# Corpora and Thai Language Studies

Wirote Aroonmanakun
Dept.of Linguistics, Chulalongkorn University
ICTL2006, Imperial Queen's Park, Nov 10-12 2006

## What is a corpus?

- a collection of written or spoken texts (Oxford)
- a large collection of written or spoken language, that is used for studying the language (Longman)
- the collection of a single writer's work or of writing about a particular subject, or a large amount of written and sometimes spoken material collected to show the state of a language (Cambridge)
- A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research. (Sinclair, 2005)

## Why need a corpus?

- Answer questions intuition alone cannot do.
  - Frequency of usages น้อย >> เล็ก เล็ก >> น้อย
- Data of analysis
  - Lexicographer : understand different senses clearly
  - Linguists : linguistic analysis
- Resources
  - Examples of usage for language learning/teaching
  - Identify problems, understand better,
  - Create a better syllabus

## Statistics of Thai Language

- Frequency of characters (pdf)
- Frequency of Thai consonants
- Frequency of Thai vowels and special characters
- 1000 most frequent word (from newspaper)
  - Lexical syllabus
- Search for specific patterns
  - *ใจ : ใจกลาง, ใจกว้าง, ใจขุ่น, ใจแข็ง, ใจความ, ใจคอ, ใจแคบ, ใจง่าย, ใจจดใจจ่อ, ใจจืด, ใจเฉื่อย, ใจชื้น, ใจดำ, ใจดี, ใจเด็ด, ใจเดียว, ใจเดียวกัน, ใจต่ำ, ใจเติบ, ใจแตก, ใจโต, ใจถึง, ใจน้อย, ใจบาน, ใจบุญ, ใจเบา, ใจปลาซิว, ใจป้ำ, ใจแป้ว, ใจฝ่อ, ใจเพชร, ใจมา, ใจมือ, ใจไม่ดี, ใจไม้ไส้ระกำ, ใจยักษ์, ใจเย็น, ใจร้อน, ใจร้าย, ใจเร็ว, ใจลอย, ใจสูง, ใจเสาะ, ใจเสีย, ใจหนักแน่น, ใจหาย, ใจเหี่ยวแห้ง, ใจใหญ่ใจโต, ใจอ่อน;

- *ใจ* : กลางใจเมือง, ข้อใหญ่ใจความ, ใจจดใจจ่อ, ใจใหญ่ใจโต, ถอนใจใหญ่, ถอยใจใหญ่, น้ำใจใคร่, นิสัยใจคอ, หน้าใหญ่ใจโต, เอาใจช่วย, เอาใจใส่;
- *ใจ : กี่ตามใจ, กระเทือนใจ, กริ่งใจ, กลั้นใจ, กลุ้มใจ, กลุ้มอกกลุ้มใจ, กำลังใจ, กินใจ, เกรงใจ, เกรงอกเกรงใจ, ข่มขืนใจ, ข่มใจ, ขวยใจ, ขวัญใจ, ข้องใจ, ข้อเตือนใจ, ขอบใจ, ขัดใจ, ขาดใจ, ขึ้นใจ, ขืนใจ, ขุ่นข้องหมองใจ, ขุ่นใจ, เขี่ญใจ, เข้าใจ, แข็งใจ, ไข้ใจ, ความจริงใจ, ความภูมิใจ, ความมั่นใจ, ความสนใจ, คับใจ, คับอกคับใจ, คู่ใจ, แคลงใจ, จงใจ, จริงใจ, จรุงใจ, จอมใจ, จับใจ, จำใจ, จิตใจ, จุใจ, จูงใจ, เจนใจ, เจ็บใจ, เจ็บช้ำน้ำใจ, เจริญตาเจริญใจ, ฉุกใจ, เฉลียวใจ, ชอบใจ, ชั่งใจ, ช่ำใจ, ช้ำใจ, ชุ่มใจ, ชูใจ, เชื่อใจ, ดูใจ, ดูวงใจ, ดีใจ, ดีเนื้อดีใจ, ดีอกดีใจ, ดึงดูดใจ, ดูใจ, ได้ใจ, ตกใจ, ตกอกตกใจ, ตรอมใจ, ต้องใจ, ตั้งใจ, ตั้งอกตั้งใจ, ตัดใจ, ตัดสินใจ, ตันอกตันใจ, ตามใจ, ตามอำเภอใจ, ตายใจ, ต่ำใจ, ติดใจ, ตื่นใจ, ตื้นตันใจ, ตื่นตาตื่นใจ, เต็มใจ, เต็มอกเต็มใจ, เตือนใจ, ถนัดใจ, ถลำใจ, ถึงใจ, ถือใจ, ถูกใจ, ถูกอกถูกใจ, ทะยานใจ, ทันใจ, ทำใจ, แทงใจ, นอกใจ, นอนใจ, น้อยใจ, น้อยเนื้อต่ำใจ, น่าตกใจ, น่าสนใจ, น้ำใจ, แน่ใจ, บริสุทธิ์ใจ, บังคับใจ, บาดใจ, เบาใจ, ประทับใจ, ประหลาดใจ, ปลงใจ, ปลอบใจ, ปล่อยใจ, ปล่อยตัวปล่อยใจ, ปลุกใจ, ปักใจ, ปันใจ, เป็นใจ, เปลี่ยนใจ, เปลืองใจ, เปิดใจ, แปลกใจ, ผิดใจ, ผิดพ้องหมองใจ, ฝังใจ, ใฝ่ใจ, พร้อมใจ, พอใจ, พิมพ์ใจ, พึงใจ, พึงพอใจ, เพลียใจ, ภาคภูมิใจ, ภูมิใจ, มั่นใจ, มิตรจิตมิตรใจ, มีแก่ใจ, มุ่นใจ, ย้อมใจ, ยับยั้งชั่งใจ, ยาใจ, ยินใจ, ยุ่งใจ, เย็นใจ, ร่วมใจ, รู้เห็นเป็นใจ, เร้าใจ, แรงคลใจ, ลมหายใจ, ลองใจ, ลำบากใจ, วางใจ, ไว้ใจ, ไว้เนื้อเชื่อใจ, ไว้วางใจ, สนใจ, สนัดใจ, สบายใจ, สมัครใจ, สองจิตสองใจ, สองใจ, สะใจ, สะดุดใจ, สะท้อนใจ, สะเทือนใจ, สายใจ, สำรวมใจ, สิ้นใจ, สินน้ำใจ, สุขใจ, สุดใจ, สุดสวาทขาดใจ, เสมอใจ, เสียกำลังใจ, เสียใจ, เสียน้ำใจ, แสลงใจ, ใส่ใจ, หนักใจ, หนาวใจ, หน่ำใจ, หมองใจ, หมางใจ, หมายใจ, หย่อนใจ, หลากใจ, หักใจ, หักอกหักใจ, ห่วงใจ, หายใจ, เห็นใจ, เหลือใจ, แหนงใจ, อดใจ, อ่อนจิตอ่อนใจ, อ่อนใจ, อ่อนอกอ่อนใจ, อัดอั้นตันใจ, อำเภอใจ, อิดหนาระอาใจ, อิ่มใจ, อิ่มอกอิ่มใจ, อึดใจ, อุ่นใจ, เอะใจ, เอาใจ, เอาแต่ใจ, เอาอกเอาใจ;

# เรื่อง “เล็ก” “น้อย” ในภาษาไทย

- From Royal dictionary
- ขนาด
  - เล็ก น้อย
- ปริมาณ
  - น้อย
- ไม่สำคัญ
  - เล็ก น้อย เล็กน้อย

**เล็ก** [เล็ก]

ว. มีขนาดย่อมกว่าเมื่อเทียบกัน เช่น ละมุดเล็กกว่าส้มโอ กล้วยไข่เล็กกว่ากล้วยหอม, มีขนาดไม่โต เช่น บ้านหลังนี้เล็ก, โดยปริยายหมายความว่า ไม่สำคัญ, สำคัญน้อยกว่า, เช่น เรื่องเล็ก.

**น้อย** ๑ [น้อย]

ว. ไม่มาก (ใช้เกี่ยวกับปริมาณ) เช่น มีเงินน้อย พูดน้อย, โดยปริยายหมายถึงลักษณะที่ไม่บริบูรณ์ เช่น น้ำน้อย ฝนน้อย.

ว. เล็ก (ใช้เกี่ยวกับขนาด) เช่น เรือน้อย.

ว. โดยปริยายหมายถึงลักษณะที่ไม่สำคัญ เช่น ครูน้อย ผู้น้อย เณรน้อย, เกี่ยวกับความรู้สึกเป็นไปในทางน่ารักน่าเอ็นดู เช่น เด็กน้อย น้องน้อย สาวน้อย หนูน้อย.

**เล็กน้อย** [เล็ก-น้อย]

ว. นิดหน่อย, ไม่สำคัญ.

# “เล็ก” and “น้อย”

Questions

- Beside the use of เล็ก for size and น้อย for quantity, are there any other differences?
- When both are used for size, are they the same?
- When used as “unimportant”, are there any differences between เล็ก น้อย? Is it the same as เล็กน้อย?
- reduplication เล็กๆ, น้อยๆ, เล็กๆน้อยๆ is used to intensify meaning like other words?
- Antonym of เล็ก is ใหญ่? Antonym of น้อย is มาก?
- What else?

# “เล็ก” and “น้อย”

Procedure

- Use Thai Concordance Online http://www.arts.chula.ac.th/~ling/ThaiConc/
- Extract 500 samples of เล็ก, and น้อย (demo)
- Categorize collocate of เล็ก and น้อย
- Search for เล็กน้อย เล็กๆน้อยๆ in the corpus
- Search dictionary for compounding of เล็ก น้อย

# Finding of เล็ก and น้อย

Basic fact :

- น้อย is used more than เล็ก about 2 times (น้อย 614, เล็ก 306 times per million words)
- Compounding น้อย >> เล็ก
  เล็ก => มหาดเล็ก, ทับเล็ก(แมลง)
  น้อย => อย่างน้อย, ชนกลุ่มน้อย,น้อยหน้า, น้อยใจ,ผู้น้อย,ไทยน้อย,นวลน้อย(หญ้า),ใจน้อย, ท้องน้อย,น้อยหน่า,แน่งน้อย,น้อยโหน่ง(พุ่มไม้),ตายน้อย(=เกือบตาย),รากสาดน้อย,กล้วยน้อย (ต้นไม้),กองทัพน้อย,มักน้อย,หมอน้อย(ปลิง),เครื่องทองน้อย,ข้าวใหม่น้อย(ไข้),ธงเยาวราชน้อย, สาวน้อยเล่นน้ำ(เพลงไทย),ธงมหาราชน้อย,ฉากน้อย(ท่าละคร),ธงบรมราชวงศ์น้อย,ธงราชินี น้อย,ฉลองพระกรน้อย,ขอคำน้อย(ว่าน),ข้าน้อย,ข้าหลวงน้อย,คูณร่วมน้อย,ชนหมู่น้อย,น้อยแง่, แอกน้อย,เมียน้อย

# Finding of เล็ก and น้อย

- found another reduplication xเล็กxน้อย
  ปลาตัวเล็กตัวน้อย คิดเล็กคิดน้อย เก็บเล็กผสมน้อย ตาเล็กตาน้อย
  some are used as an idiom

Corpus => confirm เล็ก for 'size' , น้อย for 'quantity'

- anything conceptualized by size
  บริษัทเล็ก ปลาเล็ก คนตัวเล็ก รถคันเล็ก บริษัทเล็ก ธุรกิจขนาดเล็ก ผ้าผืนเล็ก โครงสร้างเล็กลง
- anything conceptualized by quantity
  คนส่วนน้อย รายได้น้อย ซับซ้อนน้อยกว่า รู้น้อย อายุน้อย กินไฟน้อย โกรธน้อยลง พึงพอใจน้อย ประสบการณ์น้อย น่าเชื่อถือน้อย

# Finding of เล็ก and น้อย

Differences between เล็ก and น้อย

- When น้อย is used for 'size', it conveys speaker's attitude
  เด็กน้อย หนูน้อย สาวน้อย หนุ่มน้อย ช้างน้อย
  เด็กน้อย vs เด็กเล็ก ช้างตัวน้อย vs ช้างตัวเล็ก
  น้อย is found in some epithets เจ้าสัวน้อย เจ้าเสือน้อย
- น้อย used for size is found with ใหญ่ e.g. บ่อนน้อยใหญ่ ร้านค้าน้อยใหญ่ ธุรกิจใหญ่น้อย
- น้อย occurs with ไม่ ไม่ + น้อย = มาก (มีไม่น้อย vs มีมาก)
  rarely found ไม่ + เล็ก = ใหญ่
  (เรื่องนี้ไม่เล็ก vs เรื่องนี้เรื่องใหญ่)
  old language use น้อยหรือ เช่น น้อยหรือทำได้

# Finding of เล็ก and น้อย

เล็กน้อย = [1] unimportant (derived from intensifying of size), [2] intensifying of quantity

- เล็กน้อย = 'unimportant' e.g. งานเล็กน้อย เรื่องเล็กน้อย
  เล็ก = 'unimportant e.g. เรื่องเล็ก ผลงานชิ้นเล็ก
  น้อย = 'unimportant' e.g. ครูน้อย ?เมียน้อย compound?
- เล็กน้อย = intensify 'quantity'
  เล็กน้อย is related to น้อย than เล็ก, may be replaced by น้อย
  มีจำนวนเล็กน้อย มีจำนวนน้อย *มีจำนวนเล็ก
- เล็กน้อย is used for บาดเจ็บ เสียหาย
  but not เล็ก; น้อย quite possibly
  เสียหายเล็กน้อย ?เสียหายน้อย *เสียหายเล็ก,
  บาดเจ็บเล็กน้อย ?บาดเจ็บน้อย *บาดเจ็บเล็ก

# Finding of เล็ก and น้อย

- While เล็กน้อย related to น้อย, Xเล็กXน้อย is related to เล็ก
  ธุรกิจรายเล็ก ธุรกิจรายเล็กรายน้อย,
  เติบโตน้อย เติบโตเล็กน้อย,

Reduplication

- เล็ก => เล็กๆ น้อย => น้อยๆ intensify meaning
- น้อยๆ found << เล็กๆ , while น้อย found >> เล็ก
- Intensifying of size เล็กๆ => 'unimportant' meaning
  ข่าวเล็กๆ โครงการเล็กๆ สมาคมเล็กๆ เหตุการณ์เล็กๆ

# Finding of เล็ก and น้อย

- Intensifying of quantity น้อยๆ e.g. มือน้อยๆ คนกลุ่มน้อยๆ
  not => 'unimportant' meaning
  เล็กๆน้อยๆ is used instead e.g. ความขัดแย้งเล็กๆน้อยๆ รางวัลเล็กๆน้อยๆ
- เล็กๆน้อยๆ is an intensify of เล็กน้อย but they are not always replaceable
  เรื่องเล็กน้อย => เรื่องเล็กๆน้อยๆ
  ราคาสูงขึ้นเล็กน้อย => *ราคาสูงขึ้นเล็กๆน้อยๆ
  เทคนิคเล็กๆน้อยๆ => *เทคนิคเล็กน้อย

# Finding of เล็ก and น้อย

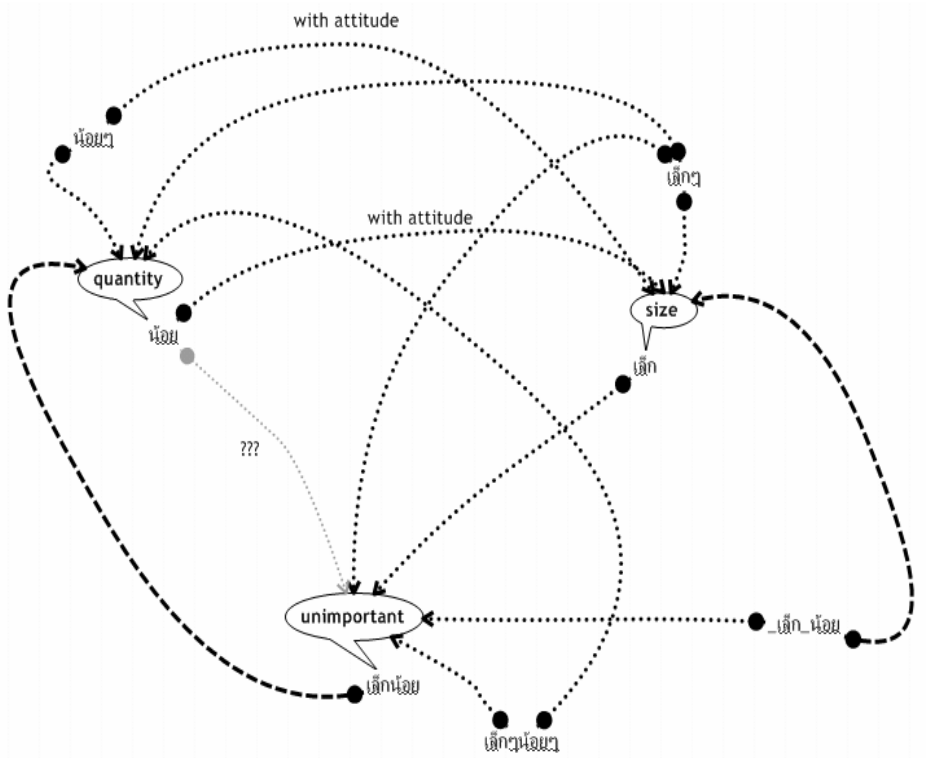
- เล็กน้อย เล็กๆน้อยๆ are replaceable if used in the same meaning
  - เอาเปรียบเล็กน้อย<quant> เอาเปรียบเล็กๆน้อยๆ<quant>
  - ข้อคิดเล็กน้อย<unimportant> ข้อคิดเล็กๆน้อยๆ<unimportant>
  - รางวัลเล็ก<quant> *รางวัลน้อย<quant> รางวัลน้อยๆ<quant> *รางวัลเล็กน้อย<quant> ?รางวัลเล็กน้อย<unimportant> รางวัลเล็กๆน้อยๆ <unimportant>
  - ทำงานเล็ก<unimportant> ทำงานน้อย<quant> ทำงานเล็กน้อย<quant> ?ทำงานเล็กน้อย<unimportant> ทำงานเล็กๆน้อยๆ <unimportant>
- X + เล็กน้อย/เล็กๆน้อยๆ => quant || unimportant

# Finding of เล็ก and น้อย

Use reduplication to change semantic category

- use เล็กๆ for things usually used with น้อย (quantity -> size)
  ความสุขน้อย -> ความสุขเล็กๆ, อิจฉาเล็กๆ
- Use น้อยๆ for things usually used with เล็ก (size -> quantity with attitude)
  มือเล็ก -> มือน้อยๆ, สวรรค์เล็กๆ -> สวรรค์น้อยๆ
- Prasithrathsint 2006 : content words can be reduplicated.
  When reduplicate noun, category is changed to verb
  e.g. ครู+ครู

| | เล็ก | น้อย | เล็กน้อย | _เล็ก_น้อย |
|---|---|---|---|---|
| Quantity | | จำนวนคนน้อย,เวลาน้อย,ยอดขายน้อย, อายุน้อย | Ok | No |
| State of mind | อิจฉาเล็กๆ,ความสุขเล็กๆ | โกรธน้อยลง,ความสามารถน้อย,ความพอใจน้อย,ความสุขน้อย, แรงบันดาลใจน้อย | Ok | No คิดเล็กคิดน้อย |
| Property | ?คุณค่าเล็กๆ | คุณค่าน้อย,จุดเด่นน้อย,สำคัญน้อย, | Ok | No |
| Consuming | | กินไฟน้อย,ใช้จ่ายน้อย,เติบโตน้อย, | Ok | No |
| Matter | เรื่อง (ไม่สำคัญ) | | Ok | Ok |
| Trouble | ปัญหาเล็กๆ (ไม่สำคัญ) | ปัญหาน้อยใหญ่ | Ok (ไม่สำคัญ) | Ok |
| Business, organization | บริษัทเล็ก,ร้านค้ารายเล็ก,ธุรกิจขนาดเล็ก,มหาวิทยาลัยเล็ก,พรรคเล็ก,องค์กรเล็ก | บริษัทน้อยใหญ่, | No | Ok |
| Object size | รถคันเล็ก,คนตัวเล็ก,ผ้าผืนเล็ก, บอลลูนลูกเล็ก, ?ทารกเล็ก … | เด็กน้อย,ผ้าผืนน้อย,หนูน้อย,ช้างน้อย ต้นไม้น้อยใหญ่, ทารกน้อย | No | Ok |
| Damage | No | Ok | บาดเจ็บ,เสียหาย | No |
| Economic, price change | No | Ok | เพิ่มขึ้น,ลดลง,กระเตื้อง,หดตัว | No |




# Finding of เล็ก and น้อย

## specific patterns

- น้อย : ลดน้อยถอยลง มากน้อยแค่ไหน น้อยอกน้อยใจ น้อยเนื้อต่ำใจ บริษัทน้อยใหญ่ แม้แต่น้อย สิ่งละอันพันละน้อย ตัดช่องน้อยแต่พอตัว น้ำน้อยแพ้ไฟ น้ำบ่อน้อย รู้น้อยพลอยรำคาญ เบี้ยน้อยหอยน้อย
- เล็ก : ลูกเด็กเล็กแดง เล็กพริกขี้หนู ปลาใหญ่กินปลาเล็ก

# Finding of เล็ก and น้อย

## Statistical Collocation (software extracted)

- น้อย + X : กว่า, ลง, มาก, เพียงใด, แค่ไหน, ที่สุด, นิด, เกินไป, ๆ, เต็มที, ถอย (ลดน้อยถอยลง), ทีเดียว, ไป, อก (น้อยอกน้อยใจ), สุด, คน, เลย (ไม่น้อยเลย), อยู่, ราย
- เล็ก + X : ๆ, ลง, กว่า, ที่สุด, เกินไป, สุด, มาก,
- X + น้อย : ไม่, เมีย, มาก, ดาวเคราะห์, แม้แต่, สำนึก, มี, สาว, ลด, เสือ, ค่อนข้าง, ข้าง, บางกอก, อายุ, รายได้, จำนวน, เหลือ, กลุ่ม, ตัว, ส่วน, เจ้าสัว, หนุ่ม, ทีละ, มีอายุ, ความสำคัญ, เด็ก, ไม่ใช่, มด, พูด, กินไฟ, ยัง, หมี, ปริมาณ, โอกาส, นก, ความสนใจ, น้ำหนัก, ได้, ลูก, บทบาท, เวลา, เพียง, จำเป็น, ให้, อยู่, ทุน, ลงทุน
- X + เล็ก : ขนาด, เด็ก, ตัว, ละคร, ราย, สายการบิน, สกรีน, หอประชุม, ชิ้น, เครื่องบิน, รถ, โรง, ขั้ว, เกร็ด, เสี่ย, ร้าน, เล่ม, ประเทศ, ร่าง, หมู่บ้าน, ปลา, ครู, กลุ่ม, เม็ด, ตั้งแต่, เมือง, ตอน, นายก, เรื่อง, หลัง, โรงเรียน, คน, แบงก์, หนัง, ยัง, บริษัท,

# Finding of เล็ก and น้อย

**Antonyms**

- Found มาก ⇔ น้อย และ ใหญ่ ⇔ เล็ก
- มาก <-> น้อย : ผิดมากหรือน้อย มากน้อยแค่ไหน/เพียงใด มากบ้างน้อยบ้าง ผิดน้อยหรือมาก *ผิดน้อยมาก <either or>
- ใหญ่ <-> เล็ก : ต้นไม้เล็กใหญ่ ลำน้ำทั้งเล็กและใหญ่ เกาะเล็กใหญ่ หนูเหมียวเล็กใหญ่ทั้งสิบ ค่ายบันเทิงใหญ่เล็ก บ่อนใหญ่และเล็ก <all,can be both> ประเทศไม่เล็กไม่ใหญ่ <neither> รูปร่างใหญ่เล็กไม่สำคัญ ลูกบอลใหญ่เล็ก ความใหญ่เล็กของประเทศ <either or>

# Finding of เล็ก and น้อย

**Antonyms**

- Could also found น้อย + ใหญ่ for size
  บ่อนน้อยใหญ่ เกาะน้อยใหญ่ สัตว์น้อยใหญ่ หัวเมืองน้อยใหญ่ ขุนนางใหญ่น้อย พระตำหนักใหญ่น้อย ธุรกิจใหญ่น้อย สวนส้มใหญ่น้อย สุภาพสตรีใหญ่น้อย ดาราใหญ่น้อย
- Unlike ไม่น้อย,
  ?ไม่ใหญ่ไม่น้อย ?ไม่น้อยไม่ใหญ่ rarely found
- Found เล็ก + โต e.g. ตาเล็กตาโต

# Further questions

- Is เล็กใหญ่ different from ใหญ่เล็ก?
- Is ใหญ่น้อย different from น้อยใหญ่?
- มาก & ใหญ่ are use in a similar way to น้อย & เล็ก?
- What is the differences between ใหญ่ & โต?
- เล็ก and น้อย are equivalent to 'small' and 'little'?

Corpus : data-driven learning, journey of discovery,

# Want to use it now?

Use existing resources

- Thai Concordance Online (http://www.arts.chula.ac.th/~ling/ThaiConc/)

Use your own data

- collect data from the internet, save as texts
- install Thai Concordance software for windows
- if need word segmented, get word segmentation program, collocation extract program http://pioneer.chula.ac.th/~awirote/ling/wire.htm