Building an Indonesian WordNet

Desmond Darma Putra, Abdul Arfan, and Ruli Manurung
Faculty of Computer Science
University of Indonesia
ddp41@ui.edu, abar40@ui.edu, maruli@cs.ui.ac.id

Abstract

A WordNet is a useful lexical resource where specific senses of words are clustered together into synonym sets, and semantic relationships between these sets are specified. This paper describes an ongoing project to create an Indonesian WordNet using the expand model approach, i.e. by mapping existing WordNet entries to Indonesian word sense definitions. We discuss some issues encountered during the development of a web-based application that facilitates this mapping. Some initial results of an experiment to measure the inter-annotator reliability are presented, along with a brief discussion on the constructed Indonesian WordNet.

1. Introduction

The original Princeton WordNet [3], or simply PWN, is a lexical resource for the English language containing a large database of open-class words, i.e. nouns, verbs, adjectives, and adverbs. These words are clustered together based on their meaning into synonym sets, or *synsets*. Each instance of a word in a synset is marked as a separate *sense* of that word. PWN further specifies semantic relations between synsets and specific word senses, among others antonymy, hypernymy, and meronymy. As a freely-available semantic resource that comprehensively catalogues distinct English word senses, PWN is widely used in various NLP research.

WordNets have since been created for many other languages, and currently there are WordNets for over 40 different languages in some shape or form. Notable efforts include EuroWordNet [11], BalkaNet [9], and the establishment of the Global WordNet Association.

A WordNet for Indonesian would serve as a useful resource for research communities in various areas, e.g. lexicography, information retrieval, machine translation, knowledge engineering, etc. Additionally, many additional language-independent resources mapped to PWN, e.g. pictorial resources for augmentative and alternative communication (AAC) users [5], would become available to Indonesian systems.

This paper presents work carried out at the Information Retrieval Lab, Faculty of Computer Science, University of Indonesia, to develop an Indonesian WordNet by mapping PWN synsets to Indonesian word sense definitions. In Section 2 we present the general methodology of the project, before discussing the major aspects in Sections 3 to 6. A discussion of the initial mapping results is presented in Section 7, along with an outlining of the next step, i.e. using these mappings to construct the Indonesian WordNet (Section 8).

2. Methodology and approach

There are two general approaches to constructing a WordNet: the *expand* model and the *merge* model. In the former approach, the synsets in PWN (or another existing WordNet) are translated to the target language. Subsequently, the semantic information from PWN, e.g. semantic relationships holding between synsets, are transferred over. In the latter approach, synsets in the target language are independently specified, along with semantic relationships holding between them. These synsets are then mapped to PWN synsets through the definition of an equivalence relation. The expand approach has the advantage of being easier to accomplish, but with the tradeoff that the resulting WordNet may be too dependent on the structure of PWN. The merge approach is perhaps the most principled one, taking into consideration semantic distinctions unique to the target language, but is more costly to develop. Many real WordNet construction projects employ a combination of the approaches [8].

As an initial attempt to building an Indonesian WordNet, we mainly adopt the expand model approach. Since this approach calls for the translation of existing WordNet synsets into a target language, one possible approach is to take the PWN and manually (i) translate its synsets and (ii) identify words that convey them in a target language.

However, instead of creating new translations of English WordNet entries into Indonesian, we cast the process as

one of identifying correspondences between English WordNet entries with entries in an **existing** Indonesian dictionary, namely the *Kamus Besar Bahasa Indonesia*, or simply the KBBI[1]. This resource is considered by many to be the definitive dictionary of Indonesian.

The advantages of this approach are two-fold. Firstly, the lexical entries and definitions in the KBBI have been well-studied and justified from a lexicographical viewpoint. Second, we believe that the task of identifying whether an existing Indonesian sense definition is indeed equivalent to an English synset gloss should intuitively be an easier task to accomplish than to create a new translation from scratch. Crucially, we argue that this is a task that does not require expert lexicographical knowledge.

The goal of our project is thus to establish the correct concept mapping between PWN synsets and KBBI sense definitions (Section 3). To achieve this, we developed a web-based application (Sections 5 and 6) and relied on a large number of users to manually perform the disambiguation (Section 4).

3. Concept mapping

As described above, WordNet is first and foremost a semantic lexicon. Thus, its basic elements are semantic elements, i.e. synsets, which are further elaborated as words that convey these concepts. On the other hand, the KBBI is a traditional dictionary, where its basic elements are orthographic elements, i.e. lemmas. These lemmas are subsequently classified into broad sense categories (*makna*), sublemmas, and specific senses of the sublemma (*definisi*).

Therefore, there is an inherent assymetry in the mapping process: whereas WordNet synsets represent unique semantic entities that are conveyed by several words, the KBBI does not cluster sublemmas that share the same meaning in the same way. The KBBI definitions therefore contain separate entries for semantic entities that are conceptually identical. We will thus view the KBBI as a WordNet with extremely fine-grained sense distinctions, with each sense forming a singleton synset. As a result of mapping WordNet synsets to KBBI sense definitions, semantic equivalence classes, i.e. synonymy relationships, will be established across the various sense definitions in the KBBI. Figure 1 illustrates this phenomenon.

For a more concrete example, Table 1 presents a selection of synsets in PWN, all of which can be conveyed by the orthographic form *time*. In Indonesian, the appropriate translation for the sense of *time* in synset `107309599` is *kali* (*"**kali** ini dia berhasil"*), in synset `115245515` it is *waktu* (*"ini **waktu**nya untuk pergi"*), and finally, in synset `115129927` it is *jam* (*"apa anda tahu **jam** berapa sekarang?"*). Table 2 presents a selection of KBBI word sense definitions of these words. Since we are mapping concepts to concepts, synset `107309599` should be mapped to the specific senses `k37192` and `k37194`, whereas synset `115245515` should be mapped to the specific sense `k33882`.

As discussed above, one PWN synset can correspond to many KBBI word sense definitions, resulting in the establishment of synonymy relations over the KBBI word senses. Theoretically, this may happen in the reverse direction: a KBBI word sense definition may be mapped to several PWN synsets. However, this should be much less likely given that synsets already represent distinct semantic entities. Nevertheless, such semantic distinctions are not completely universal, and indeed it has been pointed out that WordNet's degree of polysemy is too fine-grained in some instances [2]. As a result, such a mapping may also conflate several synsets into one.

[1]The KBBI is the copyright of *Pusat Bahasa*, Indonesian Ministry of Education

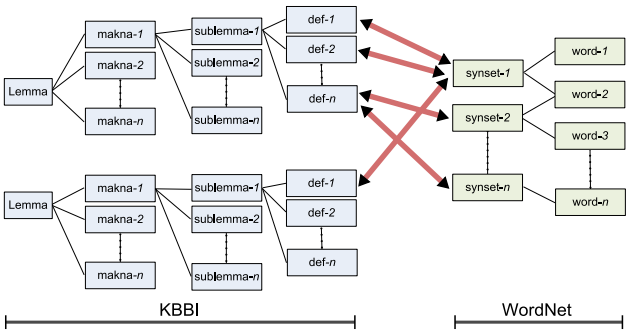

Figure 1. Diagram illustrating WordNet-to-KBBI concept mapping

Synset ID	Words	Gloss	Example
`107309599`	clip, time	an instance or single occasion for some event	*"this **time** he succeeded"*
`115245515`	time	a suitable moment	*"it is **time** to go*
`115129927`	clock_time, time	a reading of a point in time as given by a clock	*"do you know what **time** it is?"*

Table 1. A selection of PWN synsets conveyed by the word *time*

4. Obtaining annotations

Manually performing concept mapping as defined above on the entire WordNet to the entire KBBI is a costly and laborious task. Moroever, to ensure the validity of the annotations, we require multiple annotators [6]. Due to resource limitations, we are unable to employ the required number of annotators. Our strategy is therefore to create a web-based application to facilitate this annotation process, and rely on

KBBI ID	Sublemma	Gloss	Example
k37192	kali,1,1	kata untuk menyatakan kekerapan tindakan	*dl satu minggu ini, dia sudah empat kali datang ke rumahku*
k37193	kali,1,2	kata untuk menyatakan kelipatan atau perbandingan (ukuran, harga, dsb)	*harga barang kebutuhan pokok pd tahun ini dua kali lebih mahal dp harga pd tahun yg lalu*
k37194	kali,1,3	kata untuk menyatakan salah satu waktu terjadinya peristiwa yg merupakan bagian dr rangkaian peristiwa yg pernah dan masih akan terus terjadi	*untuk kali ini ia kena batunya*
k37195	kali,1,4	kata untuk menyatakan perbanyakan atau pergandaan	*dua kali dua sama dng empat*
k37209	kali,2,1	sungai	
k37211	kali,3,1	barangkali	*kali dia sakit*
k37212	kali,4,1	pejabat tinggi dl masyarakat di Sulawesi Selatan	
k33880	jam,1,1	alat pengukur waktu (spt arloji, lonceng dinding)	
k33881	jam,1,2	waktu yg lamanya 1/24 hari (dr sehari semalam)	
k33882	jam,1,3	saat tertentu yg dl arloji jarumnya yg pendek menunjuk angka tertentu dan jarum panjang menunjuk angka 12 (pd lonceng disertai dng dentang suara bandul memukul logam atau bel); pukul	*ia bangun jam lima pagi*

Table 2. A selection of KBBI word senses

the large number of web users to assist in the creation of the WordNet-to-KBBI mappings[2]. This is in some sense similar to the approach taken by the KUI project [7]. However, whereas the KUI system facilitates discussion between annotators regarding sense distinctions, our design deliberately encourages independent judgment. This was done to prevent bias during the evaluation of inter-annotator reliability.

The task asked of the user is to determine whether, for a given WordNet synset, a particular KBBI sense definition is indeed a semantic equivalent. As it would be completely unrealistic to expect a human annotator to make this judgment for every possible word sense listed in the KBBI, we made the task more feasible by determining, for each WordNet synset, a subset of *candidate* KBBI word senses. This is determined through automatic means, i.e. existing bilingual lexicons. Note that these candidates are rough suggestions that have yet to be disambiguated. Moreover, the quality of the suggestion depends heavily on the translation resource used, and some correct translations might be missing from the suggestions. To compensate for this, human annotators are also given the chance to manually search for and select KBBI word senses outwith the candidate set.

We focused the annotation efforts on a set of *common base concepts*, a subset of WordNet synsets previously identified as playing an important role across different languages. We used the common base concepts[3] identified from the EuroWordNet [11] and BalkaNet projects [9]. Since these were defined for a previous version of WordNet, we automatically mapped them to WordNet v3.0 using the provided glosses, yielding 4369 base concept synsets.

We distributed these synsets into groups of 50 synsets, with each participant being assigned to a group. To ensure that we obtained a suitable range of multiple annotations of a given synset for evaluation purposes, we designed an overlapping distribution of the synsets into 100 groups, with each group being composed of the following:

- 33 synsets from a subset of 3300 base concepts.
- 10 synsets from a subset of 500 base concepts.
- 5 synsets from a subset of 100 base concepts.
- 2 synsets from a subset of 20 base concepts.

All the above subsets are disjoint. When synsets from a subset have been exhausted, the selection wraps back to the beginning of the subset. Using this mechanism, once all the groups have been annotated, we would have 3300 synsets that are mapped once, 500 synsets mapped twice, 100 synsets mapped five times, and 20 synsets mapped ten times.

Due to consideration of the KBBI copyright, we did not open the annotation web-based application to the entire Internet. Instead, users were restricted to those with access to the University of Indonesia intranet. To encourage participation, a modest amount of incentives were offered in a lottery between the annotators.

5. Lexical database preparation

In this section we describe the various lexical resources used in our work, namely:

- An SQL database version of the most current PWN, version 3.0[4]. It contains 117659 synsets and 147306 distinct words.
- An electronic version of the *Kamus Besar Bahasa Indonesia*, or KBBI, which was developed at the Faculty of Computer Science, University of Indonesia, during the mid-90s. We preprocessed this resource by correcting several systematic errors and cleaning up incorrect tags. The resulting lexicon contains 96640 unique word sense definitions.

[2] cf. the current Web 2.0 buzzword: *crowdsourcing*

[3] `http://www.globalwordnet.org`

[4] `http://wnsqlbuilder.sourceforge.net`

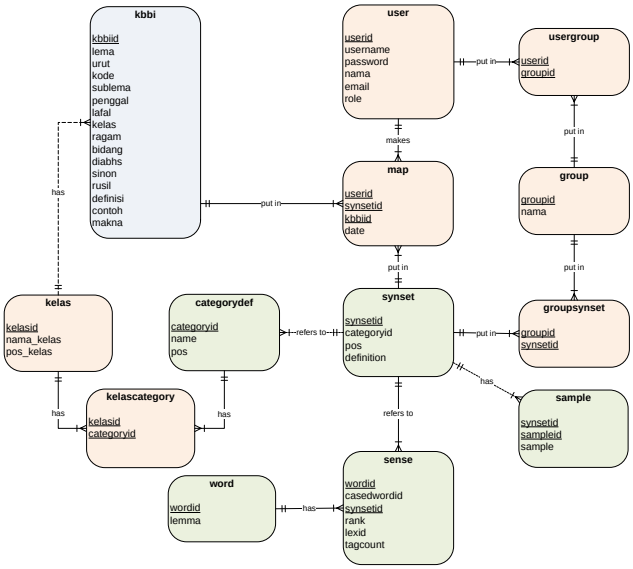

Figure 2. Complete ERD of the lexical database

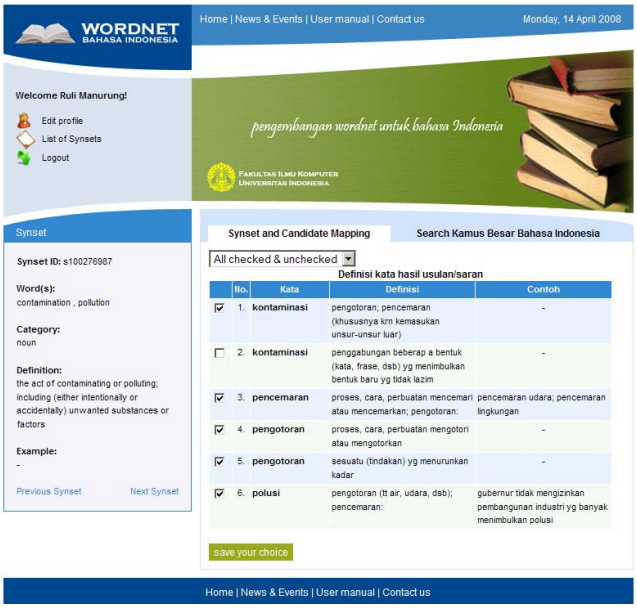

Figure 3. Screenshot of web-based mapping application

- A bilingual English-Indonesia dictionary was constructed from various online resources, including a handcrafted dictionary by Hantarto Widjaja[5] (30818 word pairs) and kamus.net (55657 word pairs).

All of the above resources were combined into a relational database implemented using the open-source MySQL DBMS[6]. Figure 2 shows the entity relationship diagram (ERD) [12] of a portion of the database. The main components are the `kbbi` and `synset` tables, which represent the KBBI senses WordNet synsets, respectively. Connecting these two resources is the `map` table, which contains human annotations defining the concept-to-concept mapping. It stores additional information, e.g. details of the annotator.

6. Web-based application

The annotation tool was developed as a web-based application using a popular collection of open source tools, i.e. the so-called LAMP stack based on the MySQL DBMS backend and PHP web scripting language.

The general desiderata of the application was as follows:

- User management, i.e. the automatic assignment of groups of synsets to be mapped to annotators.
- A usable user interface that facilitates the mapping of WordNet synsets to a set of candidate word senses.
- A facility for searching for KBBI word senses outwith the candidates.
- Reporting of statistics, i.e. number of users, number of synsets mapped, facility to extract the mappings for further analysis.

Several prototypes of the application were developed, and an initial version was tested on a small group of users before the final version was deployed. Figure 3 shows a sreenshot of the main mapping interface. The system was made available online at `http://bahasa.cs.ui.ac.id/wordnet`.

7. Initial results

At the time of writing, mappings for a total of 1441 unique base concept synsets have been obtained, to a total of 3074 distinct KBBI sense definitions, by a total of 64 different annotators. To measure inter-annotator reliability, we computed the Fleiss kappa statistic [4], a measure in the interval of [0,1] indicating the degree of agreement between several independent annotators, taking into account the probability of chance agreement.

For each synset mapped by at least 2 annotators, we compute the level of agreement between the different human annotators at the task of determining whether a given WordNet synset is semantically equivalent to each of the provided candidate KBBI senses.

[5] `http://hantarto.definitionroadsafety.org`
[6] `http://www.mysql.com`

We use the term *sense judgment* to denote an instance of a human annotator indicating that a given KBBI sense is indeed semantically equivalent to a given WordNet synset. We also introduce the concept of a **minimum judgment voting threshold** v indicating the number of annotators that must agree on a specific sense judgment before it is considered to be valid. For example, setting $v = 2$ means that if an annotator deemed KBBI sense x to be semantically equivalent to WordNet synset y, but was the only annotator to do so, we ignore that particular judgment.

We measured the Fleiss kappa whilst gradually imposing a higher minimum judgment voting threshold. The results are shown in Table 3. We experimented with two variants, i.e. $v = n - 1$ and $v = n$, where n is the minimum number of users tasked with annotating a particular synset. Thus, under the $v = n - 1$ variant, for synsets mapped by at least $2 \leq n \leq 6$ users, we only consider sense judgments agreed by at least $1 \leq v \leq 5$ users, respectively. Under the $v = n$ variant, for synsets mapped by at least $2 \leq n \leq 6$ users, we only consider sense judgments agreed by at least $2 \leq v \leq 6$ users, respectively. Column 1 under the $v = n - 1$ variant thus represents "raw" agreement.

We also computed the Fleiss kappa to account for the sense judgments that were obtained through the KBBI search facility, i.e. whereupon a user decided to add a KBBI sense not initially contained in the candidate set. We computed two variants: for the first variant we consider the range of possible judgments to span the entire set of KBBI word senses, whereas in the second variant we only consider the range of possible judgments to be the union between the suggested candidate set and the set of KBBI senses searched for by a human annotator.

There is no universally accepted method of interpreting Fleiss kappa results. The magnitude of the value is dependent upon the number of categories and subjects of an annotation task, and is thus not directly comparable across experiments. Nevertheless, there are several suggestions, for example a value < 0.4 is poor, $0.4 - 0.75$ is intermediate to good, and > 0.75 is excellent [1].

For each experiment variant, Table 3 shows the number of synsets whose Fleiss kappa is considered to be poor, intermediate to good, and excellent, as well as the average and standard deviation of the Fleiss kappa values.

8. Constructing the Indonesian WordNet

The next step is to extract the sense judgments obtained from the web-based experiment and to construct an Indonesian WordNet (IWN). Essentially, this involves conflating KBBI word senses that are judged to map to a single WordNet synset, and constructing an IWN synset whose gloss consists of the concatenation of the various KBBI word sense definitions.

Each IWN synset would then be mapped to a set of Indonesian orthographic wordforms, obtained from the union of sublemmas belonging to the various mapped KBBI senses.

Another issue to be considered is the part-of-speech (POS) compatibility of a sense judgment. All WordNet synsets are classified as a noun, verb, adjective, or adverb. KBBI entries are classified similarly, but may also belong to other categories. More problematic, however, is the fact that many entries are not marked for syntactic category. Thus, some sense judgments might incorrectly map an object to a process, which woud be semantically incongruous. Our plan is to explicitly indicate which word senses are POS-compatible with its PWN counterparts. Of the current sense judgments reported in Section 7, the POS-compatible judgments map 1237 distinct PWN base concepts to 2340 distinct KBBI senses.

Theoretically, IWN can simply use the same synset identifiers as PWN, thus establishing a trivial mapping. However, as discussed in Section 3, it is possible that a KBBI sense is mapped to several PWN synsets, thus conflating them. In this case, it is necessary to assign a different identifier to the IWN synset, and to explicitly mark it as being equivalent to the various PWN synsets. If such a scenario were to arise, it could be viewed as mitigating both the issue of WordNet's excessive polysemy and the alleged drawback of the expand model approach, i.e. that a WordNet in another language would become too dependent on the English semantic structure of PWN, which is by no means universal.

Lastly, we copy the semantic relationships holding between PWN synsets over to the IWN synsets, e.g. hypernyms/holonyms and meronym/holonyms. Using all of the currently obtained sense judgments, 903 such semantic relations are established for IWN synsets. Restricting the judgments to just the POS-compatible ones yields 735 semantic relations. This can be augmented by treating such relations to be transitive, similar to the approach in [10].

Figure 4 shows the ERD for the IWN database. An initial web interface is available at `http://bahasa.cs.ui.ac.id/iwn`.

9. Summary

We have presented the development of an Indonesian WordNet using the expand approach by harnessing large numbers of human annotators through a web-based application. We believe this can be an effective method to obtain human judgments. Initial experiments suggest that the mappings can be used to construct a fairly reliable Indonesian WordNet. The work involved in extracting the human judgments and using them to construct an Indonesian WordNet has been carried out.

	All KBBI	Search KBBI	$v = n - 1$					$v = n$				
			1	2	3	4	5	2	3	4	5	6
< 0.4	121	149	116	4	0	0	0	4	0	0	0	0
$0.4 - 0.75$	85	61	62	23	4	2	0	23	4	2	0	0
> 0.75	68	64	64	31	9	6	3	215	54	11	8	3
Average	0.476	0.357	0.425	0.713	0.840	0.914	1.000	0.931	0.964	0.947	1.000	1.000
Std. Dev	0.370	0.459	0.430	0.311	0.209	0.167	0.000	0.195	0.117	0.135	0.000	0.000

Table 3. Fleiss kappa results

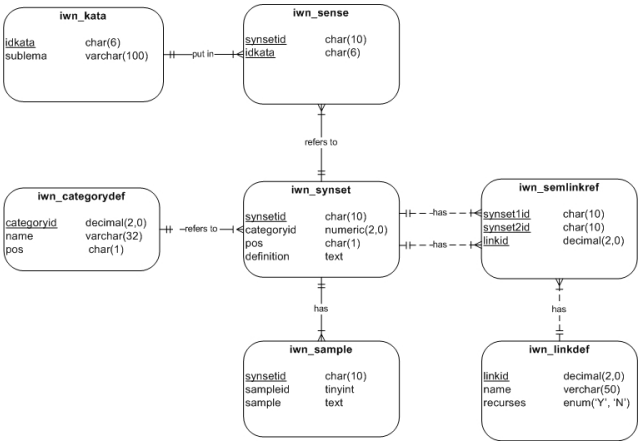

Figure 4. Diagram illustrating WordNet-to-KBBI concept mapping

Acknowledgment

The work presented in this paper is supported by an RUUI (Riset Unggulan Universitas Indonesia) 2007 research grant from DRPM UI (Direktorat Riset dan Pengabdian Masyarakat Universitas Indonesia).

References

[1] K. E. Emam. Benchmarking kappa for software process assessment reliability studies. Technical Report ISERN-98-02, International Software Engineering Research Network Technical Report, 1998.

[2] C. Fellbaum. A semantic network of english: The mother of all wordnets. *Computers and the Humanities*, 32(2):209–220, March 1998.

[3] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, May 1998.

[4] J. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.

[5] R. Manurung, D. O'Mara, H. Pain, G. Ritchie, and A. Waller. Building a lexical database for an interactive joke-generator. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy, May 2006.

[6] R. Passonneau, N. Habash, and O. Rambow. Inter-annotator agreement on a multilingual semantic annotation task. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy, May 2006.

[7] V. Sornlertlamvanich, T. Charoenporn, K. Robkop, and H. Isahara. Collaborative platform for multilingual resource development and intercultural communication. In T. Ishida, S. R. Fussell, and P. T. J. M. Vossen, editors, *IWIC*, volume 4568 of *Lecture Notes in Computer Science*, pages 91–102. Springer, 2007.

[8] D. Tufiş, E. Barbu, V. B. Mititelu, R. Ion, and L. Bozianu. The romanian wordnet. *Romanian Journal of Information Science and Technology*, 7(1–2):107–124, 2004.

[9] D. Tufiş, D. Cristea, and S. Stamou. Balkanet: Aims, methods, results and perspectives. a general overview. *Romanian Journal of Information Science and Technology*, 7(1–2):9–43, 2004.

[10] L. L. Tze and N. Hussein. Fast prototyping of a malay wordnet system. In *Proceedings of the Language, Artificial Intelligence and Computer Science for Natural Language Processing Applications (LAICS-NLP) Summer School Workshop*, pages 13–16, Bangkok, Thailand, October 2006.

[11] P. Vossen, editor. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, 1998.

[12] J. Whitten, L. Bentley, and K. Dittman. *Systems Analysis and Design for the Global Enterprise*. McGraw-Hill, 7th edition, 2007.