# Building a Gold Standard for Thai WordNet

**Dhanon Leenoi**

Department of Linguistics
Faculty of Arts
Chulalongkorn University
Bangkok, Thailand 10330
narzizsus@gmail.com

**Thepchai Supnithi**

National Electronics and Computer
Technology Center
Pathumthani, Thailand 12120
thepchai.sup@nectec.or.th

**Wirote Aroonmanakun**

Department of Linguistics
Faculty of Arts
Chulalongkorn University
Bangkok, Thailand 10330
awirote@chula.ac.th

## Abstract

This paper presents a method of building a gold standard test set of Thai WordNet. The results of this research can be utilised for evaluating or comparing the results from different approaches of Thai WordNet construction. In this research, a part of Thai WordNet is carefully handcrafted from Common Base Concepts' FirstOrderEntities with five translation resources. However, we found that to build a gold standard test set is not easy as finding words that can fit to the definition of synsets; cultural gaps between the different languages have to be aware of.

**Keywords**: WordNet, synset, Common Base Concepts

## 1 Introduction

WordNet is a lexical database in which English words are grouped into sets of synonyms called *synsets*. It provides concept definitions and records the semantic relations between these synonym sets. The objective is twofold: to produce a combination of dictionary and thesaurus being more intuitively usable, and to support automatic text analysis, natural language processing and artificial intelligence applications. WordNet is a semantic lexicon for English language whose design is inspired by psycholinguistic theories (Miller et al, 1993). WordNet provides synset glosses or definitions, and records semantic relations between these synsets. It can be called the semantic organisation, which supports synonymic, antonymic, hyponymic – hypernymic, and meronymic – holonymic relations. The significant increase of using wide coverage of ontologies for Natural Language Processing tasks drives WordNet become a *de-facto* standard for a wide range of NLP applications. WordNet is utilised as a knowledge-based approach in information retrieval (Richardson; Smeaton, 1995) to calculate similarity between query and document. WordNet is also used in query expansion (Voorheer, 1994) to increase accuracy; moreover, used in word sense disambiguation (Mihalcea; Moldovol, 2001) for selecting correct senses for each word. WordNet is an important database for many applications; however, little has been done with WordNet for Thai language.

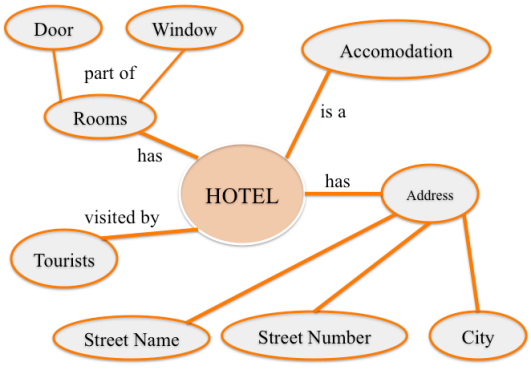


Figure 1. Semantic Relations

Sathapornrungkij (2004) proposed a semi-automatic construction of Thai WordNet from machine-readable dictionaries. However, only a random sample of the result was evaluated manually

by considering whether those Thai words match the English definition. It will be shown in this paper that it is difficult to make the correct judgement by considering only the definitions without taking the conceptual structure into account. Moreover, if other approaches of WordNet construction were implemented, it would not be possible to evaluate or compare the results from different approaches.

This research aims at manually constructing a part of Thai WordNet from the Common Base Concepts suggested by the Global WordNet Association. The result can be used as a gold standard test set for evaluating any Thai WordNet constructions. In addition, we will show that constructing a WordNet is not an easy task as finding words that can fit the definition of a synset.

## 2 The Global WordNet Association and Common Base Concepts

Due to an importance of WordNet on Natural Language Processing research, the Global WordNet Association (GWA) was established by linguists, computer scientists and computer engineers who are interested in WordNet around the world. It is a non-commercial organisation which provides a platform for discussing, sharing and connecting WordNets for all languages in the world, and promotes the development of guidelines and methodologies for building WordNets in new languages.

As the success of English WordNet, or Princeton WordNet, EuroWordNet (EWN) has been developed for several European languages, such as French, German, Italian, Spanish, and etc. The notion of *Common Base Concepts* was introduced in the Euro WordNet project to reach maximum overlap and compatibility across WordNets in different languages. The Common Base Concepts are concepts shared by at least two languages in the EuroWordNet. They are supposed to be the concepts that play the most important role in various WordNets of different languages. Additionally, they are the guideline for building WordNet in new languages suggested by the Global WordNet Association.

Following Lyons (1977), Common Base Concepts, 1,024 synsets, have been distinguished at the first level for 3 types of entities: FirstOrderEntities 493 synsets, SecondOrderEntities 498 synsets, and ThirdOrderEntities 33 synsets.

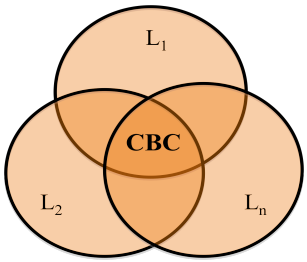


Figure 2. Common Base Concepts

FirstOrderEntities are any concrete entity (publicly) perceivable by the senses and located at any point in time, in a three-dimension space; e.g. mammal, plant, container. SecondOrderEntities are any Static Situation (property, relation) or Dynamic Situation, which cannot be grasped, heart, seen, felt as an independent physical thing. They can be located in time and occur or take place rather than exist; e.g. continue, occur, and apply. ThirdOrderEntities are any unobservable proposition that exists independently of time and space. They can be true or false rather than real. They can be asserted or denied, remembered or forgotten; e.g. idea, though, information, theory, and plan.

FirstOrderEntities are basic concepts consisting of concrete noun. Therefore, it is much better to build a gold standard test set for Thai WordNet from them.

## 3 Building a Gold Standard for TWN

Translation is the most popular approach for building WordNet in new languages. It is easier and faster than creating a new WordNet from the ground up. In this research, then, Thai WordNet is manually carefully created from Common Base Concept's FirstOrderEntities, which composed of 1,056 English words from 493 synsets. Five dictionaries are used as a resource for constructing Thai WordNet in this study. One is a Thai dictionary (the Royal Institute Dictionary). Another one is a Thai-to-English dictionary (Domnern – Satianpong Dictionary). The rest are Thai-to-English and English-to-Thai dictionary, namely SorSettabut Dictionary, Dr. Wit Thiengburanathum Dictionary, and LEXiTRON.

English words in the FirstOrderEntities of Common Base Concept will be translated into Thai

words based on these dictionaries. At this stage, we obtained 14,965 Thai words. Next, irrelevant Thai words have to be filtered out of each synset. The steps are as follow:

Firstly, retain words that their meanings fit the definition of that synset. For example, we obtained only *'เนยแข็ง'* /nɤ:ikʰɛ̌ŋ/ for the concept of a solid food prepared from the pressed curd of milk, 'CHEESE', from all five dictionaries. After considering the definition of this concept, this word will be retained.

Secondly, delete words that their meanings are irrelevant to the definition. In case of polysemy, for example, the concept of the buildings for carrying industrial labour, 'PLANT', we obtained the translations *'ต้นไม้'* /tônmái/ and *'โรงงาน'*/ro:ŋŋa:n/ for the word 'plant'. Since *'ต้นไม้'* /tônmái/ means 'TREE', it will be deleted from the list. But *'โรงงาน'*/ro:ŋŋa:n/ will be kept because it means 'PLANT'.

Thirdly, if translated words are technical terms, we will seek experts' opinion to verify the meanings of those terms. New terms, as suggested by the experts, can be added if necessary. For example, there is no translation for 'BODY PART' from all the dictionaries used. After discussion with an anatomical expert, a term *'ส่วนของร่างกาย'* /sǔankʰǒ:ŋrâ:ŋka:i/, which means 'part-of-body', is added for this concept.

However, in most cases, we found that it is necessary to consult conceptual structure to determine whether the translated words are relevant to the concept. For example, the following words, *'ตู้สินค้า'* /tû:sǐnkʰá:/, *'กระติก'*/kràtìk/, *'ภาชนะ'*/pʰa:tɕʰànáˊ/, and *'ที่ใส่'*/tʰî:sài/, are obtained from the translation of the English words for 'CONTAINER' in the concept of any object that can be used to hold things. After considering hyponym members of 'CONTAINER', such as 'dish', 'spoon', 'bag', 'vessel', 'wheeled vehicle', and etc., only one word *'ที่ใส่'* /tʰî:sài/, in this case, is the correct word for this concept because its meaning fit in the top concept which covering all hyponym members.

During the process of manually checking the translation, we found that the task is not as easy as determining whether the meaning of translated word fit the definition of the concept. Conceptual structure is needed to be consulted as stated above. The task is immensely labourious. We found that difficulties in constructing Thai WordNet are often caused by cultural gaps between Thai and English. Three types of cultural gaps are reported in this paper: categorisation, gender, and collective perception.

### 3.1 Categorisation

According to the concept of business establishment in WordNet hierarchy, we found that the concept of 'outlet' or 'retail store' is a hypernym or superordinate concept of 'store'. When translating these synsets, we obtained *'ห้าง'* /hâ:ŋ/ for 'store' and obtained *'ร้านค้าปลีก'* /rá:nkʰá:plì:k/ for 'retail store'. But for Thai, *'ห้าง'*/hâ:ŋ/ is bigger than *'ร้านค้าปลีก'* /rá:nkʰá:plì:k/. This reflects a difference world outlook between Thai and English. While the concept of 'retail store' is bigger than the concept of 'store' in English culture; conversely, the translation of 'store' or *'ห้าง'*/hâ:ŋ/ is bigger than the translation of 'retail store' or *'ร้านค้าปลีก'* /rá:nkʰá:plì:k/ in Thai culture. Therefore, if we did not pay attention to the structure of concepts, we could put Thai words incorrectly in the Thai WordNet.

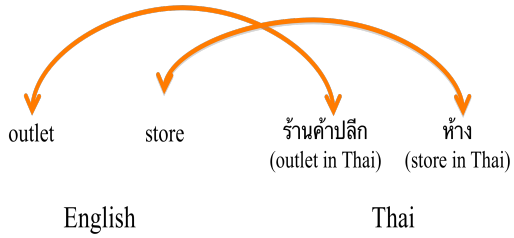


Figure 3. Cross meaning between Thai and English

Next, the instrumentality concept has many hyponym synsets such as 'DEVICE', 'INSTRUMENT', 'EQUIPMENT', 'IMPLEMENT', 'TOOL' and etc. When translating such hyponym synsets, we obtained *'เครื่องมือ'* /kʰrɯ̂əŋmɯ:/ and *'อุปกรณ์'* /ʔùppàkɔ:n/ for all of them. All instrumentality concepts can be replaced by these two Thai words because they are not distinguished in Thai. This example shows a meaning overlap between Thai and English.

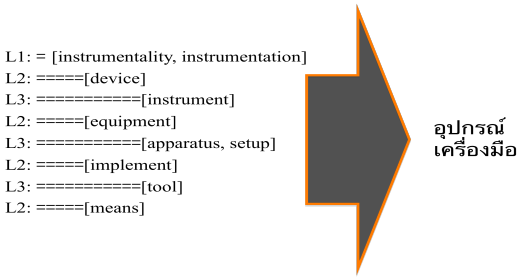


Figure 4. Meaning overlap in categorisation

### 3.2 Collective perception

The concept of a living (or once living) entity, 'LIVING THING', has two major hyponym synsets: 'BEING' and 'LIFE'. In WordNet hierarchy, 'LIVING THING' is the top concept and the hypernym. 'BEING' is the hyponym concept of individuality; 'LIFE' is the hyponym concept of collectiveness. After we translated all those terms, only one word *'สิ่งมีชีวิต'* /sìŋmi:tɕʰi:ʊít/was obtained. All concepts can be replaced by *'สิ่งมีชีวิต'* /sìŋmi:tɕʰi:ʊít/because in Thai, we do not distinguish between individual and collective being. Thus, this example indicates a meaning overlap between Thai and English.


Figure 5. Meaning overlap in collective perception

### 3.3 Gender

One English word can be mapped into two genders in Thai. The concept of a licensed medical practitioner, 'DOCTOR', can be translated to both *'นายแพทย์'* /na:ipʰɛ̂:t/ or 'MALE DOCTOR' and *'แพทย์หญิง'* /pʰɛ̂:tjǐŋ/ or 'FEMALE DOCTOR'. Furthermore, the concept of someone who believes and helps to spread the doctrine of another, 'DISCIPLE', also can be translated to both *'อุบาสก'* /ʔùba:sòk/or 'MALE DISCIPLE', and *'อุบาสิกา'* /ʔùba:sìka:/or 'FEMALE DISCIPLE'. These examples suggest that the structure of Thai WordNet is different from that of English. In these cases, two hyponym synsets, one for male and one for female, should be added.

L1: ===[doctor, doc, physician, MD, Dr., medico] [คุณหมอ, แพทย์, หมอ, หมอยา]
L2: ========== [male_doctor] [นายแพทย์, นพ.]
L2: ========== [female_doctor] [แพทย์หญิง, พญ.]

Figure 6. Hyponym synsets for gender

## 4 Discussion

We have learned from this research that creating WordNet in one language is not as easy as copying WordNet from another language and replacing equivalent words in the target language. The structure of Thai WordNet should reflect how things are conceptualised in Thai. Thus, in principle, Thai WordNet should be constructed from the ground up; that is, Thai WordNet should be constructed by analysing practical usage of all lexemes. All words in Thai corpus need to be segmented and analysed their semantic features. Then, Thai WordNet structure could be constructed on the basis of those results. By this method, a genuine Thai WordNet can reflect Thai culture; however, this approach is immensely difficult. All structure must be considered to find a word for each synset. Hence, translation is still a popular and feasible approach in construction a Thai WordNet because it is easier and faster. Beside, maintaining similar conceptual structures between Thai WordNet and English WordNet should be useful for many NLP applications, e.g. machine translation. We just need to be aware of cultural gaps and try to amend the structure to fit the Thai data.

Although this paper indicates the incompatible concepts between Thai and English, most of synsets can be mapped. To solve the problems of mismatch between English and Thai conceptual structures, we need to decide whether Thai WordNet should be lexicalised ontology or conceptual ontology. In this study, unlike EuroWordNet, we prefer conceptual ontology because we would like to have maximum correspondences between English and Thai conceptual structures. This would make more useful for applications like machine translation and information retrieval. By adopting this approach, all English concepts will be preserved even they cannot be lexicalised in Thai. English concepts that do not lexicalised in Thai will be translated into a phrase, whereas new synsets can be added if Thai has more complex concepts.

However, in cross meaning problems as shown in Figure 3, Thai conceptual structure is different from the English one. This might cause a problem for mapping between English and Thai concepts at this point. This issue will be investigated further.

## 5 Conclusion and future work

This research does not favour any translation approaches, though translation is used when manually creating a Thai WordNet for Common Base Concepts. Rather, we aim at building a gold standard that can be used a test set for anyone who wants to construct a Thai WordNet. Researchers who want to create Thai WordNet with any approaches can evaluate their results with this gold standard test set.

This research is the first manually Thai WordNet construction by linguists. We carefully consider every synset definitions. Furthermore, we consult WordNet hierarchy and we are mindful of gaps between Thai and English. Creating a Thai WordNet is not a simply task as translating English words with bilingual dictionaries and utilise such words without consulting WordNet structure. So, the result of this study, we believe, is the most accurate one and can be used as a gold standard test set for other automatic or semi-automatic approaches of Thai WordNet construction in the future.

Due to the language specifics of English and Thai, some English synset can be merged because those concepts do not exist in Thai. Conversely, some English synset must be split because Thai has more complex. We have to resolve cross meaning problems and accomplish a gold standard Thai WordNet.

## 6 Acknowledgement

Thanks to Thailand Graduate Institute of Science and Technology (TGIST), National Science and Technology Development Agency (NSTDA) for supporting the scholarship and Center of Excellence for Language Linguistics and Literature, Faculty of Arts, Chulalongkorn University for supporting registration.

## References

Lyons, J. 1977. *Semantics.* London. Cambridge University Press.

Mihalcea R.F. &Moldovol D.I. 2001."A highly accurate bootstrapping algorithm for word sense disambiguation".*International Journal on Artificial Intelligence Tools* 10 (2001): 5 – 21

Miller, G. A. 1995. "WordNet: a Lexical Database for English". *Communications of the ACM 38*, (November 1995): 39 – 41

Miller, G. A., Fellbaum, C., and Miller K. J. (1993) *Five Papers on WordNet*[Computer file] Available from: ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.ps [2006, November 2]

PatanakulSathapornrungkij. 2004. *A Semi-automatic construction of Thai wordnet lexical database from machine readable dictionaries*. Master's Thesis, Department of Computer Science, Faculty of Graduate Studies, Mahidol University.

Richardson, R. &Smeaton, A.F. 1995. Using wordnet in a knowledge-based approach to information retrieval. *Technical Report CA-0395*, School of Computer Applications, Dublin City University

Seo, H. C. et al. 2004. Unsupervised word sense disambiguation using WordNet relations. *Computer Speech and Language* 18, (May 2004): 253 – 273

E.M. Voorhees. 1994. Query expansion using lexical-semantic relations. *In Proceedings of the 17th ACM-SIGIR Conference*, 61-69.

Vossen, P. 1999. *EuroWordNet General Document Version 3 Final*.University of Amsterdam.EuroWordNet LE2-4003, LE4-8328. 1999.

Vossen, P. 2001. *EuroWordNet General Document*[Computer file] Available from: http://www.hum.uva.nl/~ewn[2006, November 2]