรงคุณค่า ... วิธี <	การถ่ายรูปแอ๊บแบ๊ว	แบบมืออาชีพ โดย ป้าแล่ม. วันน

(17/2.1%) .แอ๊บแบ๊ว

ังกร กัว กัว อ่ะ. กิ้งกือออออ... <	.แอ๊บแบ๊ว	...อม...รัยอ่ะ. กัว. กัว. ... ประกาศผ
บอกได้คำเดียว...ว่า...น่ารัก... <	.แอ๊บแบ๊ว	. Interests. หนังสือวรรณกรรมแฟนตาซี.
นวยุคผลัดใบ..แว้น..สก๊อย..อีโม. <	.แอ๊บแบ๊ว	! ... วินโดว กับแมค ใครกันที่เจ
กับ designer จาก Jaspal ... สไตล์แต่งหน้า <	.แอ๊บแบ๊ว	1. ก่อนแต่งหน้าทามอยซ์เจอไรเ
ะวังสารเคมี. เชื่อ ... มาอีกแระ <	.แอ๊บแบ๊ว	.ว.ว.ว. จาก blue_sushi พุธ, 24/1/2550 เวลา : 15:30 ...

(16/2%) หราแอ๊บแบ๊ว

หัวใจไต่ฝัน. เพลง : ... จินต <	หราแอ๊บแบ๊ว	. อัลบั้ม : จินตหรา ครบเครื่อง
จินต <	หราแอ๊บแบ๊ว	. เพลง จี่หอย - พี ... พ ว ก เ ร า ช า
ลาก) ก่อมหอยคอยก้าย จินต <	หราแอ๊บแบ๊ว	(จินตหรา พูนลาภ) หัวใจเรียกหา

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