การพัฒนาเครือข่ายคำไทย

วิโรจน์ อรุณมานะกุล

ภาควิชาภาษาศาสตร์

คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

การสัมมนา 10 ปี เล็กซิตรอน 25 ก.ย. 2551

Outline

- WordNet คืออะไร
- WordNet พัฒนาอย่างไร
- ประโยชน์ของ WordNet
- การพัฒนา WordNet ภาษาไทย

Wordnet คืออะไร

- WordNet is a large lexical database of English
  พัฒนาภายใต้การนำของ George A. Miller
  มหาวิทยาลัย Princeton
- Version 1.0 เผยแพร่ปี 1991 ปัจจุบัน version 3.0
- มีคุณลักษณะของ dictionary และ thesaurus
- Dictionary : เรียงคำตามตัวอักษร ให้ความหมายต่างๆ
  ข้อมูลไวยากรณ์ การใช้ ตัวอย่าง
- Thesaurus : จัดคำตามกลุ่มความหมาย แสดงโครงสร้าง
  ลำดับชั้นของมโนทัศน์ต่างๆ

plant[1] *noun* W1 S2

Menu | Word family | Word origin | Verb form | Word set

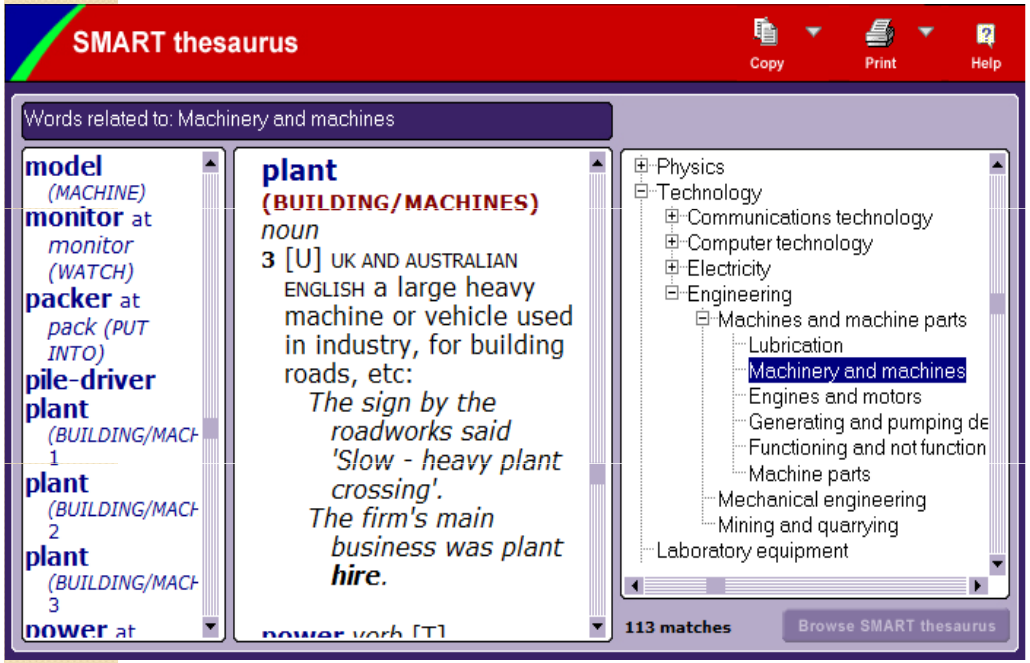
1 **LIVING THING** [countable] a living thing that has leaves and roots and grows in earth, especially one that is smaller than a tree:
Don't forget to water the plants.
a potato plant
*the forest's **plant life*** (=plants)
→ **HOUSEPLANT**

2 **FACTORY** [countable] a factory or building where an industrial process happens:
a huge chemical plant
→ **POWER PLANT**

3 **MACHINERY** [uncountable] *British English* heavy machinery that is used in industrial processes:
a plant hire business

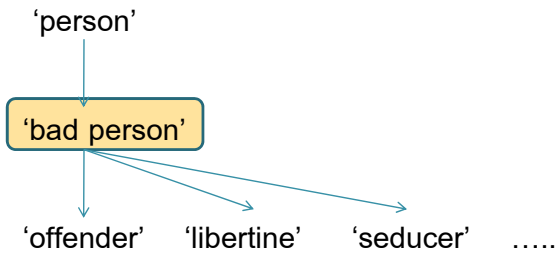
4 **SOMETHING HIDDEN** [countable usually singular] something illegal or stolen that is hidden in someone's clothes or possessions to make them seem guilty of a crime

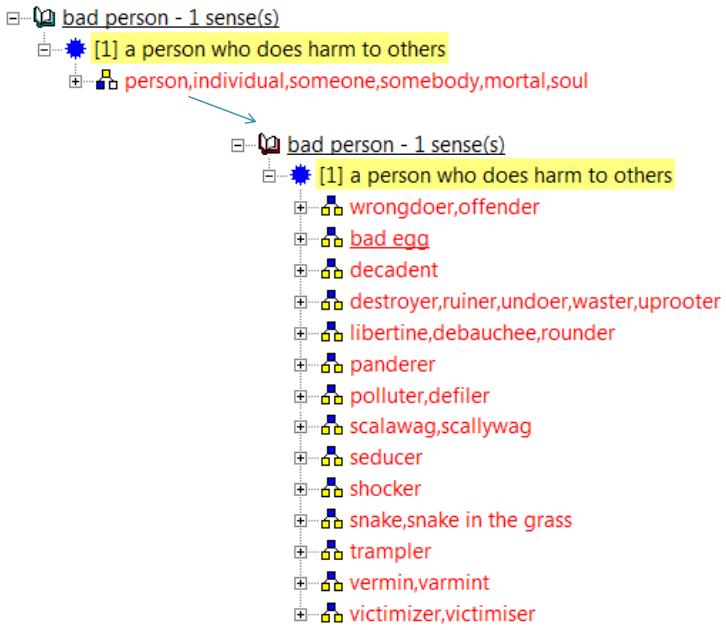
5 **PERSON** [countable] someone who is put somewhere or sent somewhere secretly to find out information



WordNet คืออะไร

- เหมือน thesaurus : หาคำที่มีความหมายแบบเดียวกันได้ ต่าง : เพิ่มการหาความสัมพันธ์ทางความหมายแบบต่างๆ, thesaurus มีเฉพาะ concept ที่เป็นคำ
- อาจมี วลี ในตำแหน่งที่เป็น lexical gap ในภาษา ภาษาอื่นอาจมีคำแทน concept นั้น ต.ย.





WordNet คืออะไร

- Wordnet = เครือข่ายคำเชื่อมโยงด้วยความสัมพันธ์ต่างๆ
- จัดกลุ่มคำพ้องความหมาย synonym เรียก synset {plant, implant, engraft, imbed} = 'to fix or set securely or deeply'
- ความสัมพันธ์ใน wordnet มี 2 แบบ conceptual กับ lexical
- Conceptual relation เชื่อมโยงระหว่าง synset เช่น hypernym, meronym, etc.
- Lexical relation เชื่อมโยงระหว่างคำ word เช่น ความสัมพันธ์ antonymy, synonymy

WordNet คืออะไร

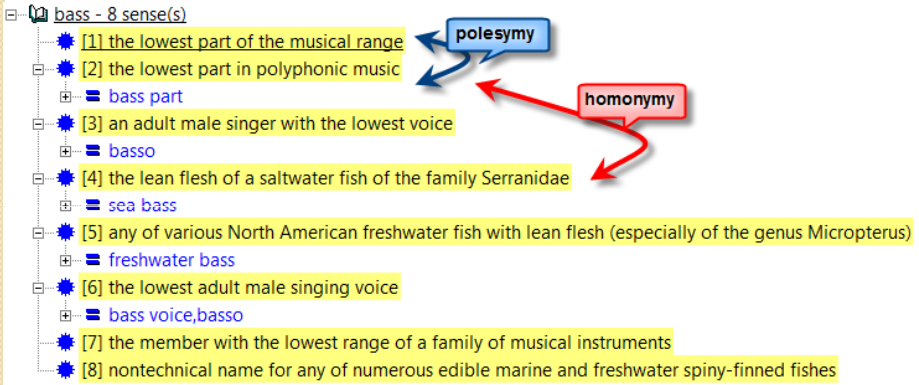
- ไม่มีการบอกความสัมพันธ์ระหว่างคำทางวากยสัมพันธ์ จึงแยกคำสี่หมวดจากกัน noun, verb, adjective, adverb
- จำนวนข้อมูลใน WordNet 3.0

POS	Unique Strings	Synsets	Total Word-Sense Pairs
Noun	117798	82115	146312
Verb	11529	13767	25047
Adjective	21479	18156	30002
Adverb	4481	3621	5580
Totals	155287	117659	206941

WordNet คืออะไร

- ไม่แยกความต่างระหว่าง polysemy กับ homonymy

ตย. bass1 - bass2 - bass4



WordNet คืออะไร

- Dictionary แยกความต่างระหว่าง polysemy กับ homonymy

bass[1] *noun*

Menu | Word family | Word origin | Verb form | Word set

1 [countable] a very low male singing voice, or a man with a voice like this

2 [singular] the part of a musical work that is written for a singer with a bass voice ➔ ALTO[2] (2), BARITONE[1] (2), SOPRANO[1] (2), TENOR[1] (2)

3 [uncountable] the lower half of the whole range of musical notes [↦ treble]

4 [countable] a BASS GUITAR:
The band features Johnson ***on bass*** (=playing the bass guitar).

5 [countable] a DOUBLE BASS

! Do not confuse **bass** and **base**, although they have the same pronunciation.

bass[2] *adjective*

bass[3] *noun*

WordNet คืออะไร

- Dictionary แยกความต่างระหว่าง polysemy กับ homonymy

bass[1] *noun*

bass[2] *adjective*

bass[3] *noun*

Menu | Word family | Word origin | Verb form | Word set

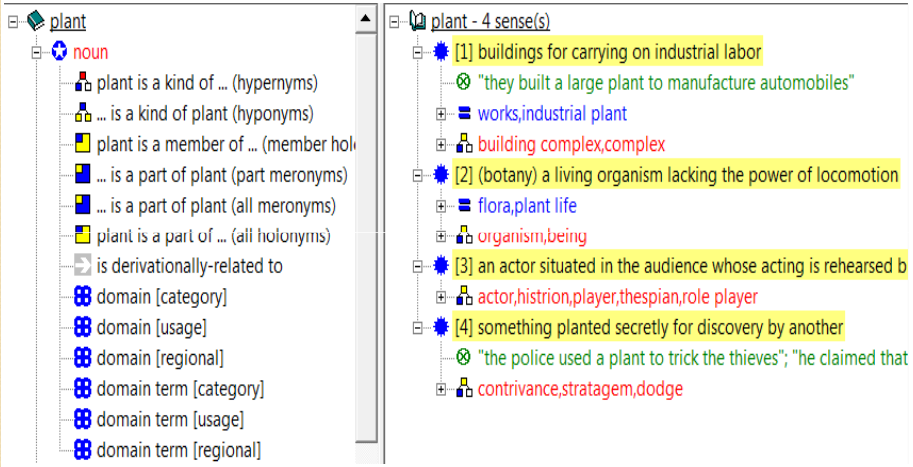
plural **bass** [countable]
a fish that can be eaten and lives in both rivers and the sea

bass clef *noun*

bass guitar *noun*

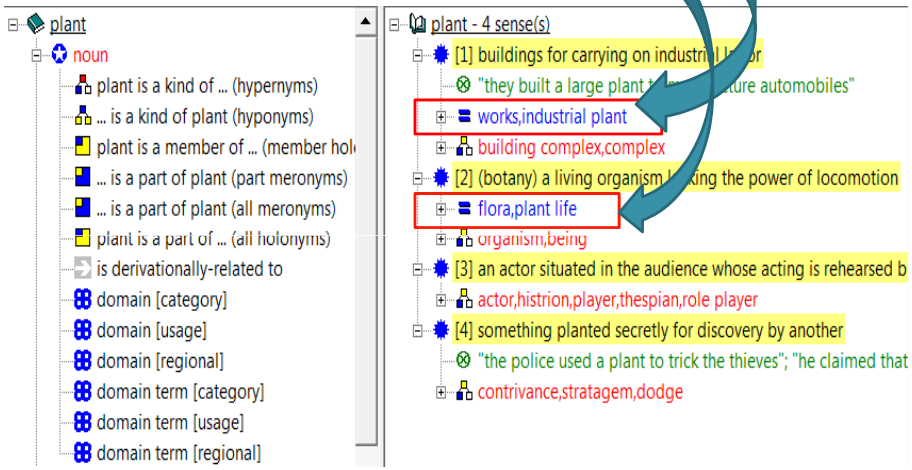
WordNet คืออะไร

- แต่ละ synset มีรายการคำที่เป็น synonym, gloss อธิบายความ, อาจมีตัวอย่าง, มีเชื่อมโยง synset อื่นๆ



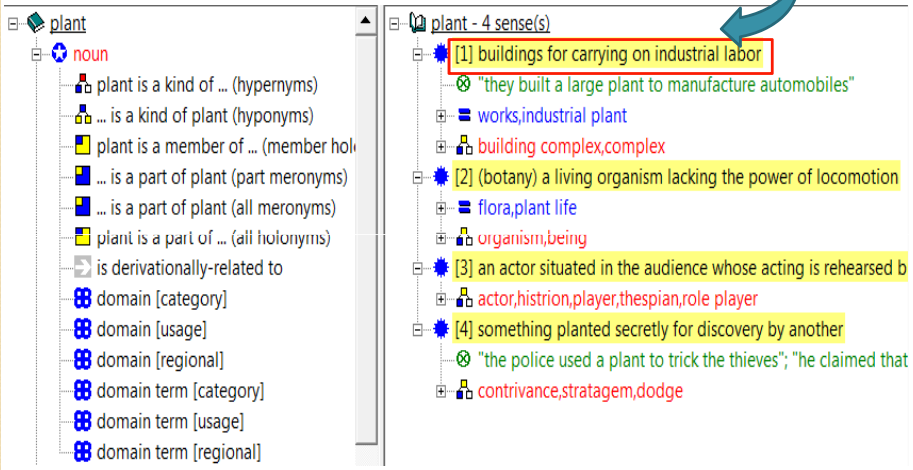
WordNet คืออะไร

- แต่ละ synset มีรายการคำที่เป็น synonym, gloss อธิบายความ, อาจมีตัวอย่าง, มีเชื่อมโยง synset อื่นๆ



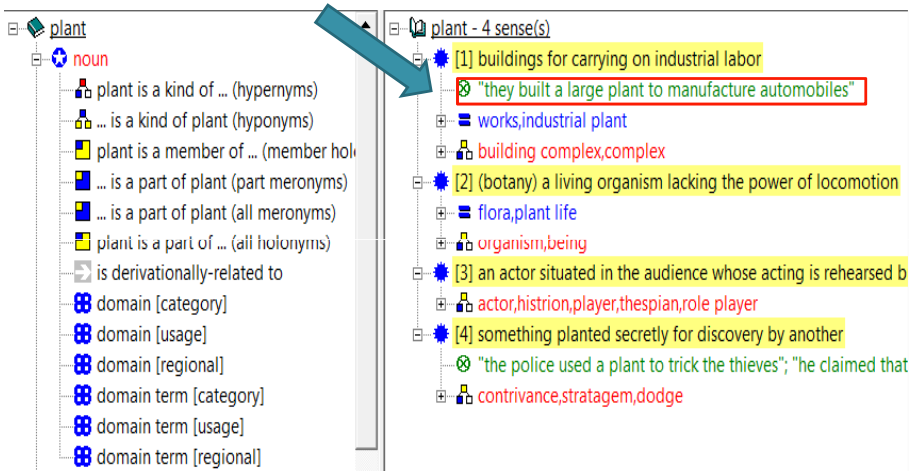
WordNet คืออะไร

- แต่ละ synset มีรายการคำที่เป็น synonym, gloss อธิบายความ, อาจมีตัวอย่าง, มีเชื่อมโยง synset อื่นๆ



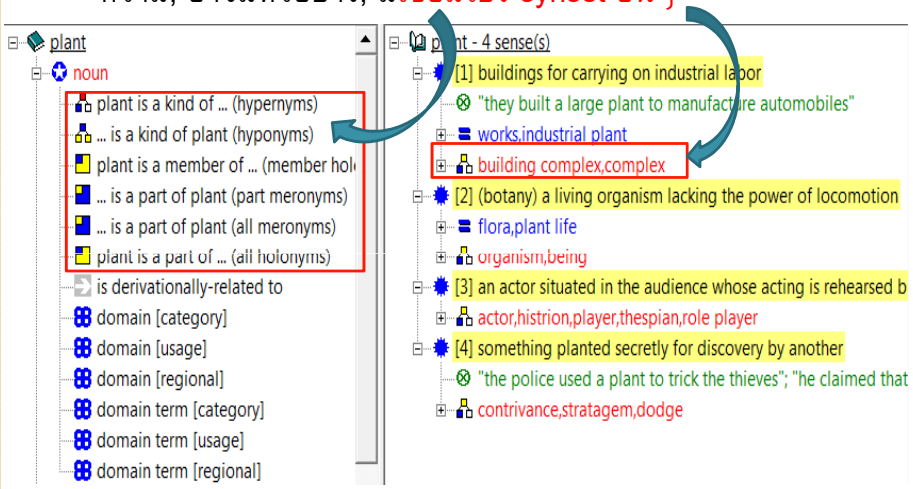
WordNet คืออะไร

- แต่ละ synset มีรายการคำที่เป็น synonym, gloss อธิบายความ, อาจมีตัวอย่าง, มีเชื่อมโยง synset อื่นๆ



WordNet คืออะไร

- แต่ละ synset มีรายการคำที่เป็น synonym, gloss อธิบายความ, อาจมีตัวอย่าง, มีเชื่อมโยง synset อื่นๆ



Wordnet คืออะไร

- ความสัมพันธ์ในกลุ่มคำนาม
  - hypernyms: Y เป็น hypernym ของ X ถ้า X เป็นประเภทย่อยของ Y (feline is a hypernym of cat)
  - hyponyms: Y เป็น hyponym ของ X ถ้า Y เป็นประเภทย่อยของ X (cat is a hyponym of feline)
  - holonym: Y เป็น holonym ของ X ถ้า X เป็นส่วนหนึ่งของ Y (building is a holonym of window)
  - meronym: Y เป็น meronym ของ X ถ้า Y เป็นส่วนหนึ่งของ X (window is a meronym of building)
  - ทำให้จัดมโนทัศน์เป็นลำดับชั้นได้ เป็นเหมือน ontology หรือ knowledge structure ที่มโนทัศน์ล่าง inherit property มโนทัศน์บนได้

Wordnet คืออะไร

- ความสัมพันธ์ในกลุ่มคำนาม
- Hypernym ใน WordNet ไม่ได้แยก
  - is_a_kind_of (taxonomic) : {robin} @-> {bird}
  - is_used_as_a_kind_of (functional) : {chicken} @-> {food}
- Meronymy ใน WordNet
  - Component part : {branch} #p -> {tree}
  - Member of : {tree} #m-> {forest}
  - Made from : {aluminum} #s-> {airplane}

Wordnet คืออะไร

- ความสัมพันธ์ในกลุ่มคำกริยา
  - hypernym: Y เป็น hypernym ของกริยา X ถ้า activity X เป็นประเภทหนึ่งของ Y (travel เป็น hypernym ของ walk)
  - troponym: Y เป็น troponym ของกริยา X ถ้า activity Y เป็นการทำ X ด้วยลักษณะอาการแบบหนึ่ง (march เป็น troponym ของ walk)
  - entailment: Y is entailed by X ถ้า ในการกระทำ X เราต้องทำ Y ด้วย (to sleep is entailed by to snore)
  - cause: การที่ Y cause X จะทำให้ Y entail X ด้วย เช่น show ทำให้เกิด see

Wordnet คืออะไร

- ความสัมพันธ์ในกลุ่มคำคุณศัพท์
  - antonymy: light != heavy
  - คำที่มีความหมายใกล้กันไม่มีคำตรงข้ามเหมือนกัน
    ไม่ใช่ {heavy, weighty, ponderous} != {light, weightless, airy}
    heavy!=light, weighty!=weightless, ponderous!=airy
  - similar: จัดคำเป็น cluster ที่มี semantic similarity
    ตย. fast, rapid, quick, prompt, swift มี slow เป็น antonym ของ fast และเป็น indirect antonym ของคำที่เหลือ
  - attribute: บอกความสัมพันธ์กับคำนาม
    fast เป็น attribute ของ speed

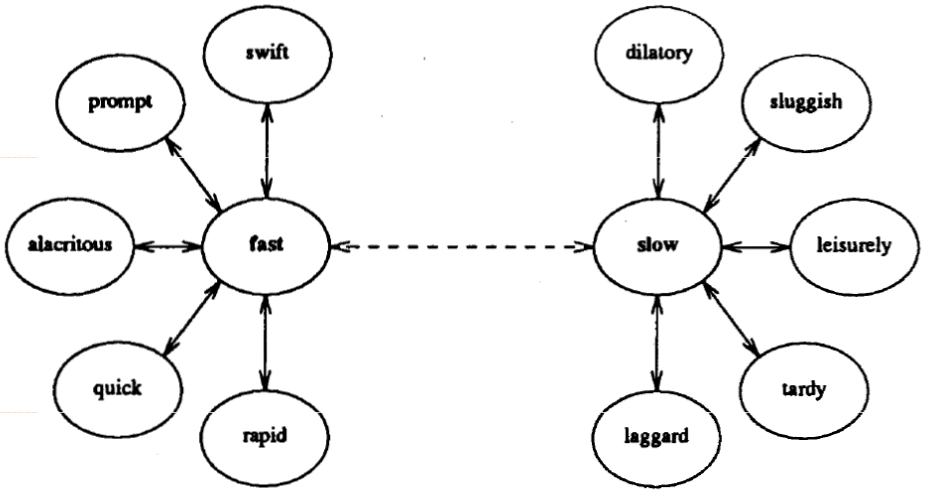
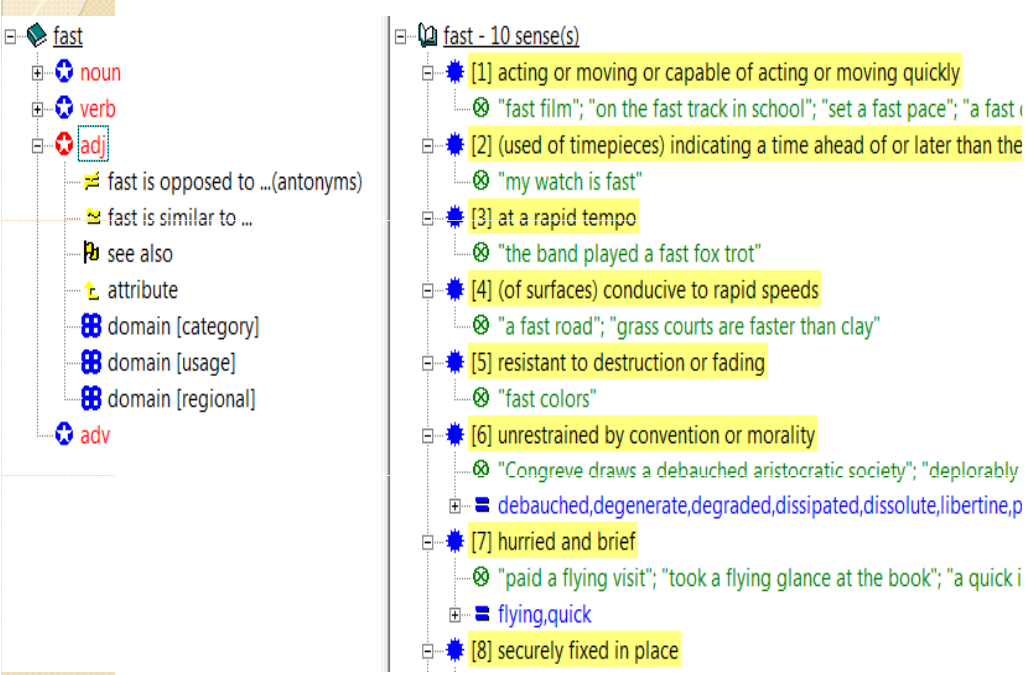
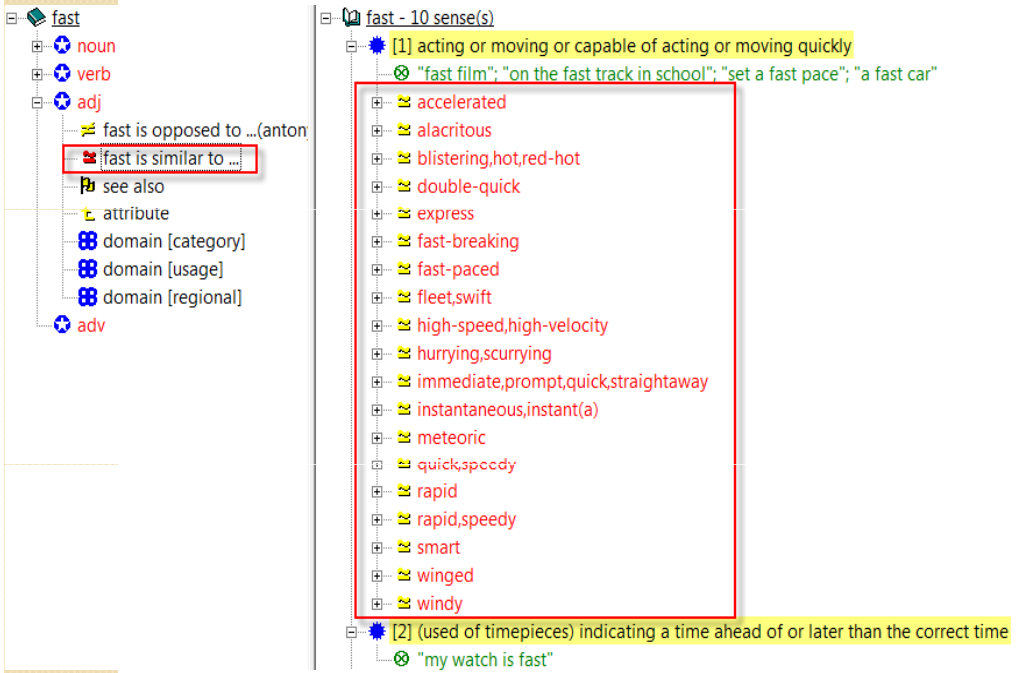
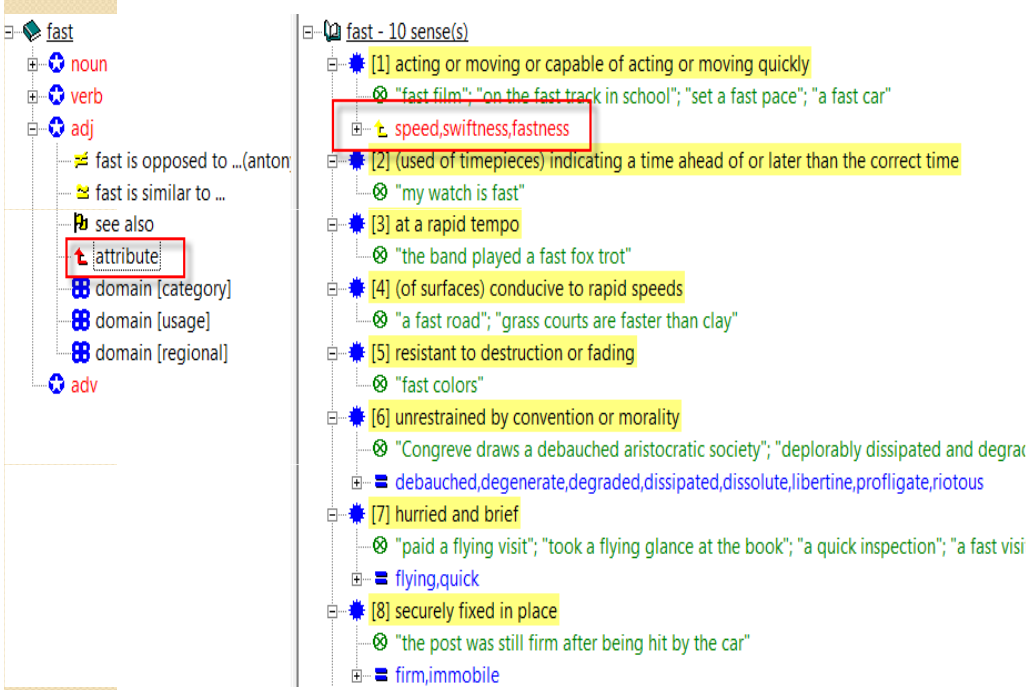


Figure 2.1
Bipolar adjective structure. (⟶ = similarity; ⟶ = antonymy)

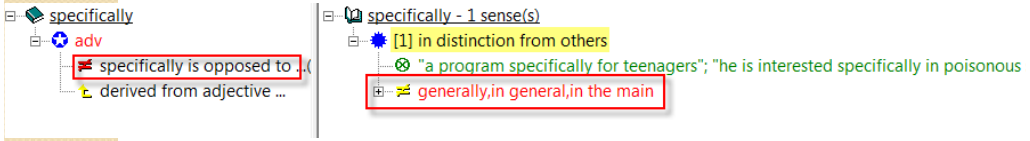






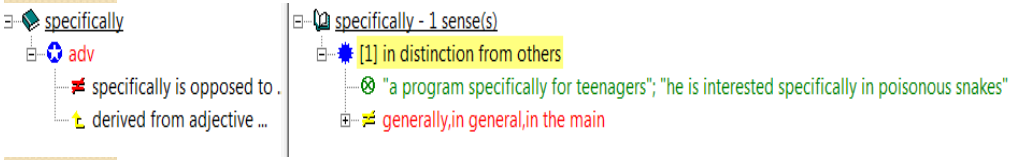
WordNet คืออะไร

- ความสัมพันธ์ใน adverb
  - Antonym: คำตรงข้าม

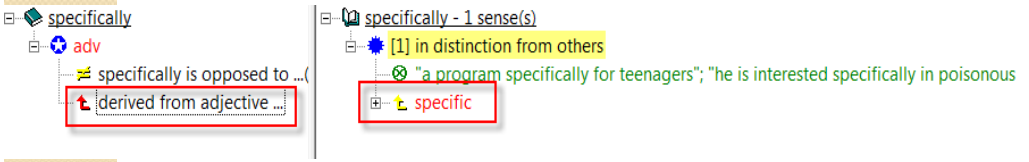


WordNet คืออะไร

- ความสัมพันธ์ใน adverb
  - Antonym: คำตรงข้าม



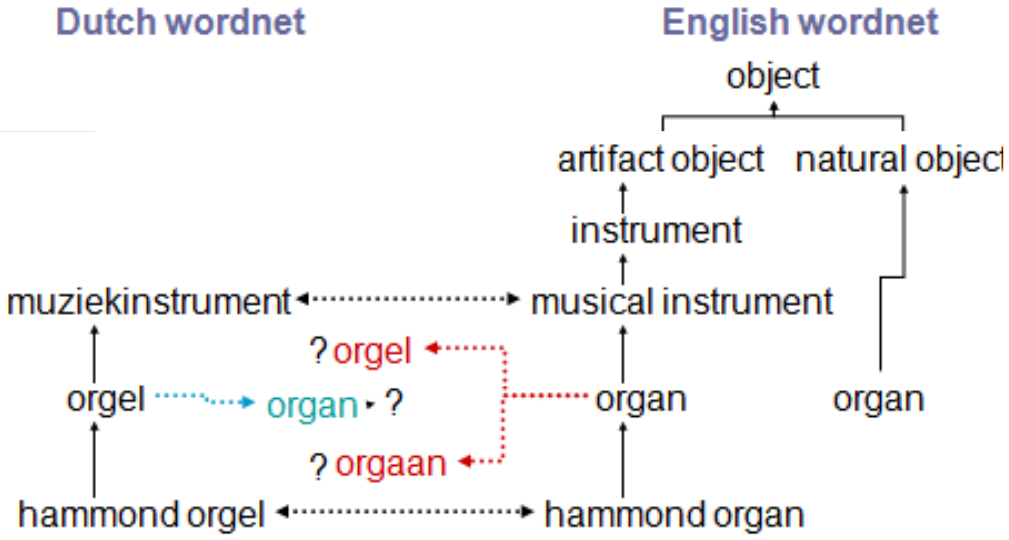
  - Derived from: แปลงมาจาก adjective



การสร้าง WordNet

- Expand approach : แปลคำใน synset ของ WordNet และใช้โครงสร้างนั้น
  - ง่ายและสะดวก ได้ WordNet ใหม่ที่มีโครงสร้างเหมือนกัน compatible กับ resources ต่างๆ ที่ link กับ WordNet eg SUMO
  - ไม่เป็นโครงสร้างของภาษานั้นเอง
- Merge approach : สร้าง WordNet ของภาษานั้นก่อนแล้วค่อย align กับอีก WordNet โดยดูเทียบหาคำแปลที่สอดคล้องกัน
  - ใช้แรงงานสูง ได้โครงสร้างที่แตกต่างกันไป
  - ได้ลักษณะเฉพาะของภาษานั้นเอง

Aligning wordnets



IRION

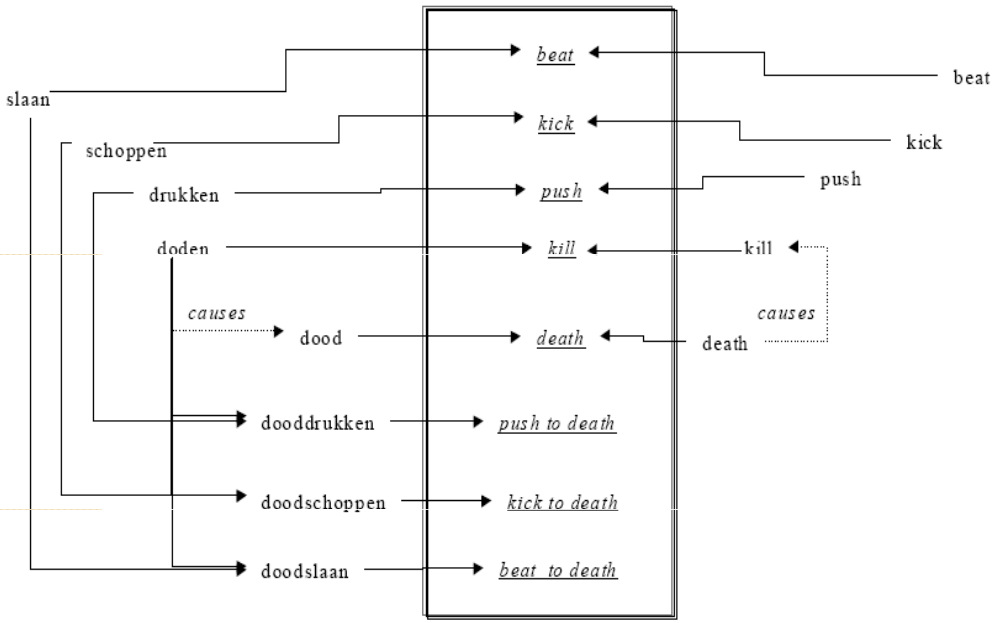


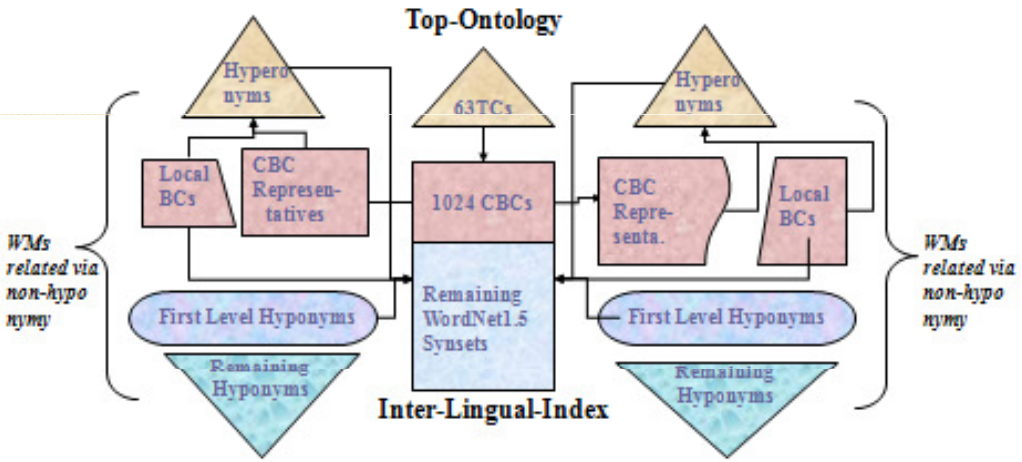
Figure 6: Ways of "killing" *lexicalized in Dutch and not in English.*

Global WordNet Association

- กำหนด core wordnet หรือ base concept เป็น top ontology
  - Common base concept เป็น concept ร่วมในมากกว่า 1 wordnet
  - Local base concept เป็น concept เฉพาะในภาษา
- EuroWordNet มี 1024 synset ที่เป็น CBC คือ share ตั้งแต่ 2 ภาษาขึ้นไป ให้ใช้เป็นพื้นการสร้าง wordnet ในอีกภาษา
- แปล common base concept โดยใช้คำอังกฤษใน synset นั้น
- ตรวจสอบและสร้าง hypernym relation ใน common base concept

  เติม local base concept และขยายลง hyponym

Top-down methodology



IRION

ประโยชน์ของ WordNet

- เพราะบอกความสัมพันธ์ทางความหมายแบบต่างๆ จึงใช้เป็น resource สำหรับงาน NLP ต่างๆ
- Word Sense Disambiguation
- Information retrieval
- Question Answering
- Text Summarization
- Knowledge engineering

การสร้าง Thai WordNet

- P. Sathapornrungkij. 2004. A Semi-Automatic Construction of Thai WordNet Lexical Database from Machine-Readable Dictionaries, Master's thesis, Mahidol University.
- วิทยานิพนธ์มหาบัณฑิต จุฬาฯ ทุน TGIST 2007-2008
  - ธนนท์ หลีน้อย การสร้างเครือข่ายคำไทยของมโนทัศน์พื้นฐานร่วมของเอนทิตีลำดับที่หนึ่งด้วยวิธีการแปลสองทาง และการใช้พจนานุกรมที่สร้างด้วยวิธีการแตกต่างกัน
  - ปริศนา อัครพุทธิพร การสร้างเครือข่ายคำไทยของมโนทัศน์พื้นฐานร่วมของเอนทิตีลำดับที่สองด้วยวิธีการแปลสองทาง การศึกษาปัจจัยความหลากหลายของความหมายที่มีต่อความถูกต้องของการแปล วิทยานิพนธ์มหาบัณฑิต จุฬาฯ ทุน TGIST
- Lexitron, TCL Lexicon, KU etc.

การสร้าง Thai WordNet

- เลือก Lexical ontology หรือ conceptual ontology
  - Lexical ontology มองคำเป็นหลัก ใส่เฉพาะคำ โครงสร้างแต่ละภาษาจึงต่างกัน ต้องมีการ map ผ่าน inter-lingual index : แนวทางของ Euro WordNet
  - Conceptual ontology ยอมให้มี concept ที่ไม่ได้เป็นคำในภาษา แนวทางของ Princeton WordNet
- ยึดตาม GWA ทำส่วนที่เป็น base concept ก่อน
- แปลคำอังกฤษในแต่ละ synset เป็นไทย
  - ใช้พจนานุกรมอังกฤษ-ไทยแบบต่างๆ ทำแบบ senmi-automatic ได้

การสร้าง Thai WordNet

- เลือกคำแปลไทยที่เข้ากับ synset นั้น
  - ดูว่าคำแปลไทยเข้ากับความหมายตาม gloss ที่ให้ไว้ ‘container’ = any object that can be used to hold things (especially a large metal boxlike object of standardized dimensions that can be loaded from one form of transport to another)
  - => *ตู้สินค้า กระติก ภาชนะ ที่ใส่*
  - ดูคำลูกกลุ่ม ‘cargo container’, ‘dish’, ‘spoon’, ‘bag’, ‘vessel’, ‘wheeled vehicle’, etc.
  - เหลือคำเดียวที่เหมาะ *ที่ใส่*
  - *** ต้องพิจารณาโครงสร้างมโนทัศน์ไปด้วย ***

การสร้าง Thai WordNet

- ปัญหาที่พบ
  - ความไม่สอดคล้องระหว่างโครงสร้างมโนทัศน์ไทย อังกฤษ

การสร้าง Thai WordNet

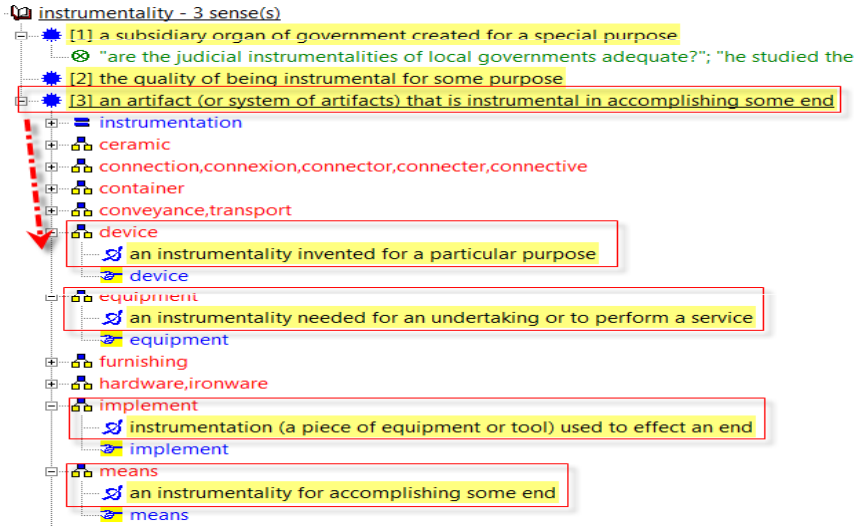
- ปัญหาที่พบ
  - ความไม่สอดคล้องระหว่างโครงสร้างมโนทัศน์ไทย อังกฤษ

การสร้าง Thai WordNet

- ปัญหาที่พบ
  - ความไม่สอดคล้องระหว่างโครงสร้างมโนทัศน์ไทย อังกฤษ

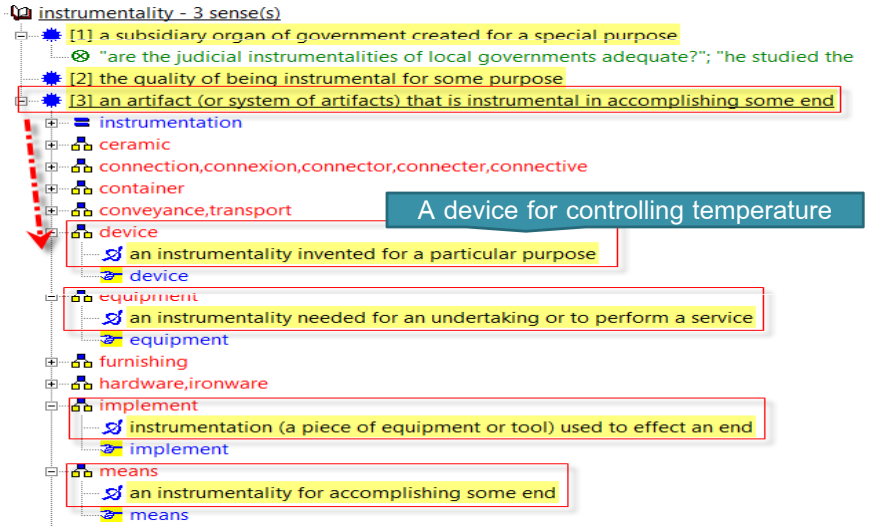
การสร้าง Thai WordNet

- ความไม่สอดคล้องระหว่างโครงสร้างมโนทัศน์ไทย อังกฤษ



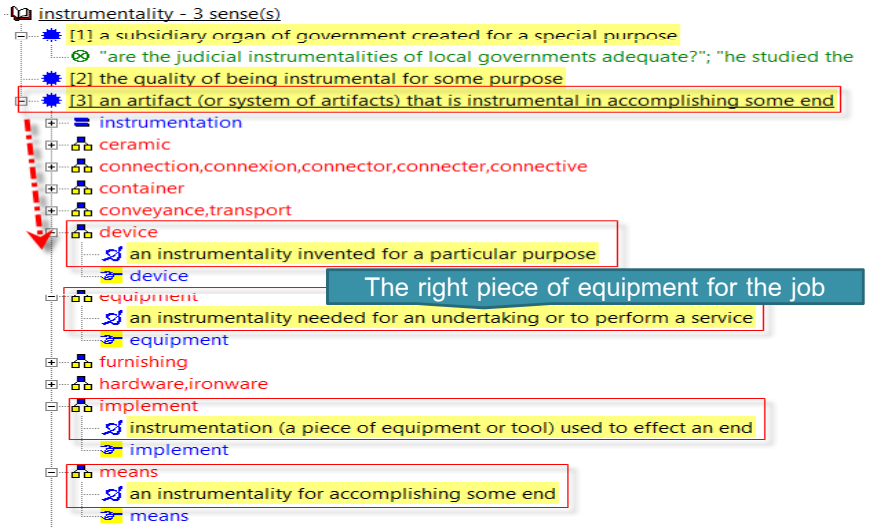
การสร้าง Thai WordNet

- ความไม่สอดคล้องระหว่างโครงสร้างมโนทัศน์ไทย อังกฤษ



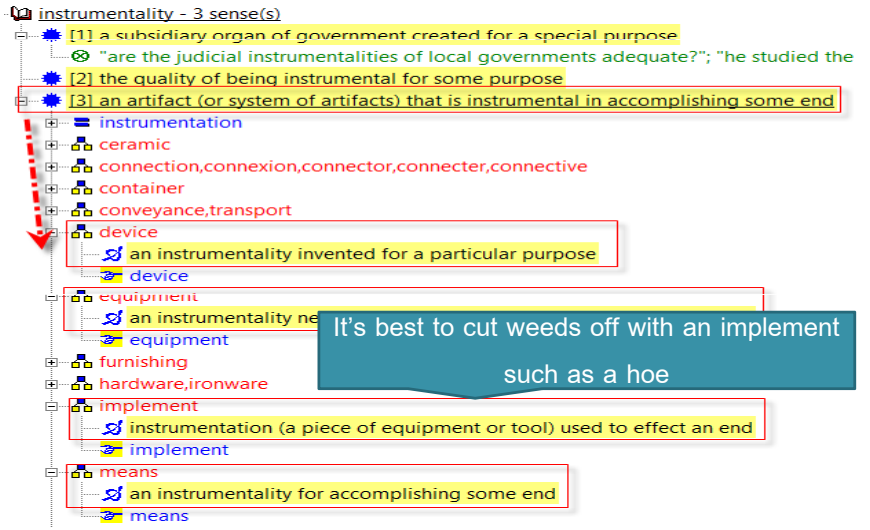
การสร้าง Thai WordNet

- ความไม่สอดคล้องระหว่างโครงสร้างมโนทัศน์ไทย อังกฤษ



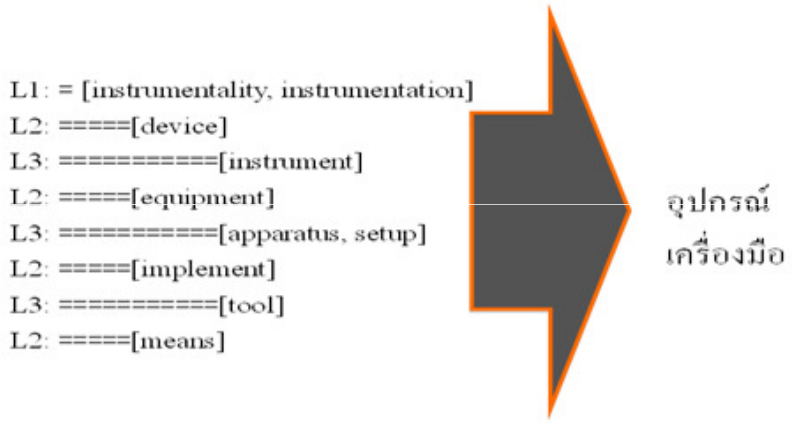
การสร้าง Thai WordNet

- ความไม่สอดคล้องระหว่างโครงสร้างมโนทัศน์ไทย อังกฤษ



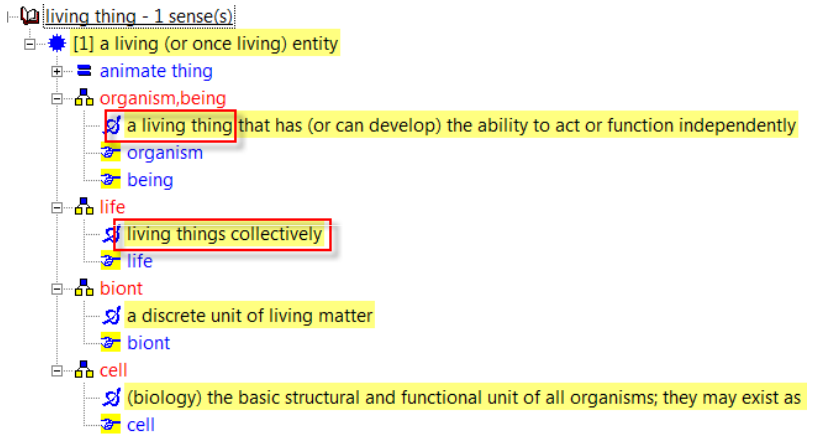
การสร้าง Thai WordNet

- ปัญหาที่พบ
  - ความไม่สอดคล้องระหว่างโครงสร้างมโนทัศน์ไทย อังกฤษ



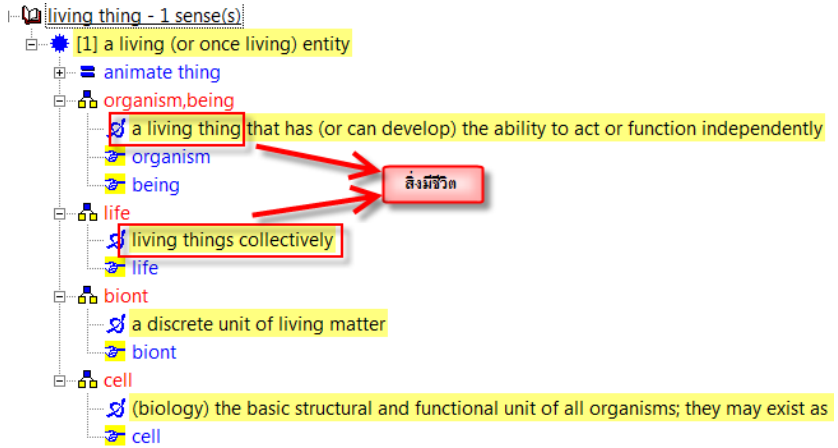
การสร้าง Thai WordNet

- ปัญหาที่พบ
  - ความไม่สอดคล้องระหว่างโครงสร้างมโนทัศน์ไทย อังกฤษ



การสร้าง Thai WordNet

- ปัญหาที่พบ
  - ความไม่สอดคล้องระหว่างโครงสร้างมโนทัศน์ไทย อังกฤษ



การสร้าง Thai WordNet

- ไม่สามารถแปลเป็นไทยแล้วดูเฉพาะส่วนหรือใน synset นั้น
- ต้องพิจารณา synset บนและล่างประกอบกันไป
- โครงสร้าง Thai WordNet ต่างจาก English WordNet
- ต่างมากหรือน้อย? ปัญหา lexical gap ?
  - สำหรับการแปล อยากได้ equivalence ที่ไม่เป็นคำได้
- การทำแบบ manual สิ้นเปลืองมาก
- สร้าง gold standard test set ที่สร้างแบบ manual
- หา semi-automatic approach ต่างๆ ที่ได้ผลดีสุด
- ปรับแก้ไขแบบ manual

สิ่งที่ไม่มีใน WordNet

- ความสัมพันธ์ระหว่างนามกับกริยาและ world knowledge
- FrameNet

Attack

Definition:

An Assailant physically attacks a Victim (which is usually but not always sentient), causing or intending to cause the Victim physical damage. A Weapon used by the Assailant may also be mentioned, in addition to the usual Place, Time, Purpose, Reason, etc. Sometimes a location is used metonymically to stand for the Assailant or the Victim, and in such cases the Place FE will be annotated on a second FE layer.

As soon as he stepped out of the bar he was SET upon by four men in ski-masks.

Is he INVADING Iraq just to cover other shortcomings?

Then Jon-O's forces AMBUSHED them on the left flank from a line of low hills.

FEs:

Core:

Assailant [Asl]
Semantic Type
Sentient

The person (or other self-directed entity) that is attempting physical harm to the Victim.

~~The mysterious fighter~~ ~~ATTACKED~~ the guardsmen with a sabre.

Victim [Vic]
Semantic Type
Sentient

This FE is the being or entity that is injured by the Assailant's attack.

The mysterious fighter ATTACKED the guardsmen with a sabre.

Non-Core:

Circumstances []

Circumstances describe the state of the world (at a particular time and place) which is specifically independent of the event itself and any of its participants.

Containing_event []

This FE denotes an event that occurs or state of affairs that holds at a time that includes the time during which the event or state of affairs reported by the target occurs and of which it is taken to be a part.

Depictive [Dep]

This FE is used for any Depictive phrase describing the state of the Assailant or Victim while the Attack