Issues in determining and annotating linguistic information for the corpus-based 100 high frequency Thai words for learners

Yuphaphann Hoonchamlong
University of Hawaii-Manoa

yuphapha@hawaii.edu

SEALS22. May 31, 2012

Long term Goal: a frequency dictionary for Thai language learners

- Inspired by:
  - COBUILD English Dictionary for Advanced Learners
  - Routledge's Frequency Dictionary: Core Vocabulary for Learners
- Dictionary for Thai Language learners
  - Mary Haas: Thai-English Student's Dictionary. Stanford University Press. 1964. Last reprint 1967.

A PROGRESS REPORT

- Preliminary research

1. determine the 100 most frequent Thai words used spreading across the range of various genres of texts, based on the raw frequency list extracted from the Thai National Corpus (TNC)
2. linguistic annotation
   - part of speech,
   - word meaning in English, grammatical explanation, a sample sentence illustrating the use for each function/meaning,
3. indices grouped by a) Thai alphabetical order; b) part of speech ; and c) theme/semantic domain

Word Frequency Information and Language Learning

- sequencing and grading of learning materials in the language curriculum design
  - first 1000 high frequency words cover 84.3% of conversation, 82.3% of fiction, 75.6% of newspapers and 73.5% of academic text (Nation 2001)
- prioritizing the words that learners are likely to encounter most often in language use
- frequency of forms
- frequency of meaning and uses

Thai National Corpus

http://ling.arts.chula.ac.th/TNC/

http://ling.arts.chula.ac.th/tnc2/ for corpus search

- Launched in 2006
- Designed to be comparable with the British National Corpus (BNC)
- 90% of data from 1998 on
- Annotations:
  - Document & textual information
  - Linguistic annotation: limited to word boundary, pronunciation transcription, names and foreign words
  - **NO Part of Speech tag**

<name type="person">อานันท์
ปัญญารชุน</name>

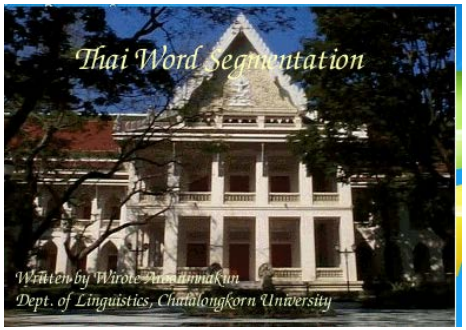
<w tran="khlaN0khOO2muun0">
คลังข้อมล</w>

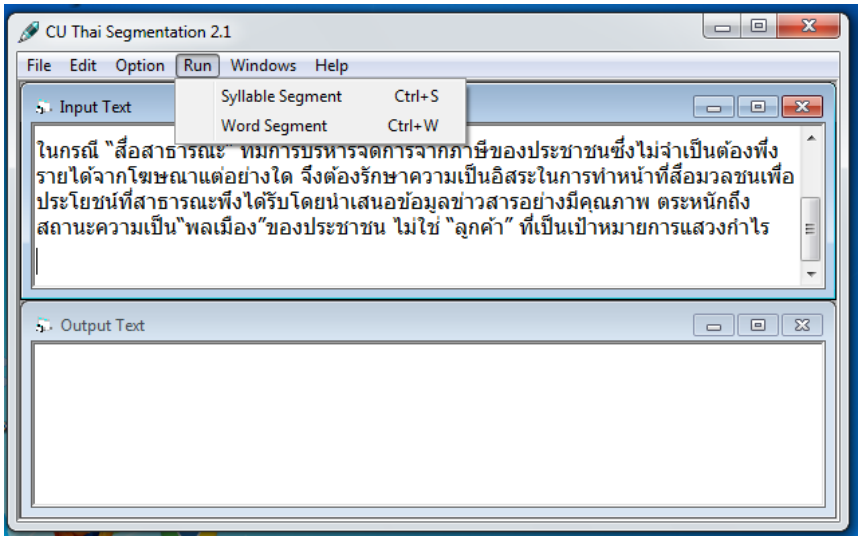
<w tran="kha1na1thii2">ขณะที่</w><w tran="khaa2">ค่า</w><w
tran="rak3saa4">รักษา</w><w tran="suuan1">ส่วน</w><w
tran="k@@n0">เกิน</w><w tran="thii2">ที่</w><w
tran="rooN0pha3jaa0baan0">โรงพยาบาล</w><w tran="ton2saN4kat1">
ต้นสังกัด</w><w tran="tOON2">ต้อง</w><w tran="caaj1">จ่าย</w>

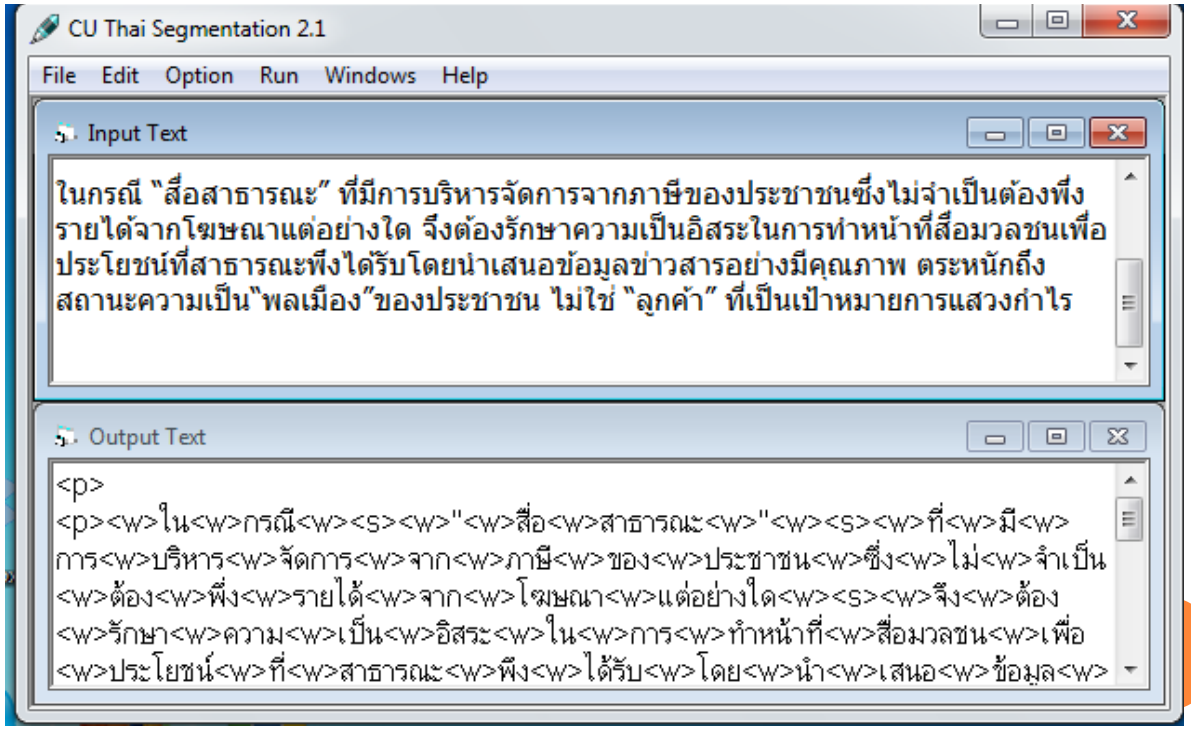
- Annotation encoding generated by "tagging program",
- then manually corrected
- Maximum size of each text = 40,000 words

Problems noted by developer

- Word segmentation
  - Word segmentation program (CU Thai Segmentation)
  - Manually corrected
  - Aroonmanakun (2002): 2 processes in word segmentation
    1. Syllable segmentation (trigram statistics)
    2. Syllable merging (collocation between syllables)
- Typo errors
- Spelling variations (7 variations for “internet”)


Thai Word Segmentation
Written by Wirote Aroonmanakun
Dept. of Linguistics, Chulalongkorn University


CU Thai Segmentation 2.1
File Edit Option Run Windows Help
Syllable Segment Ctrl+S
Word Segment Ctrl+W
Input Text
ในกรณี "สื่อสาธารณะ" ที่มีการบริหารจัดการจากภาษีของประชาชนซึ่งไม่จำเป็นต้องพึ่งรายได้จากโฆษณาแต่อย่างใด จึงต้องรักษาความเป็นอิสระในการทำหน้าที่สื่อมวลชนเพื่อประโยชน์ที่สาธารณะพึงได้รับโดยนำเสนอข้อมูลข่าวสารอย่างมีคุณภาพ ตระหนักถึงสถานะความเป็น"พลเมือง"ของประชาชน ไม่ใช่ "ลูกค้า" ที่เป็นเป้าหมายการแสวงกำไร
Output Text


CU Thai Segmentation 2.1
File Edit Option Run Windows Help
Input Text
ในกรณี "สื่อสาธารณะ" ที่มีการบริหารจัดการจากภาษีของประชาชนซึ่งไม่จำเป็นต้องพึ่งรายได้จากโฆษณาแต่อย่างใด จึงต้องรักษาความเป็นอิสระในการทำหน้าที่สื่อมวลชนเพื่อประโยชน์ที่สาธารณะพึงได้รับโดยนำเสนอข้อมูลข่าวสารอย่างมีคุณภาพ ตระหนักถึงสถานะความเป็น"พลเมือง"ของประชาชน ไม่ใช่ "ลูกค้า" ที่เป็นเป้าหมายการแสวงกำไร
Output Text
<p>
<p><w>ใน<w>กรณี<w><s><w>"<w>สื่อ<w>สาธารณะ<w>"<w><s><w>ที่<w>มี<w>การ<w>บริหาร<w>จัดการ<w>จาก<w>ภาษี<w>ของ<w>ประชาชน<w>ซึ่ง<w>ไม่<w>จำเป็น<w>ต้อง<w>พึ่ง<w>รายได้<w>จาก<w>โฆษณา<w>แต่อย่างใด<w><s><w>จึง<w>ต้อง<w>รักษา<w>ความ<w>เป็น<w>อิสระ<w>ใน<w>การ<w>ทำหน้าที่<w>สื่อมวลชน<w>เพื่อ<w>ประโยชน์<w>ที่<w>สาธารณะ<w>พึง<w>ได้รับ<w>โดย<w>นำ<w>เสนอ<w>ข้อมูล<w>

Part of Speech in Thai (1)

- Main problematic issue
- Characteristics of Thai
  - Isolating language, no morphological marking
  - Compounding and reduplication as main morphological processes
  - Serial verb construction
  - Zero pronoun language
  - Prevalent homonymy & polysemy

Part of Speech in Thai (2)

- Homonymy:
  - Group of words sharing same spelling and same pronunciation but have different meanings
  - e.g. ***bank*** (of a river); ***bank*** (financial institution)
- Polysemy
  - Word or phrase with ***different but related*** senses.
  - Different but related meanings/functions of a given form associated with distinct grammatical contexts
  - e.g. walk (N, V)

Part of Speech in Thai (3)

- Grammaticalization as important factor of polysemy in Thai
- Common polysemous categories in Thai
  - Verbs → prepositions เหมือน mUan4 'resemble/as'
  - Verbs→ adverbs ไป pay0 'go/ away (direction)'
  - Nouns → prepositions เวลา wee0-laa0 'time/when'
  - Nouns → adverbs หลัง laN4 'back/ afterwards'

Part of Speech in Thai (4)

- Unresolved issue in Thai
- Reference criteria used for determining POS in this project:

Amara Prasithrathsint (2010). Parts of Speech in Thai: a Syntactic Analysis.

- 1998-2000 study based on 2 million word Chula corpus (precursor of TNC)
- Identified 152 high frequency words that are problematic in determining POS
- Use 2 main syntactic criteria
  1. Word (positional) distribution
  2. Co-occurrence

Part of Speech in Thai (5)

proposed 8 word classes

	POS		
1	VERB (V)	+[may2__]	
2	NOUN (N)	-[may2__]	+[__V], +{V__], +[P__], +[__AJ]
3	ADJECTIVE (AJ)	-[may2__]	+ [N__]
4	PREPOSITION (P)	-[may2__]	{+[__N] + [__V]}
5	ADVERB (AV)	-[may2__]	{+[__V] + [V__]}
6	CONJUNCTION (C)	-[may2__]	{+[N__N] + [V__V] + [P__P] }
7	QUANTIFIER (Q)	-[may2__]	+[__N], -[V__], - [__V], -[P__]
8	PARTICLE (PT)	-[may2__]	+[//__], + [__#]

ไม่ may2 = negation

Part of Speech in Thai (6)

10 verb subcategories

		example	
1.1	<__#>	ยิ้ม /yim3/	smile
1.2	<__N>	ขาย /khaay4/	sell
1.3	<__N(COR)>	เป็น /pen0/	be
1.4	<__N(LOC)>	ไป /pay0/	go
1.5	<__V>	ต้อง /tOON2/	must
1.6	<__V(hay2)V>	แนะนำ /nE3nam0/	advise
1.7	<__P(waa2)V>	ชม /chom0/	praise
1.8	<__P(thii2)V>	ดีใจ /dii0cay0/	glad
1.9	<__P N>	ประกอบ /pra0kOOp1/	consist
1.10	<__kan0>/<__kap1 N>	ทะเลาะ /tha0lO3/	argue

Part of Speech in Thai (7)

- Verbs include "adjectival verbs" (attributive) e.g. ดี dii0 'good'
- Nouns include pronouns, classifiers and numbers.
  - Nouns can modify other nouns e.g. sea-food
- Limited number of "Adjectives"
  - Determiners (นี้ nii3 'this'; นั้น nan3 'that')
- Preposition phrase must be 'exocentric construction'
- Many adverbs are polysemous with (motion) verbs e.g. ไป pay0 'go'; มา maa0 'come'
  - Discourse markers are adverbs: e.g. อย่างไรก็ตาม yaaN1-ray0-kO2-taam0 'however'

Part of Speech in Thai (8)

- A preliminary proposal
- Remaining issues
  - Ambiguity between verbs & adverbs due to serial verb construction
  - **คือ** khUU0 'BE' does not fit into any category according to the criteria

A PROGRESS REPORT

- 100+2 raw high frequency word forms extracted from TNC corpus in July 2011 and March 2012
- corpus size: 30 million words
- 102 word forms
- 60 word forms are on Amara's list of problematic 152 words.

Initial observations (1)

Potential content nouns: 16

Location: ที่ thii2 place, ทาง thaaN0 way, ประเทศ pra0theet2 country

Time: เวลา wee0laa0 time, วัน wan0 day, ปี pii0 year

People: คน khon0 person, นาย naay0 boss, เด็ก dek1 young person

ผม phom4

Thing: ของ khOON4 thing,

Misc. แบบ bEEp1 form, model อย่าง yaaN1 type ส่วน suan1 part

งาน Naan0 work,event หน้า naa2 face เรื่อง rUaN2 story

Proper N ไทย thay0 Thai

Number หนึ่ง nUN1 one, สอง sOON4 two

Potential Pronouns: (6)

เรา raw0 we, ผม phom4 I (m.), ฉัน chan4 I (neutral, เธอ th@0 she, you, มัน man0 it, กัน kan0 each other

Initial Observations (2)

Potential Verbs 31 + function verbs 5

- **motion** (11)

ไป pay0 go, มา maa0 come, จาก caak1 leave, ขึ้น khUn2 rise, ตาม taam0 follow, เข้า khaw2 enter, ถึง thUN4 arrive, ออก ?OOk1 exit (out), เลย l@@y0 beyond, ลง lON0 down, กลับ klap1 return,

- **speech** (2)

ว่า waa2 say, บอก bOOk1 tell

- **Cognition** (3)

เห็น hen4 see, ดู duu0 look, รู้ ruu3 know

- **Attributive (Adjectival)** 5

ดี dii0 good, ถูก thuuk1 cheap, จน con0 poor, บาง baaN0 thin, เด็ก dek1 young, childish

- **Other 10**

ให้ hay2 give, คน khon0 stir, กัน kan0 prevent, ใช้ chay3 use, ทำ tham0 do, เอา ?aw0 take, นำ nam0 bring, เหมือน mUan4 resemble, ต่อ tOO1 connect, เกิด k@@t1 to be born (?)

- **Function verbs**
- **Copula** BE: เป็น pen0, อยู่ yuu1, 1, คือ khUU0
    - **Existential:** เกิด k@@t1

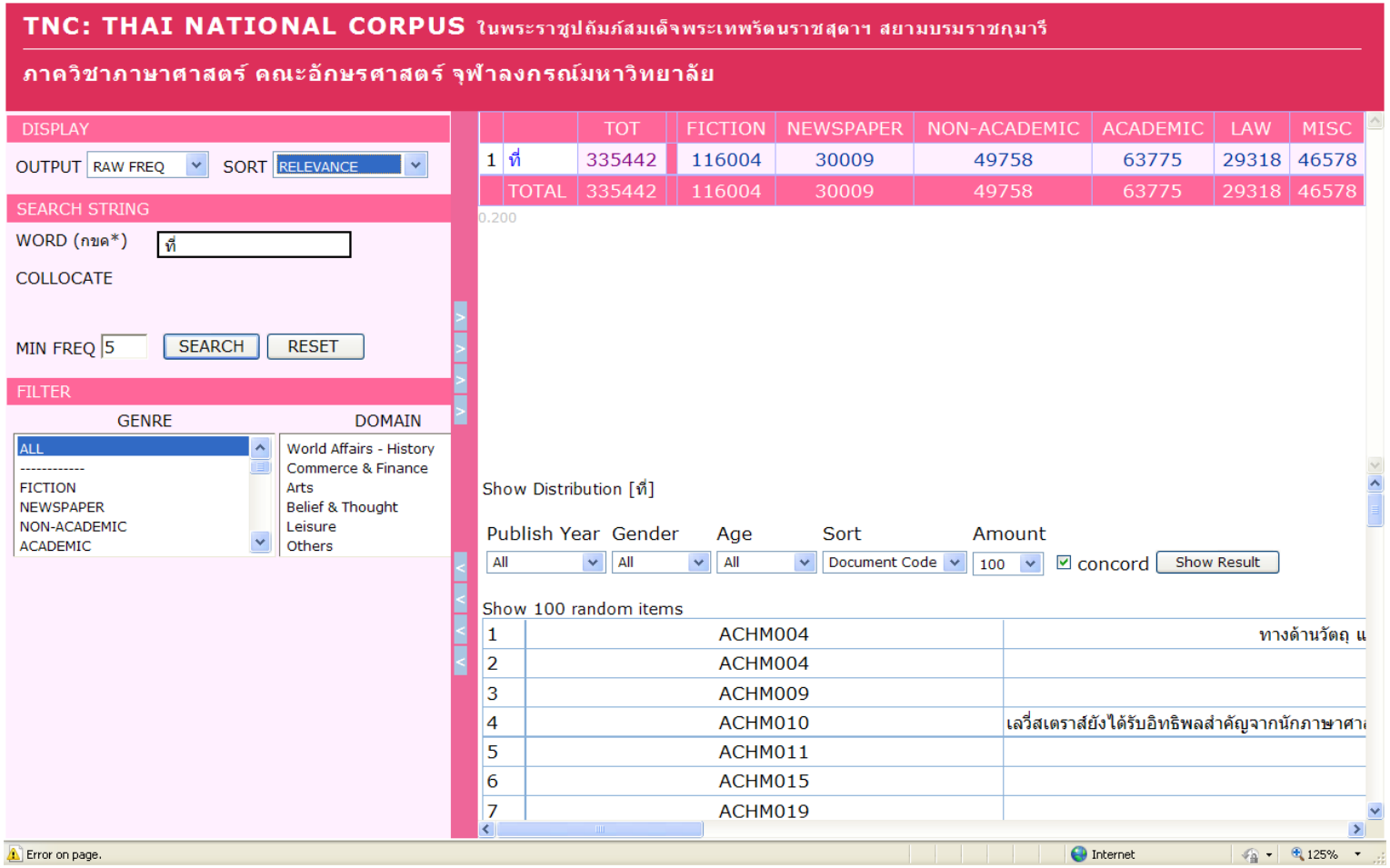
Particle: นะ **na3**

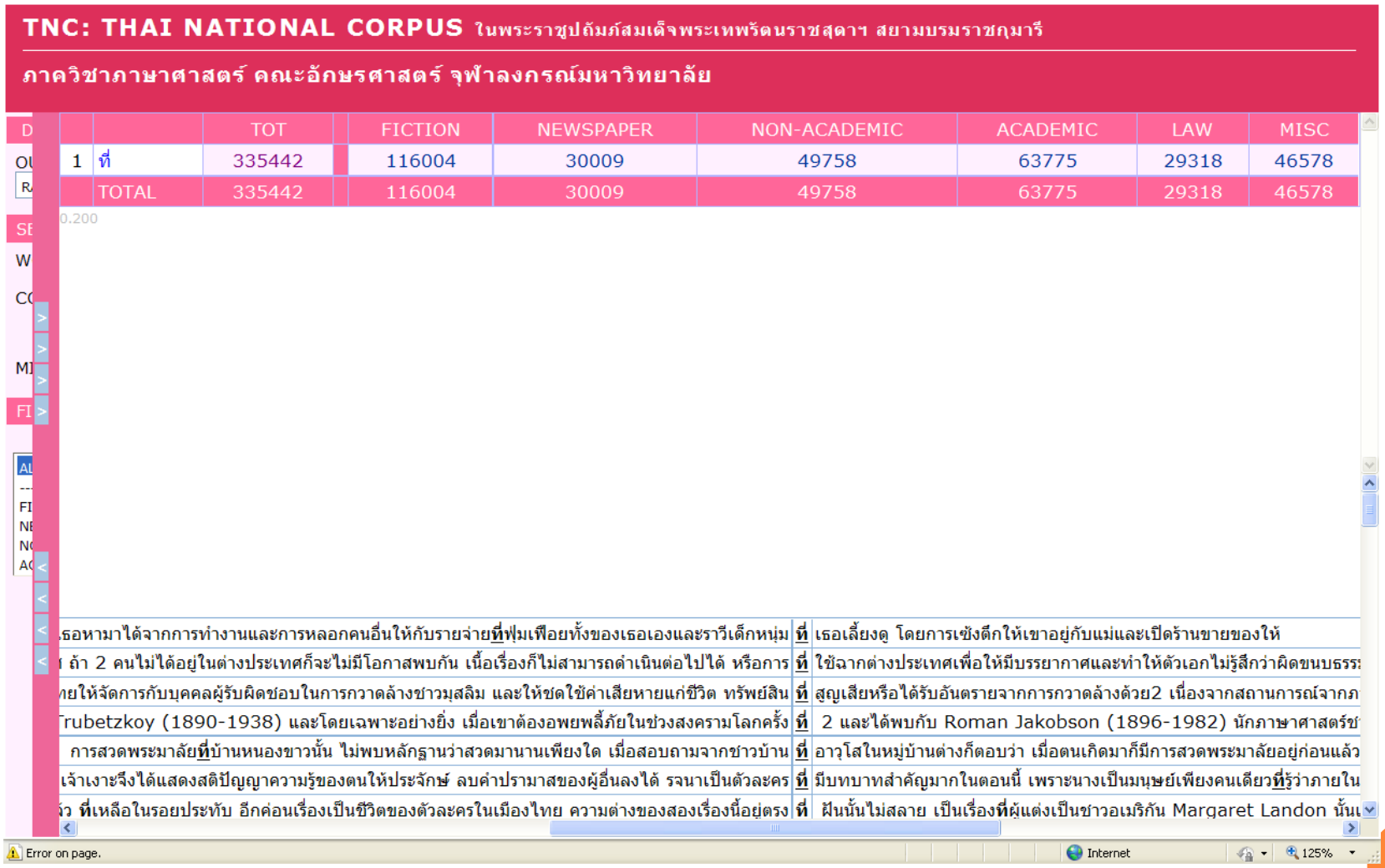
Statistics of actual usage of "word forms" (1)

- **Word function: cf. POS criteria of Amara + YH adjustment**
  - **Potential Polysemy: 48**
  - **Potential Homonymy: same POS 6, diff POS 9**
- **For polysemous forms, need to determine percentage of frequency of each POS**
  - **useful information for learners**
- **14 polysymous forms on the list have been topics of graduate degree theses at Thai dept. and linguistics dept. at Chulalongkorn university**

Statistics of actual usage of "word forms" (2)

- **Data: Concordances of 100 items per word form from TNC**
- **Analysis of percentage of occurrence of various functions (POS) per word form.**
- Polysemy-list-cropped.pdf


TNC: THAI NATIONAL CORPUS ในพระราชูปถัมภ์สมเด็จพระเทพรัตนราชสุดาฯ สยามบรมราชกุมารี
ภาควิชาภาษาศาสตร์ คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
DISPLAY
OUTPUT RAW FREQ SORT RELEVANCE
SEARCH STRING
WORD (กขค*) ที่
COLLOCATE
MIN FREQ 5 SEARCH RESET
FILTER
GENRE
ALL
FICTION
NEWSPAPER
NON-ACADEMIC
ACADEMIC
DOMAIN
World Affairs - History
Commerce & Finance
Arts
Belief & Thought
Leisure
Others
TOT FICTION NEWSPAPER NON-ACADEMIC ACADEMIC LAW MISC
1 ที่ 335442 116004 30009 49758 63775 29318 46578
TOTAL 335442 116004 30009 49758 63775 29318 46578
0.200
Show Distribution [ที่]
Publish Year Gender Age Sort Amount
All All All Document Code 100 concord Show Result
Show 100 random items
1 ACHM004 ทางด้านวัตถุ แ
2 ACHM004
3 ACHM009
4 ACHM010 เลวี่สเตราส์ยังได้รับอิทธิพลสำคัญจากนักภาษาศา
5 ACHM011
6 ACHM015
7 ACHM019
Error on page.
Internet
125%


TNC: THAI NATIONAL CORPUS ในพระราชูปถัมภ์สมเด็จพระเทพรัตนราชสุดาฯ สยามบรมราชกุมารี
ภาควิชาภาษาศาสตร์ คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
TOT FICTION NEWSPAPER NON-ACADEMIC ACADEMIC LAW MISC
1 ที่ 335442 116004 30009 49758 63775 29318 46578
TOTAL 335442 116004 30009 49758 63775 29318 46578
0.200
เธอหามาได้จากการทำงานและการหลอกคนอื่นให้กับรายจ่ายที่ฟุ่มเฟือยทั้งของเธอเองและราวีเด็กหนุ่ม ที่ เธอเลี้ยงดู โดยการเซ้งตึกให้เขาอยู่กับแม่และเปิดร้านขายของให้
ถ้า 2 คนไม่ได้อยู่ในต่างประเทศก็จะไม่มีโอกาสพบกัน เนื้อเรื่องก็ไม่สามารถดำเนินต่อไปได้ หรือการ ที่ ใช้ฉากต่างประเทศเพื่อให้มีบรรยากาศและทำให้ตัวเอกไม่รู้สึกว่าผิดขนบธรรม
ายให้จัดการกับบุคคลผู้รับผิดชอบในการกวาดล้างชาวมุสลิม และให้ชดใช้ค่าเสียหายแก่ชีวิต ทรัพย์สิน ที่ สูญเสียหรือได้รับอันตรายจากการกวาดล้างด้วย2 เนื่องจากสถานการณ์จากภ
Trubetzkoy (1890-1938) และโดยเฉพาะอย่างยิ่ง เมื่อเขาต้องอพยพลี้ภัยในช่วงสงครามโลกครั้ง ที่ 2 และได้พบกับ Roman Jakobson (1896-1982) นักภาษาศาสตร์ช
การสวดพระมาลัยที่บ้านหนองขาวนั้น ไม่พบหลักฐานว่าสวดมานานเพียงใด เมื่อสอบถามจากชาวบ้าน ที่ อาวุโสในหมู่บ้านต่างก็ตอบว่า เมื่อตนเกิดมาก็มีการสวดพระมาลัยอยู่ก่อนแล้ว
เจ้าเงาะจึงได้แสดงสติปัญญาความรู้ของตนให้ประจักษ์ ลบคำปรามาสของผู้อื่นลงได้ รจนาเป็นตัวละคร ที่ มีบทบาทสำคัญมากในตอนนี้ เพราะนางเป็นมนุษย์เพียงคนเดียวที่รู้ว่าภายใน
ว ที่เหลือในรอยประทับ อีกค่อนเรื่องเป็นชีวิตของตัวละครในเมืองไทย ความต่างของสองเรื่องนี้อยู่ตรง ที่ ฝันนั้นไม่สลาย เป็นเรื่องที่ผู้แต่งเป็นชาวอเมริกัน Margaret Landon นั้น
Error on page.
Internet
125%

Statistics of actual usage of "word forms" (3)

- Analysis of percentage of occurrence of various functions (POS) per word form.
- Polysemy-list-cropped.pdf

Statistics of actual usage of "word forms" (4)

- Issues

  1) word segmentation

  - The form is a single lexical item (morpheme) VS part of a lexical item (morpheme)

the string สิ่งที่ siN2 thii2 → siN2 thii2 (thing that)

VS the string ทั้งๆที่ thaN3 thaN3 thii2 (in spite of, despite)

TNC (which uses collocation strength algorithm) list ทั้งๆที่ as two words.

Need to improve algorithm for measuring collocational strength bwteen syllables

  - causative forms in Thai:

    ทำ tham0

    ให้ hay2

    ทำให้ tham0 hay2

    Amara treats ให้ in ทำให้ as V[Comp], an argument of ทำ

    But TNC treats ทำให้ as a lexical unit

- **The decision will affect word frequency count**

ISSUES (CONT.)

2) Determining Word class (POS) using Amara's proposed criteria

- ambiguity in making distinction between V/AV for motion verbs in post main verb position. Subjectivity if POS assignment is determined by semantic
- Need more fine grained criteria for determining subcategories of main word classes in order to get a better picture of actual frequency of function/usage of polysemous forms.

ที่	thii2	place	N, P[Loc], P[Comp], P[Rel]	P[Rel] 65 P[Loc] 16 P[Comp] 14 N 3

Issues (cont.)

- Cf. Word classes in COBUILD dictionary for learners, based on the BNC corpus

- potential direction of further development of TNC corpus

****