Construction of Thai WordNet Lexical Database from Machine Readable Dictionaries

Patanakul Sathapornrungkij
Department of Computer Science
Faculty of Science, Mahidol University
Rama6 Road, Ratchathewi
Bangkok, Thailand, 10600
patanakul@yahoo.com

Charnyote Pluempitiwiriyawej
Department of Computer Science
Faculty of Science, Mahidol University
Rama6 Road, Ratchathewi
Bangkok, Thailand, 10600
cccpt@mahidol.ac.th

Abstract

We describe a method of constructing Thai WordNet, a lexical database in which Thai words are organized by their meanings. Our methodology takes WordNet and LEXiTRON machine-readable dictionaries into account. The semantic relations between English words in WordNet and the translation relations between English and Thai words in LEXiTRON are considered. Our methodology is operated via *WordNet Builder* system. This paper provides an overview of the WordNet Builder architecture and reports on some of our experience with the prototype implementation.

1 Introduction

Due to the significant increase in the use of lexical databases, WordNet (Fellbaum 1998) becomes one of lexical databases that are widely used as a lexical information source for many applications: e.g. information retrieval, text classification, semantic disambiguation and etc. (Harabagiu, 1998 and Rila, 1998). The database design structure of WordNet has been extended to other languages such as Dutch (Vossen, 1999), Italian (Pianta et. al., 1990), Spanish (Atserias et. al., 1997), Hungarian (Prószéky and Miháltz, 2002) and Chinese (Chen et. al., 2000). For Thai language, the Thai WordNet is unavailable up to now although the related idea of constructing Thai lexical database has been expressed in (Sornlertlamvanich and Boriboon 1995).

Construction of word net lexical database, particularly the Thai WordNet, is a long term project. The manual construction requires a large number of lexicographers to hand-build the lexical database. It has been generally estimated that the average time needed to manually construct a lexical entry in a lexical database is about 30 minutes (Neff et. al., 1993).

An alternative is to construct the word net lexical database from WordNet and other existing lexical resources (Farreres et. al., 1998). The existing lexical resources can be natural language corpora and machine-readable dictionaries. The corpora are collections of words, texts and sample sentences which are collected from particular materials such as newspapers and scripts of radio broadcast. The machine-readable dictionaries are basically electronic representations of standard printed dictionaries. They can be monolingual e.g., COLLIN, bilingual e.g., LEXiTRON (Palingoon et. al. 2002) or multilingual e.g., Babylon. One may consider a lexical database to be a type of machine-readable dictionary since lexical entries in the database are typically similar to that in machine-readable dictionary but the structure of the lexical database is more precise and predictable.

In this paper, we describe a semi-automatic approach to construct the Thai WordNet lexical database from WordNet and LEXiTRON machine-readable dictionaries. WordNet provides lexical and semantic relations between English words whereas LEXiTRON provides translation relations between Thai and English words. Our approach is operated via *WordNet Builder* system which has been designed and implemented. WordNet and LEXiTRON are extracted to obtain lexical, semantic and translation relations. A sample set of interesting relations are analyzed and evaluated to construct a model which is used to construct the Thai WordNet lexical database.

2 Overview of WordNet Builder Architecture

A conceptual overview of the WordNet Builder system is shown in Figure 1. The MRDs and the Thai WordNet are the input and the output of the system, respectively. System components include the MRD extractors, the link analyzer and the WordNet constructor.

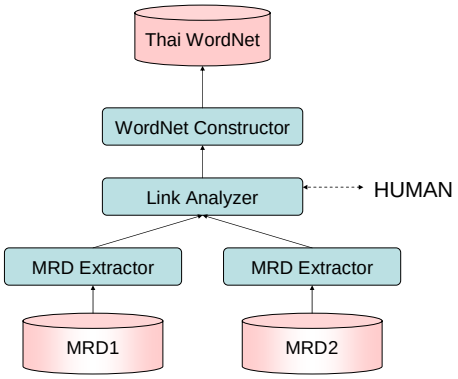

Figure 1 Schematic desciption of the WordNet Builder architecture and its main components

2.1 Machine Readable Dictionaries (MRDs)

We relate the MRDs as the electronic version of standard dictionaries which may contain other lexicographic information that does not appear in the printed version. MRDs can be monolingual, bilingual, and multilingual. Any number of MRDs can be taken into account. However, the English WordNet (Fellbaum 1998) which is one of monolingual MRDs and the LEXiTRON which is one of bilingual MRDs are our current focus. The English WordNet provides a lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory. The LEXiTRON provides a bilingual corpus-based lexical reference system.

2.2 Thai WordNet

We relate the Thai WordNet as a collection of Thai words and their relationships. In this research, we modeled the Thai WordNet after the English WordNet which organizes words by their meaning and identifies the relationships among words via reference links. Our Thai WordNet can be considered as a lexical database which contains information about Thai and English words. We use the Database Management System (DBMS) to help manage the lexical database.

2.3 MRD Extractors

Source-specific MRD Extractors provide access to the underlying machine-readable dictionaries and support data restructuring and data cleansing. Specifically, an MRD Extractor converts data format into a common format and cleans the noisy, erroneous, missing, irrelevant and duplicate data. It joins and aligns the scattered data for smoothly access and selects the relevant data. In this research, we consider the nominal data set to be relevant with respect to the multiple criteria used in the Link Analyzer.

2.4 Link Analyzer

The Link Analyzer supports classifying *translation links* with respect to *semantic links* and supports deriving *candidate links*. The translation links are referred to as the relationships between Thai and English words. The semantic links are referred to as the relationships between English words and their meaning. The candidate links are referred to as the relationships between Thai words and their meaning. We use multiple criteria for the classification and derivation which will be described in Section 3. For each criterion, a set of sample translation links are verified and a statistical classification model is constructed. The model is deployed to classify and validate the remaining translation links. This approach can reduce human intervention which is time consuming. The verification of translation links will be described in Section 4 while the model construction and evaluation will be described in Section 5.

2.5 WordNet Constructor

The WordNet Constructor support relating word forms to their meaning, inferring the relationships between words and meanings, and attaching the glossary. In this research, Thai words are classified regarding the synonym sets defined in WordNet and the glossary is obtained from the LEXiTRON.

3 Classification and Derivation of Links

In English WordNet, words are organized by *synset* which is a set of words having the same meaning. In this research, we aim at constructing Thai WordNet in which Thai words are organized by synsets. The synsets in Thai WordNet can be derived from the English WordNet. The candidate links between Thai words and synsets can be derived from the semantic links which are obtained from WordNet and the translation links which are obtained from LEXiTRON.

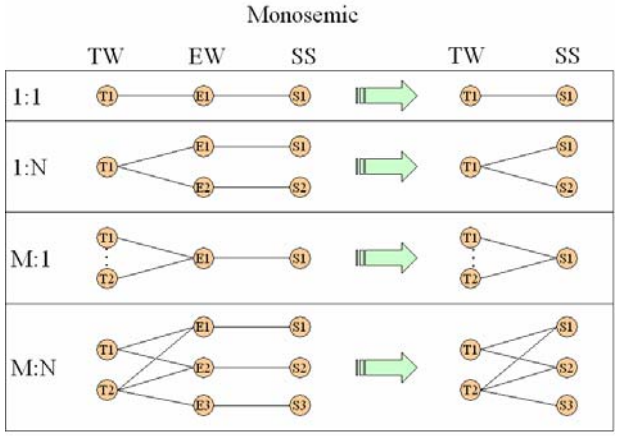

Figure 2: Conceptual description of monosemic criteria

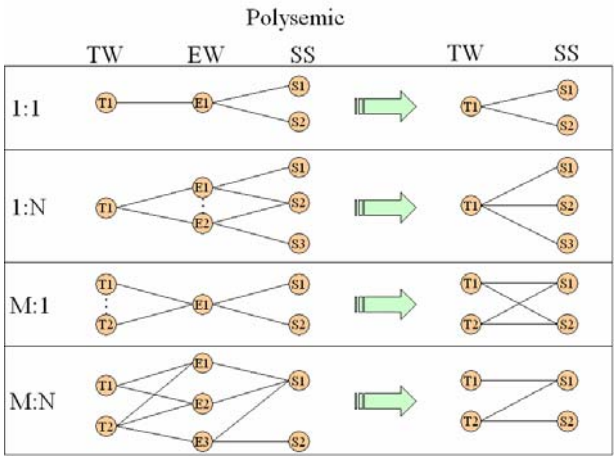

Figure 3: Conceptual description of polysemic criteria

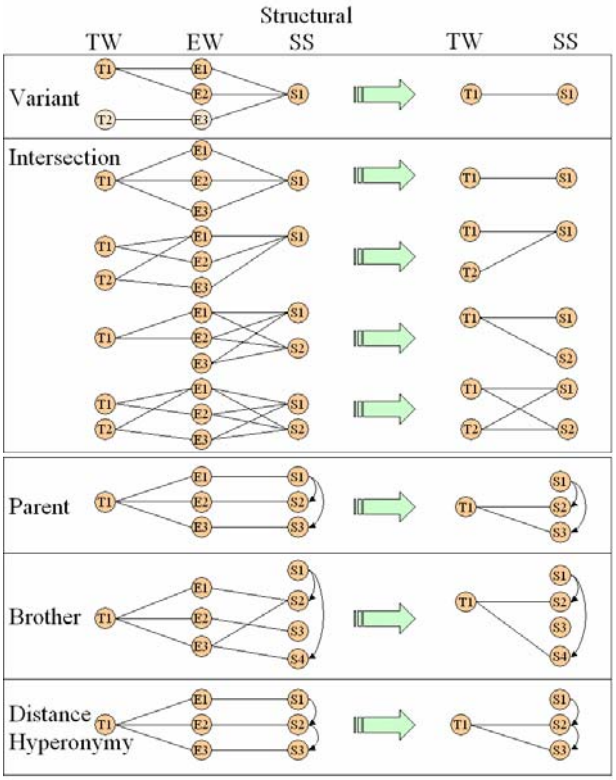

Figure 4: Conceptual description of structural criteria

Like (Farreres et. al., 1998), we derived links between Thai words and synsets with respect to 13 criteria which are categorized into the following three groups:

- *Monosemic criteria* focus on an English word which has only one meaning. Such English word is classified into only one synset with respect to WordNet 1.7.
- *Polysemic criteria* focus on an English word which has multiple meaning. Such English word is classified into multiple synsets with respect to WordNet 1.7.
- *Structural criteria* focus on the structural relations among synsets with respect to WordNet 1.7.

Figures 2, 3 and 4 show the conceptual description of the monosemic, the polysemic and the structural criteria, respectively. In the Figures, T_i represents a particular Thai word, E_j represents a particular English word, and S_k represents a particular synset. The links between T_i and E_j represent the translation links which is extracted from LEXiTRON database. The links between E_j and S_k represent the semantic relations which are extracted from WordNet. The links between a T_i and S_k are the candidate links which will be used to construct Thai WordNet.

Now we can describe the 13 criteria as follows:

Criterion 1. ***Monosemic one-to-one***: If the English word is associated with one single synset and only one translation exists between the Thai word and the English word, then a candidate link between the Thai word and the synset is derived.

Criterion 2. ***Monosemic one-to-many***: If the Thai word can be translated into multiple English words each of which is translated into one particular Thai word and each English word is associated with one single synset, then candidate links between the Thai word and all synsets are derived.

Criterion 3. ***Monosemic many-to-one***: If one or more Thai words can be translated into a single English word and the English word is associated with one single synset, then candidate links between all Thai words and the synset are derived.

Criterion 4. ***Monosemic many-to-many***: If a Thai word can be translated into multiple English words, each English word can be translated into multiple Thai words and each English word is associated with one single synset, then candidate links between all Thai words and all synsets are derived.

Criterion 5. ***Polysemic one-to-one***: If the English word is associated with multiple synsets and only one translation exists between the Thai word and the English word, then candidate links between the Thai word and all synsets are derived.

Criterion 6. ***Polysemic one-to-many***: If the Thai word can be translated into multiple English words each of which is translated into one particular Thai word and each English word is associated with one single synset, then candidate links between the Thai word and all synsets are derived.

Criterion 7. ***Polysemic many-to-one***: If one or more Thai words can be translated into a single English word and the English word is associated with one single synset, then candidate links between all Thai words and the synset are derived.

Criterion 8. ***Polysemic many-to-many***: If a Thai word can be translated into multiple English words, each English word can be translated into multiple Thai words and each English word is associated with one single synset, then candidate links between all Thai words and all synsets are derived.

Criterion 9. ***Variant***: If a synset contains multiple English words at least two of which link to the same Thai word, then a candidate link between the synset and the Thai word is derived.

Criterion 10. ***Intersection***: If a Thai word links to multiple English words all of which share at least one common synset, then a candidate link between the common synset and the Thai word is derived.

Criterion 11. ***Parent***: If a Thai word links to multiple English words and a synset of the English words is the direct parent of remaining synsets (i.e., hyponym sets), then candidate links between the Thai word and all hyponym sets are derived.

Criterion 12. ***Brother***: If a Thai word links to English words whose synsets are brother with respect to a common parent link (i.e., co-hyponym sets), then candidate links between the Thai word and all co-hyponym sets are derived.

Criterion 13. ***Distance hyponym***: If a Thai word links to English words whose synsets are distance hypernym sets of the rest of the English word links, then candidate links between the Thai word and the lower level synsets are derived.

Table 1: A snapshot of the link existence matrix

Eword	Tword	c01	c02	c03	c04	c05	c06	c07	c08	c09	c10	c11	c12	c13
girl	เด็กหญิง	0	0	0	0	0	0	0	1	0	0	0	1	0
abandonment	การสละ	0	0	0	0	0	0	0	1	0	0	0	1	0
Thai	ภาษาไทย	0	0	0	0	0	0	0	1	0	1	0	0	0
desertion	การทอดทิ้ง	0	0	0	0	0	0	0	0	0	1	0	0	0

Table 2: Number of links created in each criterion

No	Criteria	# of Links
1.	Monosemic one-to-one	7,784
2.	Monosemic one-to-many	1,688
3.	Monosemic many-to-one	5,463
4.	Monosemic many-to-many	1,802
5.	Polysemic one-to-one	856
6.	Polysemic one-to-many	332
7.	Polysemic many-to-one	1,230
8.	Polysemic many-to-many	88,958
9.	Variant	5,067
10.	Intersection	26,929
11.	Parent	6,130
12.	Brother	21,277
13.	Distance Hyponymy	489

In this research, the classification of translation links and the derivation of candidate links are two of the processes in the Link Analyzer. Table 1 shows a snapshot of the classification result which is presented as a link existence matrix. Table 2 summarizes the total number of candidate links in each criterion.

4 Verification of Links

As shown in Table 2, a tremendous amount of candidate links has been classified. If all translation links are verified, this will take enormous time-consuming. To overcome such problem, we apply the stratified sampling technique (Krejcie and Morgan, 1970). With 95% confidence level, the 400 translation links in each criterion are randomly selected

Table 3: Number of correct links in each criterion (sample size = 400 links)

Criterion	Number of correct links	%
1	368	92.00
2	332	83.00
3	319	79.75
4	254	63.50
5	360	90.00
6	304	76.00
7	292	73.00
8	197	49.25
9	356	89.00
10	322	80.50
11	351	87.75
12	202	50.50
13	317	79.25

After sampling, the translation links are verified with respect to three available dictionaries: the Thaisoft So Sethaputra, the LEXiTRON, and the WordNet 2.0. The verification of translation links considers mainly on Thai words which match to English concept. Table 3 shows the summary of link verification for each criterion.

Table 4 Summary existence matrix

NTOT	NOK	c01	c02	c03	C04	c05	c06	c07	c08	c09	c10	c11	c12	c13
345	345	1	0	0	0	0	0	0	0	0	0	0	0	0
301	301	0	0	1	0	0	0	0	0	0	0	0	0	0
281	281	0	0	0	0	0	0	1	0	0	0	0	0	0
258	258	0	0	0	0	0	0	0	0	1	1	0	0	0
1	0	0	0	0	1	0	0	0	0	0	0	1	0	0
14	0	0	0	0	1	0	0	0	0	0	1	0	0	0
41	0	0	1	0	0	0	0	0	0	0	0	0	0	0

The result of verification is represented as a summary existence matrix as shown in Table 4. Note that the *NTOT* is the total number of verified links and the *NOK* is the number of correct verified links. The summary existence matrix will be used to construct a model for predicting the correctness of the remaining translation links and identifying the correlation and the significance of the multiple criteria.

5 Model Construction and Evaluation

The probability that a link would be correct can be estimated by *P(OK) = NOK/NTOT*. In this research, we use the logistic regression model (Javaras and Wiesner 2002) to predict the correctness of the remaining links. In general, the logistic regression is used to predict a discrete outcome from a set of binary variables. The linear logistic regression model can be defined as

$$log(P(nok/ntot)) = \beta_0+\beta_1 C_1 +\ldots+\beta_{13} C_{13}$$

where *nok* is the number of correct evaluation for the set of solutions of every group of methods, *ntot* accumulates the total number of evaluations, C_i is a boolean variable representing the existence of link in the i^{th} criterion, and β_i is unknown parameter which required the least square criterion.

To evaluate the model, a statistical approach is used. A P-value of each criterion describes the significant of that criterion in the model by explaining the probability of a link of being correct. For a P-value lower than 0.05, the criterion is significant in the model.

Our goal is to find a minimal set of significant criteria. We have applied the backward method to find a local optimum iteratively by deleting the less informative variable between the non-significant ones (Sathapornrungkij 2004).

Table 5 Evaluation result of the model when all 13 criteria are considered.

Criterion	Beta	P-value
0	-0.2500	0.032
1	2.6940	0.000
2	1.0314	0.000
3	1.5946	0.000
4	0.3891	0.008
5	2.4082	0.000
6	1.9264	0.000
7	1.2209	0.000
8	-0.0255	0.817
9	0.5221	0.002
10	1.4567	0.000
11	1.6409	0.000
12	0.3825	0.000
13	0.5202	0.000

We start constructing the first logistic regression model by considering all 13 criteria. The model is evaluated using the P-values as shown in Table 5. The evaluation result has shown that the 8^{th} criterion (i.e., the Polysemic many-to-many criterion) is insignificant since its corresponding P-value is no less than 0.05. Therefore, our second model is constructed without consideration of the 8^{th} criterion. Table 6 shows the evaluation result. The model is accepted since all P-values are less than 0.05.

Table 6 Evaluation result of the model when excluding the 8^{th} criterion.

Criterion	Beta	P-value
0	-0.3877	0.003
1	2.8314	0.000
2	1.0943	0.000
3	1.7286	0.000
4	0.4997	0.001
5	2.5447	0.000
6	2.0337	0.000
7	1.3575	0.000
9	0.5440	0.001
10	1.5421	0.000
11	1.7096	0.000
12	0.4664	0.000
13	0.5774	0.000

6 Thai WordNet Construction

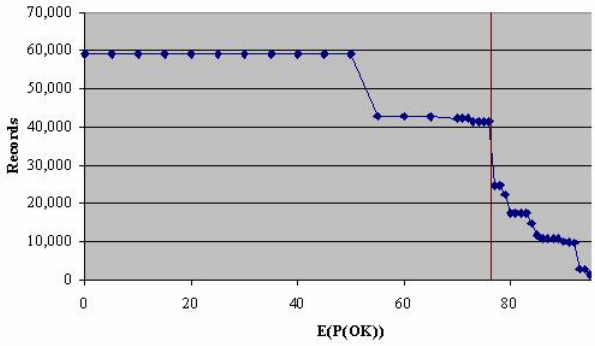

Figure 6 The coverage and the accuracy of the accepted model to construct the Thai WordNet

The accepted model is deployed to construct the Thai WordNet. Trading off between the coverage and the accuracy need to be considered. Figure 6 shows the coverage and the accuracy of the model when it is applied to each translation link. We can see that the model provides 80% coverage with 76% accuracy. At this point, our Thai WordNet contains 44,844 semantic links describing relations among 13,730 words and 19,582 synsets.

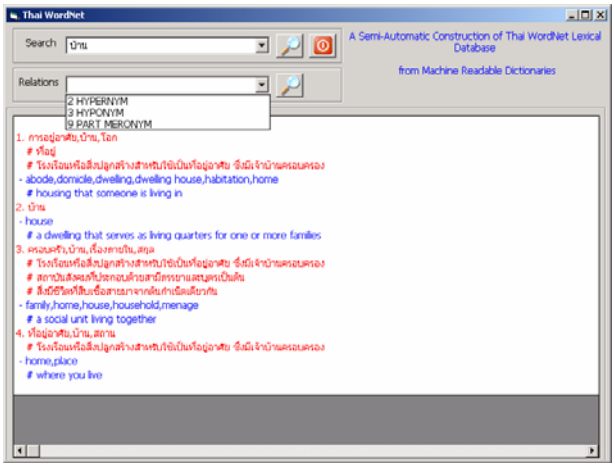

Figure 7 User Interface of the Thai WordNet

By applying the model to all remaining translation links, the semantic relations between

Thai words and synsets are constructed according to each criterion. In addition, the lexical relations among Thai words are inferred with respect to the lexical relation defined in the English WordNet. Finally, other information such as word descriptions and usage example are attached. Figure 7 captures the user interface of our Thai WordNet.

7 Conclusion and Future Work

We have described a semi-automatic method to construct Thai WordNet lexical database from machine-readable dictionaries. Our method is operated via WordNet Builder system which takes into account the WordNet and the LEXiTRON machine-readable dictionaries. The semantic and the translation relations are extracted and evaluated according to multiple criteria. Our Thai WordNet lexical database is an electronic dictionary in which words are organized around their semantics instead of alphabetical order. Therefore, Thai WordNet lexical database can be used as a lexical information source for many applications such as information retrieval, text classification, semantic disambiguation and machine translation.

Our Thai WordNet is available for further extension. Other lexical resources could be added to increase the volume of word in the database. Other Thai taxonomy such as the Thai concept type hierarchy can be combined. Our approach for constructing Thai WordNet lexical database allows further technical and methodological improvements elsewhere in acquisition process. Alternative methods such as conceptual distance method can be considered to improve the methodology of constructing the Thai WordNet. To eliminate the barrier of mankind's communication, the ultimate goal of constructing WordNet is to link all languages in the world together.

8 Acknowledgements

Thanks to Mr. Xavier Farreres for valuable advices.

References

J. Atserias, S. Climent, X. Farreres, G. Rigau and H. Rodríguez. 1997. *Combining Multiple Methods for the automatic Construction of Multilingual WordNets*, in proceedings of International Conference "Recent Advances in Natural Language Processing" (RANLP'97). Tzigov Chark, Bulgaria.

H .H. Chen et. al. 2000. *Construction of a Chinese-English WordNet and its application to CLIR*, in Proceedings of the Fifth International Workshop on Information Retrieval with Asian Languages (Hong Kong, Sept.-Oct. 2000). ACM Press, New York, NY, 189-196.

C. Fellbaum. 1998. *WordNet: An electronic lexical database*. MIT Press, Cambridge, Massachusetts.

X. Farreres, G. Rigau, and H.Rodríguez. 1998. *Using WordNet for Building WordNets*, in Proceedings of COLING/ACL Workshop on the Usage of WordNet in Natural Language Processing Systems, pages 65-72.

S. Harabagiu. 1998. *Usage of WordNet in Natural Language Processing Systems*, In Proceedings of the Workshop, Montreal, Quebec.

K. Javaras and V. Wiesner. 2002. *Introduction to Generalized Linear Models*, Lecture notes in Introduction to Statistical Modeling, Trinity, Department of Statistics Oxford University.

R. V. Krejcie and D. W. Morgan. 1970. Determining Sample Size for Research Activities. Educational and Psychological Measurement (30), pages 607-610.

M. Neff et. al. 1993. *Get It Where You Can: Acquiring and Maintaining Bilingual Lexicons for Machine Translation*. In Proceeding of AAAI Spring Simposium on Building Lexicons fro Manchine Translation, Stanford University.

P. Palingoon et. al. 2002. *Qualitative and Quantitative Approaches in Bilingual Corpus-Based Dictionary*, In Proceeding of SNLP-Oriental COCOSDA.

E. Pianta et. al. 1990, *MultiWordNet Developing an aligned multilingual database*, ITC-irst, Centro per la Ricerca Scientifica e Tecnologica

G. Prószéky and M. Mihăltz. 2002. *Semi-automatic Development of the Hungarian WordNet*, Proceedings of the LREC 2002 Workshop on WordNet Structures and Standardization, Las Palmas.

M. Rila.1998. *The User of WordNet in Information Retrieval*. In Proceedings of ACL Workshop on the Usage of WordNet in Natural Language Processing Systems, pages 31-37.

P. Sathapornrungkij. 2004. *A Semi-Automatic Construction of Thai WordNet Lexical Database from Machine-Readable Dictionaries*, Master's thesis, Mahidol University.

V. Sornlertlamvanich and M. Boriboon. 1995. *Technical Report - Thai Concept Classification.* Linguistics and Knowledge Science.

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