Creating the Open Wordnet Bahasa[⋆]

Nurril Hirfana Mohamed Noor and Suerya Sapuan and Francis Bond

School of Humanities and Social Sciences
Nanyang University of Technology
hirfananoor@yahoo.com, suerya.sapuan@gmail.com, bond@ieee.org

Abstract. This paper outlines the creation of the Wordnet Bahasa as a resource for the study of lexical semantics in the Malay language. It is created by combining information from several lexical resources: the French-English-Malay dictionary **FEM**, the KAmus Melayu-Inggeris **KAMI**, and wordnets for English, French and Chinese. Construction went through three steps: (i) automatic building of word candidates; (ii) evaluation and selection of acceptable candidates from merging of lexicons; (iii) final hand check of the 5,000 core synsets. Our Wordnet Bahasa is only in the first phase of building a full fledged wordNet and needs to be further expanded, however it is already large enough to be useful for sense tagging both Malay and Indonesian.

Keywords: Wordnet, Malaysian, Indonesian, Hyponymy, Open Source

1 Introduction

The dictionary is a very important lexical resource in any field of studies. However, WordNet, originally created by academics at Princeton University, is just as important if not greater (Fellbaum, 1998). In fact, it is a source of reference that takes the traditional dictionary to a whole new level. While a dictionary can provide information such as the meaning, synonyms and parts of speech, and can organise them in alphabetical order, a wordnet is able to organise the words into a set of cognitive synonyms (synsets) which express distinct concepts. This reason has been the motivation for the creation of the various wordnets for various languages.

There is currently no wordnet available for Malay despite the great number of wordnets available for many languages. Hence, this paper will attempt to create a lexical database for the Malay language based on alignments with other lexical resources — the French-English-Malay (**FEM**) dictionary, the English wordnet, **KAMI** and wordnets for Chinese and French. Crossing lexicons over several languages contributes to the accuracy of the Wordnet Bahasa. This wordnet will be released under an open source license (Creative Commons Attribution) in order to make it fully accessible to all potential users.

Bahasa Melayu "the Malay language" is one that had been standardized over time with the aim of formal usage of the language. It derived from the variety of Malay languages that exist in the different parts of the Malay Archipelago, and is now widely used in Malaysia, Singapore, parts of Thailand and Brunei. The language spoken in Indonesia (Bahasa Indonesia) is very similar, and largely mutually intelligible. In this paper we will use **Malay** for standard Malay (the official language of Malaysia, ISO 639-3 code **`zsm`**), **Indonesian** to refer to the official language of Indonesia (**`ind`**) and **Bahasa** to refer to the generic Malay language that includes both (**`msa`**). Bahasa is the official language of four South Eastern Asian countries, namely Malaysia, Indonesia, Brunei and

[⋆] We would like to thank Key-Sun Choi for allowing us to use the Goi-Taikei–CoreNet–Wordnet mappings; Nippon Telegraph and Telephone Corporation for allowing us to use **KAMI**, the **FEM** developers for allowing us to use **FEM**, and all the developers of the various Wordnets. This research was supported in part by the joint JSPS/NTU grant on *Revealing Meaning Using Multiple Languages*.

Singapore. Some people from The Philippines, Thailand, Burma, Sri Lanka, Cocos Island and Christmas Island also use it. There are about 40 million native Bahasa speakers worldwide.[1]

Spelling reforms in the 1970s harmonized the orthographic conventions of Malay and Indonesian, making the written forms very similar (Asmah Haji Omar, 1975). Because of the enormous overlap in vocabulary (close to 98% by our measure, see Section 4.3) we decided it was possible to create a single wordnet for both languages: the Wordnet Bahasa. The vast majority of words are usable for both Malay and Indonesian and we specially mark those words that are used exclusively in one language. We hope that by building a single, open wordnet for both Malay and Indonesian we can help to create a strong lexical resource for the region.

2 Previous Work

The most common approaches to building a wordnet for a new language are automatic or semi automatic approaches. There are two main methods: the merge and the extend approach (Vossen, 2005). The merge approach would require the construction of an independent lexicon for a certain language based on monolingual resources, after which, it is mapped to other wordnets. The extend approach on the other hand is executed by obtaining a set of synsets from Princeton WordNet (PWN), and then translating it into the target language. This method allows the preservation of the original structure of the wordnet. We have opted for the extend approach both because of its simplicity and because the resulting wordnet is automatically aligned to all other wordnets.

The idea of extending the synsets with reference from not just the PWN but at least one other wordnet in a different language provides a much stronger foundation laid before the construction of a new wordnet. In Bond *et al.* (2008), the authors pointed out that by using wordnets in multiple languages to disambiguate the target language (Japanese in their study), a more reliable prototype could be provided. This multiple-pivot technique was then adapted to suit the needs of the Wordnet Bahasa, as will be explained in the next section.

There has already been some work on building wordnets for Malay and Indonesian. Lim and Hussein (2006) serves as a good head start for the building of a Malay wordnet. The paper suggests finding the prototype based on sense alignments with Kamus Inggeris Melayu Dewan (KIMD) and the English wordnet.

According to Lim and Hussein (2006), this "… fast prototyping exercise (would require the creation of) semantic relations between the Malay synsets based on the existing relations between their English equivalents". This method is an elaboration of the merge methodology. Lim and Hussein (2006) managed to build 12,429 noun synsets and 5,805 verb synsets. While this is by no means exhaustive, it is at the very least a rough gage of the minimum possible range of words in a Malay wordnet. In the final discussion of the paper, Lim and Hussein (2006) point out that the bottleneck for their prototype "is in the dictionary used". Unfortunately, we do not have access to the same Malay lexicon, so we cannot directly implement their approach.

There have been two approaches to building an Indonesian wordnet. The first was an expand approach, and created a small prototype (Putra *et al.*, 2008). The second also used an expand approach, and then corrected entries using the infrastructure from the Asian Wordnet Project (Riza *et al.*, 2010). The Indonesian Wordnet at the Asian Wordnet currently has 33,726 synsets; 38,394 words and 65,206 senses (word-synset pairs).[2] The lexicons used to expand were bilingual English-Indonesian and thus did not enable the use of multiple pivots.

3 Resources

We used two lexicons: **FEM**, which contains entries with French, English and Malay as well as hypernyms in French; and **KAMI**, which contains Malay, English and Chinese as well as semantic classes from the Goi-Taikei ontology.

[1] `http://www.ethnologue.com/show_language.asp?code=msa`

[2] `http://id.asianwordnet.org/`

We used four wordnets: one for English, one for Chinese and two for French as the original French Wordnet has not been maintained, so we supplemented it with the new Wordnet Liberé du Français (WOLF). As these map to different versions of the English WordNet, we used mappings to harmonize them (Daude *et al.*, 2003). To map between the Goi-Taikei ontology and wordnet, we used the mappings produced by CoreNet (Kang *et al.*, 2010).

3.1 Malay Lexicons

We used two lexicons **FEM** and **KAMI**.

The Malay-English Dictionary **KAMI**: KAmus Melayu-Inggeris was compiled by NTT-MSC (Quah *et al.*, 2001), based on a dictionary produced originally by a translation company. The dictionary currently has 67,670 Malay words with English translations. 69% have only one translation, 19% have two, 7% have three; the average number of translations is 1.57, giving 106,558 Malay-English pairs.

Each entry in the dictionary consists of the following fields: (1) Malay index word; (2) Malay root word; (3) Malay POS; (4) detailed syntactic features; (5) semantic classes; (6) English translation; (7) English comments; (8) Chinese translation. All entries have values for fields 1,2 and 3; most have syntactic features. 22% have Chinese translations and 28% have semantic classes from the Goi-Taikei (**GT**) ontology (Ikehara *et al.*, 1997). The Goi-Taikei ontology consists of 2,710 semantic classes, providing an upper level ontology. It was originally designed for Japanese, but has also been used for Chinese, English, Korean and Malay.

English and Chinese translations and comments are provided for use in a machine translation system, as well as an aid for non-Malay speakers. Semantic classes were automatically produced from a variety of sources, including deducing them from the associated classifiers and finding them in other lexicons or resources such as International Standard Organization (ISO) language and currency names (Quah *et al.*, 2001), and still contains some errors.

We also used **FEM**: the French-English-Malay Lexicon (Lafourcade *et al.*, 2003). We combined the general lexicon and a specialist lexicon of computational terms, giving 33,022 lexical entries. Each entry comes with: (1) French headword; (2) pronunciation; (3) part of speech; (4) superordinate term in French (46% of entries); (5) English equivalent; (6) Malay equivalent; (7) French example (30%); (8) English example (30%); (9) Malay example (30%). The dictionary had been automatically compiled and hand-corrected with some errors remaining, especially in the Malay equivalents.

We converted both lexicons to the following format (ignoring fields that we won't use):

(1)
$$\begin{bmatrix} \textit{lexical entry} & \\ \text{Malay} & m_0, \ldots m_n \\ \text{English} & e_0, \ldots e_m \\ \text{French/Chinese} & f_0, \ldots f_o \\ \text{Part-of-Speech} & \left\{ \text{noun, verb, adjective, adverb, other} \right\} \\ \text{Hypernym} & \left\{ \text{French word} \mid \textbf{GT} \text{ class} \right\} \end{bmatrix}$$

Each entry has one or more words in Malay, English and French/Chinese plus possibly a hypernym, expressed either as a French word or as Goi-Taikei semantic class. They also have a part-of-speech which we map into either one of the four open classes used in WordNet, or the class **other** which is used for closed class words.

3.2 WordNets and Mappings

Because we had dictionaries linking Malay to English, Chinese and French, we needed wordnets for these three languages, summarized in Table 1. For English, we used the Princeton WordNet

(Fellbaum, 1998), the original wordnet, and the largest so far. For Chinese, we used the Chinese Wordnet created by (Xu *et al.*, 2008), with some normalization (removing bracketed data, leading and trailing punctuation and white space, removing affixes attached to adjectives and adverbs such as 的 *de* and 地 *zi*). For French, we created a new wordnet (which we will just call the French Wordnet) by combining entries from the French part of Euro WordNet (Vossen, 1998) and the Wordnet Liberé du Français (Sagot and Fišer, 2008). The combined wordnet had considerably better coverage than either of its components.

All of the wordnets were linked to some version of the English wordnet (shown in Table 1). We used the mappings produced by Daude *et al.* (2003) to harmonize them.

Language	English	Chinese	French		
Wordnet	Princeton		Combined	Euro WordNet	WOLF
Synsets	117,659	109,140	44,914	31,601	21,951
Senses	206,941	161,655	77,015	44,920	32,689
Words	155,287	102,364	49,420	37,364	18,787
version	3.0	2.0	3.0	1.5	2.0

Table 1: Wordnet Sizes

To map between the Goi-Taikei (**GT**) ontology and PWN, we used the mappings produced by CoreNet (Kang *et al.*, 2010). CoreNet is an extension of Goi-Taikei to Chinese and Korean. These consist of a table matching CoreNet classes to one or more wordnet synsets. We were also given a table matching **GT** classes to CoreNet classes. The **GT**-CoreNet mapping is very accurate, as CoreNet design was strongly influenced by Goi-Taikei (Korterm, 2005). The CoreNet-wordnet mapping is automatically produced, we found it quite accurate. We crossed the two tables to get a single **GT**-corenet-wordnet mapping.

The combined wordnets can be thought of as having entries like the following (ignoring irrelevant information).

(2)
$$\begin{bmatrix} \textit{synset} & \\ \text{Lexemes} & \begin{bmatrix} \text{English} & e_0, \ldots e_m \\ \text{Chinese} & c_0, \ldots c_n \\ \text{French} & f_0, \ldots f_o \end{bmatrix} \\ \text{Part-of-Speech} & \left\{ \text{noun, verb, adjective, adverb} \right\} \\ \text{Relations} & \begin{bmatrix} \text{Hypernym} & \textit{synset} \\ \text{Meronym} & \textit{synset} \\ \ldots & \end{bmatrix} \end{bmatrix}$$

4 Method

Building the Wordnet Bahasa was done in three steps: (i) automatically building candidates; (ii) evaluating and selecting acceptable groups; (iii) hand correcting the 5,000 most common concepts (core synsets).

4.1 Automatic Construction

The construction broadly follows the matching through multiple pivot approach of Bond and Ogura (2007). We want to match lexical entries (which have Malay words associated with them) to wordnet synsets.

For each word in the lexicon, we try to link to each synset that has the same part-of-speech. We have three pivots for this: the English term, the French or Chinese term and the hypernym.

We first link through the terms, and then, for each synset that matched, we see if the hypernym is compatible with the synset's hypernyms.

We will give an example for the following entries.

(3) Entry in **FEM**

lexical entry	
Malay	busur
English	bow
French	arc
Part-of-Speech	noun
Hypernym	arme "weapon"

(4) Entry in **KAMI**

lexical entry	
Malay	busur
English	bow
Chinese	弓
Part-of-Speech	noun
Hypernym	$\langle 940 : worktool \rangle$

(5) Wordnet candidates (only two of many)

a.

synset	
Lexemes	English bow Chinese 弓 French arc
Part-of-Speech	noun
Relations	Hypernym *weapon* ...
Definition	a weapon for shooting arrows, ...

b.

synset	
Lexemes	English bowing, obeisance, bow Chinese 鞠躬, 弯腰, 运弓法[3] French révérence
Part-of-Speech	noun
Relations	Hypernym *reverence, motion* ...
Definition	bending the head or body or knee as a sign of reverence ...

Considering the **FEM** entry for {busur, bow, arc} (3), we look up the combined wordnet and find one entry (5a) that matches in two languages, and several that match in only one (we only show 5b). We then look at the semantic class, and using the combined wordnet, find that *arme* "weapon" gives a synset which is a hypernym of (5a), but not (5b). We thus have a strong match to the correct synset.

When we come to the **KAMI** entry for {busur, bow, 弓} (4), we look up wordnet and also find one entry (5a) that matches in two languages, and several that match in only one (we only

[3] This is in fact an error, it means "archery" and should be in a different sysnet.

show 5b). When we look up the semantic class, the **GT**-corenet-wordnet mapping leads to the synset for `tool` "an implement used in the practice of a vocation", which is not a hypernym of any of the candidates.[4] We thus have a reasonable link to the correct synset, and only weak links to the others.

The process of matching is straightforward, the major effort was in getting all the lexical resources into compatible formats. As was shown in this example, typically there would be small errors in one or more of the resources. Actual matching was done with a series of one-off python scripts using the Natural Language Toolkit's wordnet interface (Bird *et al.*, 2009) to calculate the hypernym relation.

4.2 Selection

After matching all the candidates, we wanted to identify those that could be used as is, with an acceptable level of error. We considered the following criteria in selection:

uniq lexical entry matched only one synset
in this case we considered it monosemous so the match should be good

multi lexical entry matched through two languages
as ambiguity is expressed differently in different languages, matching through two gives a much stronger match

more lexical entry matched more than one word (in one languages)
for entries with multiple words in the same language, if these all matched the same synset it suggests it is a better match

sem lexical entry's hypernym was compatible
If a word and its hypernym both match, then it should be semantically compatible

We took a random sample of a hundred entries from each combination of these features. The major groups are shown in Table 2, including those entries that just matched through one word (one) which we did not check for accuracy as we expected the accuracy to be low. Any combination that had fewer than 100 candidates was completely hand checked, there were 417 examples of these (such as **sem+uniq+multi**). Checking was done by the first and second authors, who are bilingual in Malaysian and English. When one author was unsure, they checked with the other, with standard reference lexicons for Malaysian and Indonesian (Dewan Bahasa dan Pustaka, 2005; Pusat Bahasa, 2008) and by checking usage examples on-line.

Lexicon	**KAMI**		**FEM**	
Match	Size	Accuracy (%)	Size	Accuracy (%)
one	340,537	—	210,443	—
more	5,920	75	409	78
sem	7,137	69	**12,208**	**93**
uniq	**7,381**	**85**	4,723	79
sem+uniq	**1,340**	**86**	204	79
multi	**8,870**	**96**	**21,213**	**85**
sem+multi	**684**	**93**	**2,533**	**89**

Table 2: Lexical Entry-Synset Match Accuracy
Subsets marked in bold were included in the Wordnet Bahasa as good.

We chose the fairly low threshold of 85% accuracy, as we judged coverage to be extremely important, and it is easier to remove bad entries than add new ones.

[4] The semantic class in **KAMI** is incorrect, it should be the immediate hypernym of this class

We merged the candidates from the two dictionaries, grouping things in to only four groups: **good** according to the selection above. **ok** in that it matched two or more criteria and **—** if there was only one supporting match.

When we merged if each dictionary marked a sense as **ok**, we upgraded it to **good**, based on a random sample of a 100 such entries. This happened to a further 3,533 entries.

Type	Senses
—	497,911
ok	23,257
good	42,050

Table 3: Merged results of the automatic construction

Because of overlap in the two resources, the numbers in the merged lexicon are less than the sum of the individual lexicons.

4.3 Correction

In order to make sure of the reliability of the most common synsets, we hand corrected the 5,000 core synsets: the most common synsets used in the British National Corpus[5] (Fellbaum and Vossen, 2007). After mapping to WordNet 3.0, the actual list has the 4,960 synsets. All candidates for these entries were hand-checked, regardless of how well they matched. There were a total of 99,061 sense candidates, of which 15,951 were judged to be good.

Type	Senses	
rejected	83,365	
—	413,899	
ok	18,172	
good	30,805	Release
checked	17,524	

Table 4: Merged results of the automatic construction

During this process, candidates that were only used in either Malay or Indonesian were marked as such. The default assumption is that a sense (synset-word) mapping can be used in either Malay or Indonesian (which we tag as Bahasa). If it is restricted to use in one or the other, then we tag it as Malay or Indonesian.

5 Results and Discussion

The resulting Wordnet Bahasa counting hand-checked and high-quality automatic candidates has 19,207 synsets, 48,111 senses and 19,460 unique words. This is still quite small, in terms of types, but as the high frequency synsets are all in, it should have high token coverage when used to tag text. The average ambiguity is high ($\frac{|\text{senses}|}{|\text{words}|} = 2.47$), but this because of the high frequency (and thus highly polysemous) entries. If we take out the high frequency synsets and consider just the average ambiguity of the high-quality automatic candidates it is only 1.05.

Looking at the results in section 4.2, we can see that adding the hypernym matching gave us over a quarter of the good entries (the **sem** cell for **FEM** in Table 2). The hypernym matching was less useful for **KAMI**— an analysis of errors showed that this was mainly due to errors in the (automatically assigned) semantic classes. The classes tended to be too general, and this gave them little disambiguating power. Matching through multiple pivots was much more effective for **KAMI**. In this case, we hypothesize that the more different language (Chinese) gives more disambiguating power than French, when combined with English. Because French and English are closely related, they often show the same ambiguity.

[5] `http://wordnet.cs.princeton.edu/downloads.html`

We measured how close Malay and Indonesian are by calculating the distribution of the language tags. These only exist for the hand checked entries, of these 17,150 (97.9%) were marked as acceptable in both languages, 158 (0.9%) as acceptable only in Indonesian and 216 (1.2%) as acceptable only in Malay.

(6)

synset		
Lexemes	English	dragonfly, mosquito hawk, . . .
	Chinese	蜻蜓
	French	libellule
	Bahasa	capung
	Malay	sibur-sibur
	Indonesian	sibar-sibar
Part-of-Speech	noun	
Relations	Hypernym	*odonate*

Further investigation in this phenomenon shows that differences in Malay and Indonesian words mostly lie with nouns, other than minor spelling differences of various words. (6) is an example of this.

As can be seen above, a dragonfly in translated as *sibur-sibur* which is identified only as a Malay word, since in Indonesian, a dragonfly is a *sibar-sibar*. However, in both languages *capung* can also be used to describe this insect, showing that the two languages are highly interrelated in terms of meaning and spelling.

Another example of a difference is in translation of worms. When translated in Bahasa (both Malay and Indonesian), a worm is *cacing*. However, once the basic word divides in subordinate categories, the two Bahasa languages also divide. The Indonesian language has *cacing parasit* "roundworm" as a subordinate word for *cacing* whereas the Malay language uses *cacing keruit/cacing kerawit* "threadworm" to describe the same creature. In wordnet, *threadworm* is a hyponym of *roundworm*. This shows that on top of having slight variations in spelling and nouns, the two languages sometimes have different hierarchies.

This research was made possible by the availability of a wide variety of lexical resources: the original lexicons, wordnets of various languages, mappings between different versions of wordnet and wordnet and different ontologies. Many of these have been released freely, some of these we were granted permission to use for research. Granting access to resources makes possible entirely new applications and so should be encouraged.

The Wordnet Bahasa is released under the MIT license[6] (equivalent to the original wordnet license: it allows the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies so long as copyright is attributed to the original authors). It can be freely downloaded from `wn-msa.sourceforge.net`. We have three reasons for choosing an open license. The first is practical, creating the wordnet was a significant investment in time and labor, so we want it to be used as widely as possible, getting us the highest return on our investment. The second is moral, we were able to create the Wordnet Bahasa quickly and accurately due to the wealth of lexical resources people allowed us to use, therefore feel we should also let others build upon our work. The final reason is also practical, maintaining and extending a lexical resource is an unending struggle, by making it open we hope to get more useful feedback and user contributions.

[6] `http://www.opensource.org/licenses/mit-license.php`

6 Further Work

As this is only the first phase step toward creating a wordnet for Malay and Indonesian, much more can be done to improve it. Firstly, the Malay languages have very rich derivational morphology — we would like to extend the Wordnet Bahasa to cover derivational morphology and link the words to their stem form (which may require an extension of the data structure, the root form does not fit cleanly into the part of speech categories). Secondly, we intend to add numeral classifier relations. Thirdly, we would like to add Malay and Indonesian definition sentences. Finally, tagging a corpus with this WordNet will allow us both to get frequency information and also to check for gaps in coverage.

Currently we under-specify the language for most entries in our master database, and output two fully specified versions of the dictionary (Malay and Indonesian) for applications. As these are 98% the same, this is inefficient. We would like to enhance our lexical search interface so that we can have a combined wordnet, and extend the **`domain:usage`** relation to languages, linking individual senses to the synsets for either Malay or Indonesian as required.

Finally, we intend to continue our research on the Wordnet Bahasa in cooperation with other groups in Indonesia and Malaysia, so that we can all contribute to a single rich lexical resource.

7 Conclusions

We were able to make a rapid start in building the Wordnet Bahasa using several existing lexical resources (**FEM**, **KAMI** and many wordnets). We extend the standard matching through multiple pivot languages to also consider hypernym compatibility. We also combine Standard Malay and Indonesian into a single Wordnet Bahasa only marking those entries where the Malay language and Indonesian language were differentiated. This wordnet will serve as a platform for further work in those two languages and we intend to cooperate with teams in both Malaysia and Indonesia for future expansion.

References

Asmah Haji Omar. 1975. Supranational standardisation of spelling system: the case of Malaysia and Indonesia. In *Essays in Malaysian Linguistics*, pp. 84–101. Dewan Bahasa dan Pustaka, Kuala Lumpar.

Bird, Stephen, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly. (`www.nltk.org/book`).

Bond, Francis, Hitoshi Isahara, Kyoko Kanzaki, and Kiyotaka Uchimoto. 2008. Boot-strapping a WordNet using multiple existing WordNets. In *Sixth International conference on Language Resources and Evaluation (LREC 2008)*, Marrakech.

Bond, Francis and Kentaro Ogura. 2007. Combining linguistic resources to create a machine-tractable Japanese-Malay dictionary. *Language Resources and Evaluation*, 42(2), 127–136. (Special issue on Asian language technology).

Daude, Jordi, Lluis Padro, and German Rigau. 2003. Validation and tuning of Wordnet mapping techniques. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'03)*, Borovets, Bulgaria.

Dewan Bahasa dan Pustaka. 2005. *Kamus Dewan [The Institute Dictionary]*. Dewan Bahasa dan Pustaka, Kuala Lumpar, 4 edition.

Fellbaum, Christiane and Piek Vossen. 2007. Connecting the universal to the specific: Towards the global grid. In *First International Workshop on Intercultural Collaboration (IWIC-2007)*, pp. 2–16, Kyoto.

Fellbaum, Christine, ed. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Ikehara, Satoru, Masahiro Miyazaki, Satoshi Shirai, Akio Yokoo, Hiromi Nakaiwa, Kentaro Ogura, Yoshifumi Ooyama, and Yoshihiko Hayashi. 1997. *Goi-Taikei — A Japanese Lexicon*. Iwanami Shoten, Tokyo. 5 volumes/CDROM.

Kang, In-Su, Sin-Jae Kang, Se-Jin Nam, and Key-Sun Choi. 2010. Linking CoreNet to WordNet through KorLex — some aspects and interim consideration. In Pushpak Bhattacharyya, Christiane Fellbaum, and Piek Vossen, eds., *5th Global Wordnet Conference: GWC-2010*, Mumbai.

Korterm. 2005. *CoreNet: Multilingual WordNet*. KAIST Press. (in Korean).

Lafourcade, M., G. Sérasset, L. Metzger, A. Rahman, and C. K. Chuah. 2003. Dictionnaire Français-Anglais-Malais (FeM) - version 2. CD-ROM, Dictionnaire en version XML et Application Java. (online at `http://www-clips.imag.fr/cgi-bin/geta/fem/fem.pl?lang=en`).

Lim, Lian Tze and Nur Hussein. 2006. Fast prototyping of a Malay wordnet system. In *Proceedings of the Language, Artificial Intelligence and Computer Science for Natural Language Processing (LAICS-NLP) Summer School Workshop*, pp. 13–16.

Pusat Bahasa. 2008. *Kamus Besar Bahasa Indonesia*. Pusat Bahasa, Jakarta, 3 edition.

Putra, Desmond Darma, Abdul Arfan, and Ruli Manurung. 2008. Building an Indonesian wordnet. In *Proceedings of the 2nd International MALINDO Workshop*, CyberJaya.

Quah, Chiew Kin, Francis Bond, and Takefumi Yamazaki. 2001. Design and construction of a machine-tractable Malay-English lexicon. In *Asialex 2001 Proceedings*, pp. 200–205, Seoul.

Riza, Hammam, Budiono, and Chairil Hakim. 2010. Collaborative work on Indonesian wordnet through Asian wordnet (awm). In *Proceedings of the 8th Workshop on Asian Language Resources*, pp. 9–13, Beijing.

Sagot, Benoît and Darja Fišer. 2008. Building a free French wordnet from multilingual resources. In European Language Resources Association (ELRA), ed., *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

Vossen, Piek, ed. 1998. *Euro WordNet*. Kluwer.

Vossen, Piek. 2005. Building wordnets. `http://www.globalwordnet.org/gwa/BuildingWordnets.ppt`.

Xu, Renjie, Zhiqiang Gao, Yuzhong Qu, and Zhisheng Huang. 2008. An integrated approach for automatic construction of bilingual Chinese-English WordNet. In *3rd Asian Semantic Web Conference (ASWC 2008)*, pp. 302–341.